

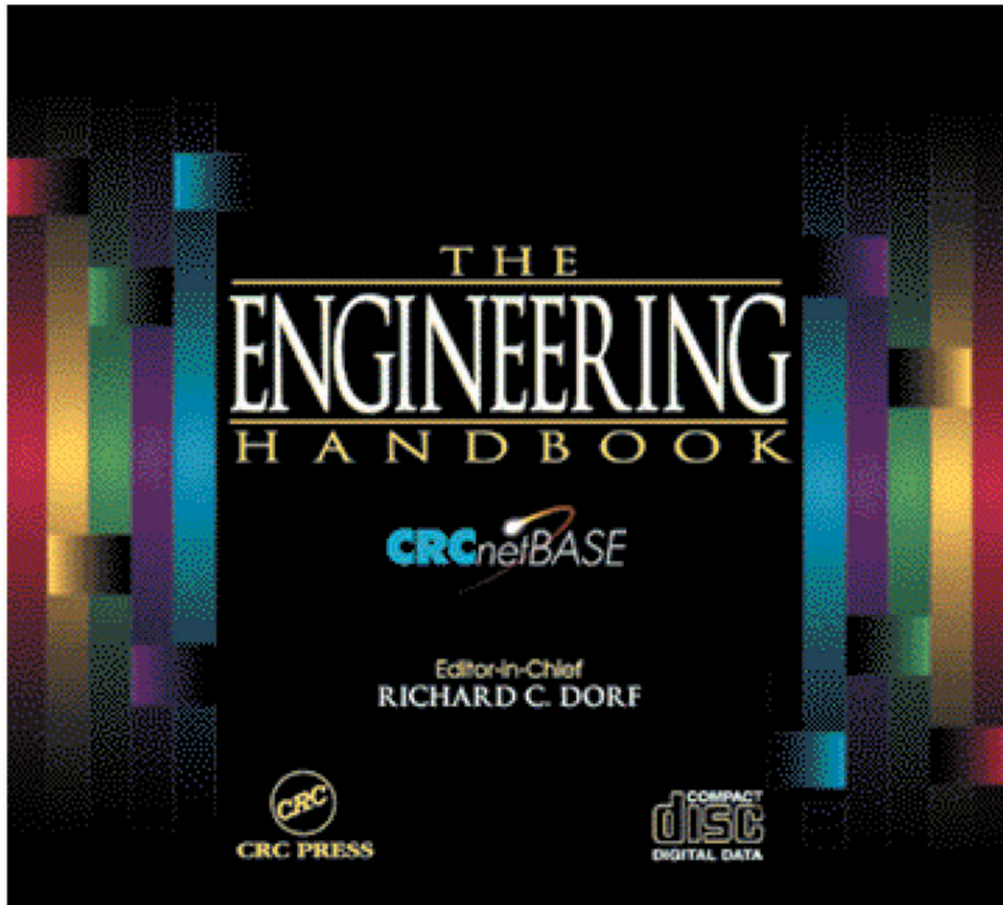
“Frontmatter”

*The Engineering Handbook.*

Ed. Richard C. Dorf

Boca Raton: CRC Press LLC, 2000

# The Engineering Handbook





### **Library of Congress Cataloging-in-Publication Data**

The engineering handbook [computer file] / Richard C. Dorf, [editor-in-chief].--CD-ROM version.

1 computer laser optical disc: 4 3/4 in.

Computer data and program.

System requirements: IBM PC; 8MB RAM;  
Windows 3.1 or higher; VGA graphics capabilities; color monitor; CD-ROM drive.

Title from title screen

Audience: Engineering professionals and postgraduate students.

Summary: Electronic version of The Engineering Handbook. Features search capabilities, zoom option, hypertext links, line drawings, photographs, bookmark and notebook functions, and the ability to print, save, and copy information into word processing program.

ISBN 0-8493-8576-8

1. Engineering--Handbooks, manual, etc. I. Dorf, Richard C. II. Engineering handbook.

TA151, 1997 00577> <MRC>

620--DC12a

97-4535

CIP

This CD-ROM contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use. Neither this CD-ROM nor any part may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, microfilming, and recording, or by any information storage or retrieval system, without prior permission in writing from the publisher. The consent of CRC Press does not extend to copying for general distribution, for promotion, for creating new works, or for resale, nor does it extend to portions of the work taken from other sources with permission of their respective copyright holders. Specific permission must be obtained in writing from CRC Press or the copyright holder for such copying. Direct all inquiries, suggestions, or problems to CRC Press LLC, 2000 Corporate Blvd., NW, Boca Raton, Florida 33431. If there are questions or problems regarding the operation of *The Engineering Handbook CD-ROM Version*, call CRC Press at 561-994-0555 extension 2515 (e-mail: [epd@crcpress.com](mailto:epd@crcpress.com)).

Library of Congress Card Number 97-4535

ISBN 0-8493-8576-8

© 1998 by CRC Press LLC

# Preface

---

## Purpose

The purpose of *The Engineering Handbook* is to provide in a single volume a ready reference for the practicing engineer in industry, government, and academia. The book in its comprehensive format is divided into 30 sections which encompass the field of engineering. The goal is to provide the most up-to-date information in the classical fields that comprise mechanical, electrical, civil, chemical, industrial, and aerospace engineering as well as the underlying fields of mathematics and materials. This book should serve the information needs of all professional engineers engaged in the practice of the profession whether in industry, education, or government. The goal of this comprehensive handbook is to replace a myriad of books with one highly informative, well-organized, definitive source of fundamental knowledge.

## Organization

The fundamentals of engineering have evolved to include a wide range of knowledge, substantial empirical data, and a broad range of practice. The focus of the handbook is on the key concepts, models, and equations that enable the engineer to analyze, design, and predict the behavior of complex devices, circuits, instruments, systems, structures, plants, computers, fuels, and the environment. While data and formulae are summarized, the main focus is the provision of the underlying theories and concepts and the appropriate application of these theories to the field of engineering. Thus, the reader will find the key concepts defined, described, and illustrated in order to serve the needs of the engineer over many years.

With equal emphasis placed on materials, structures, mechanics, dynamics, fluids, thermodynamics, fuels and energy, transportation, environmental systems, circuits and systems, computers and instruments, manufacturing, aeronautical and aerospace, and economics and management as well as mathematics, the engineer should encounter a wide range of concepts and considerable depth of exploration of these concepts as they lead to application and design.

The level of conceptual development of each topic is challenging, but tutorial and relatively fundamental. Each of the more than 200 chapters is written to enlighten the expert, refresh the knowledge of the mature engineer, and educate the novice.

The information is organized into 30 major sections. The 30 sections encompass 211 chapters, and the Appendix summarizes the applicable mathematics, symbols, and physical constants. Each section contains one or more historical vignettes that serve to enliven and illuminate the history of the subject of that section. Furthermore, each section is preceded by a photo of a device, circuit, or system that demonstrates an application illustrative of the material in the section.

Each chapter includes three important and useful categories: defining terms, references, and further information. *Defining terms* are key definitions, and the first occurrence of each term defined is indicated in boldface in the text. The definitions of these terms are summarized as a list at the end of each chapter. The *references* provide a list of useful books and articles for follow-up reading. Finally, *further information* provides some general and useful sources of additional

information on the topic.

## Locating Your Topic

Numerous avenues of access to information contained in the handbook are provided. A complete table of contents is presented at the front of the book. In addition, an individual table of contents precedes each of the 30 sections. Finally, each chapter begins with its own table of contents. The reader should look over these tables of contents to become familiar with the structure, organization, and content of the book. For example, see Section III, Dynamics and Vibration, and then Chapter 15, Forced Vibration. This tree-and-branch table of contents enables the reader to move up the tree to locate information on the topic of interest.

Three alphabetical indexes have been compiled to provide multiple means of accessing information: (1) index of contributing authors, (2) index of key equations by title or name, and (3) subject index. The subject index can also be used to locate key definitions. The page on which the definition appears for each key (defining) term is clearly identified in the subject index.

*The Engineering Handbook* is designed to provide answers to most inquiries and direct the inquirer to further sources and references. We hope that this handbook will be referred to often and that informational requirements will be satisfied effectively.

## Acknowledgments

This handbook is testimony to the dedication of the associate editors, the publishers, and my editorial associates. I particularly wish to acknowledge at CRC Press Joel Claypool, Publisher; Kristen Maus, Developmental Editor; and Carol Whitehead, Senior Project Editor. Finally I am indebted to the assistance of Sara Hare, who served as editorial assistant.

**Richard C. Dorf**

Editor-in-Chief

## Editor-in-Chief

---



**Richard C. Dorf**, professor of electrical and computer engineering at the University of California, Davis, teaches graduate and undergraduate courses in electrical engineering in the fields of circuits and control systems. He earned a Ph.D. in electrical engineering from the U.S. Naval Postgraduate School, an M.S. from the University of Colorado, and a B.S. from Clarkson University. Highly concerned with the discipline of engineering and its wide value to social and economic needs, he has written and lectured internationally on the contributions and advances in engineering and their value to society.

Professor Dorf has extensive experience with education and industry and is professionally active in the fields of robotics, automation, electric circuits, and communications. He has served as a visiting professor at the University of Edinburgh, Scotland; the Massachusetts Institute of Technology; Stanford University; and the University of California, Berkeley.

A Fellow of The Institute of Electrical and Electronics Engineers, Dr. Dorf is widely known to the profession for his *Modern Control System*, 7th edition (Addison-Wesley, 1995) and *The International Encyclopedia of Robotics* (Wiley, 1988). Dr. Dorf is also the co-author of *Circuits, Devices and Systems* (with Ralph Smith), 5th edition (Wiley, 1992). Dr. Dorf is the editor-in-chief of the widely recognized *Electrical Engineering Handbook* (CRC, 1993) and the *Handbook of Manufacturing and Automation* (Wiley, 1994).

# Advisory Board

---

**William F. Ames**

School of Mathematics  
Georgia Institute of Technology  
Atlanta, Georgia

**Jack McCormac**

Civil Engineering Department  
Clemson University  
Clemson, South Carolina

**John Steadman**

Electrical Engineering Department  
University of Wyoming  
Laramie, Wyoming

**Frank Kreith**

University of Colorado(Retired)  
Boulder, Colorado

**James F. Shackelford**

Materials Science  
University of California  
Davis, California

**Klaus Timmerhaus**

Chemical Engineering Department  
University of Colorado  
Boulder, Colorado

# Contributors

---

**Ramesh K. Agarwal**

Wichita State University  
Wichita, Kansas

**William F. Ames**

Georgia Institute of Technology  
Atlanta, Georgia

**James E. Amrhein**

Masonry Institute of America  
Los Angeles, California

**Thalia Anagnos**

San Jose State University  
Palo Alto, California

**Tung Au**

Carnegie Mellon University  
Pittsburgh, Pennsylvania

**Terrence W. Baird**

Hewlett-Packard Company  
Boise, Idaho

**Randall F. Barron**

Louisiana Tech University  
Ruston, Louisiana

**Yildiz Bayazitoglu**

Rice University  
Houston, Texas

**Paul A. Beck**

Paul A. Beck & Associates  
Pittsburgh, Pennsylvania

**Richard C. Bennett**

Swenson Process Equipment, Inc.  
Harvey, Illinois

**Pallab Bhattacharya**

University of Michigan  
Ann Arbor, Michigan

**F. Chris Alley**

Clemson University (Emeritus)  
Clemson, South Carolina

**Appiah Amirtharajah**

Georgia Institute of Technology  
Atlanta, Georgia

**Ted L. Anderson**

Structural Reliability Technology  
Boulder, Colorado

**Roger E. A. Arndt**

St. Anthony Falls Laboratory  
University of Minnesota  
Minneapolis, Minnesota

**A. Terry Bahill**

University of Arizona  
Tucson, Arizona

**Norman Balabanian**

University of Florida  
Gainesville, Florida

**Nelson R. Bauld, Jr.**

Clemson University  
Clemson, South Carolina

**Robert G. Beaves**

Robert Morris College  
Coraopolis, Pennsylvania

**R. R. Beck**

U.S. Army Tank Automotive Research Development  
and Engineering Center  
Warren, Michigan

**Jim Bethel**

Purdue University  
West Lafayette, Indiana

**Bharat Bhushan**

Ohio State University  
Columbus, Ohio

**Peter Bilito**

University of California  
Lawrence Livermore National Laboratory  
Livermore, California

**Benjamin S. Blanchard**

Virginia Polytechnic Institute & State University  
Blacksburg, Virginia

**Bruce W. Bomar**

University of Tennessee  
Space Institute  
Tullahoma, Tennessee

**Carol I. Bordas**

Thorp, Reed & Armstrong  
Pittsburgh, Pennsylvania

**Edwin R. Braun**

University of North Carolina  
Charlotte, North Carolina

**Robert Broadwater**

Virginia Polytechnic Institute & State University  
Blacksburg, Virginia

**George R. Buchanan**

Tennessee Technological University  
Cookeville, Tennessee

**Michael Buehrer**

Virginia Polytechnic Institute & State University  
Blacksburg, Virginia

**George Cain**

Georgia Institute of Technology  
Atlanta, Georgia

**William L. Chapman**

Hughes Aircraft Company  
Tucson, Arizona

**Peter M. Chen**

University of Michigan  
Ann Arbor, Michigan

**Kenneth B. Black**

University of Massachusetts  
Amherst, Massachusetts

**Robert F. Boehm**

University of Nevada  
Las Vegas, Nevada

**Robert G. Bonitz**

University of California  
Davis, California

**Charles Borzileri**

University of California  
Lawrence Livermore National Laboratory  
Livermore, California

**Donald E. Breyer**

California State Polytechnic University  
Pomona, California

**William L. Brogan**

University of Nevada  
Las Vegas, Nevada

**R. Ben Buckner**

Surveying Education Consultant  
Johnson City, Tennessee

**Luis-Felipe Cabrera**

IBM Almaden Research Center  
Almaden, California

**Shiao-Hung Chiang**

University of Pittsburgh  
Pittsburgh, Pennsylvania

**James M. Chavez**

Sandia National Laboratories  
Albuquerque, New Mexico

**Wai-Kai Chen**

University of Illinois  
Chicago, Illinois

**Tony M. Cigic**  
University of British Columbia  
Vancouver, British Columbia, Canada

**William J. Cook**  
Iowa State University  
Ames, Iowa

**C. David Cooper**  
University of Central Florida  
Orlando, Florida

**Harold M. Cota**  
California Polytechnic State University  
San Luis Obispo, California

**Dennis J. Cronin**  
Iowa State University  
Ames, Iowa

**John N. Daigle**  
University of Mississippi  
University, Mississippi

**Kevin A. Delin**  
Jet Propulsion Laboratory  
Pasadena, California

**Anca Deliu**  
Georgia Institute of Technology  
Atlanta, Georgia

**Henry Domingos**  
Clarkson University  
Potsdam, New York

**Anil Doradla**  
Virginia Polytechnic Institute & State University  
Blacksburg, Virginia

**Richard C. Dorf**  
University of California  
Davis, California

**William G. Duff**  
Computer Sciences Corporation  
Springfield, Virginia

**Michael D. Ciletti**  
University of Colorado  
Colorado Springs, Colorado

**David H. Cooke**  
Power and Cogeneration Consultant  
2507 Palo Pinto Drive  
Houston, Texas

**William C. Corder**  
CONSOL, Inc.  
Pittsburgh, Pennsylvania

**Leon W. Couch II**  
University of Florida  
Gainesville, Florida

**J. B. Cropley**  
Union Carbide Corporate Fellow (Retired)  
Scott Depot, West Virginia

**Braja M. Das**  
California State University  
Sacramento, California

**Jacques W. Delleur**  
Purdue University  
West Lafayette, Indiana

**Bon A. DeWitt**  
University of Florida  
Gainesville, Florida

**John F. Donovan**  
McDonnell Douglas Corporation  
St. Louis, Missouri

**Deepak Doraiswamy**  
E. I. du Pont de Nemours & Co.  
Wilmington, Delaware

**C. Nelson Dorny**  
University of Pennsylvania  
Philadelphia, Pennsylvania

**Stephen A. Dyer**  
Kansas State University  
Manhattan, Kansas



**William M. Edwards**

Consultant  
Houston, Texas

**Wolter J. Fabrycky**

Virginia Polytechnic Institute & State University  
Blacksburg, Virginia

**John L. Falconer**

University of Colorado  
Boulder, Colorado

**Chang-Xue Feng**

University of Iowa  
Iowa City, Iowa

**Samuel W. Fordyce**

Consultare Technology Group  
Rockville, Maryland

**Robert D. Franceschinis**

Doyen & Associates, Inc.  
Chicago, Illinois

**A. Keith Furr**

Virginia Polytechnic Institute & State University  
(Retired)  
Blacksburg, Virginia

**Pierre Gehlen**

Seattle University  
Seattle, Washington

**Peter Gergely**

Cornell University  
Ithaca, New York

**Walter J. Grantham**

Washington State University  
Pullman, Washington

**Jerry C. Hamann**

University of Wyoming  
Laramie, Wyoming

**Milton E. Harr**

Purdue University

**Mohammed M. El-Wakil**

University of Wisconsin  
Madison, Wisconsin

**James R. Fair**

University of Texas  
Austin, Texas

**Charles Fazzi**

Robert Morris College  
Coraopolis, Pennsylvania

**H. Scott Fogler**

University of Michigan  
Ann Arbor, Michigan

**Wallace T. Fowler**

University of Texas  
Austin, Texas

**Rich Freitas**

IBM Almaden Research Center  
Almaden, California

**Richard S. Gallagher**

R. S. Gallagher and Associates  
Ithaca, New York

**James M. Gere**

Stanford University  
Stanford, California

**Victor W. Goldschmidt**

Purdue University  
West Lafayette, Indiana

**Francis Joseph Hale**

North Carolina State University  
Raleigh, North Carolina

**Simon P. Hanson**

CONSOL, Inc.  
Pittsburgh, Pennsylvania

**Steve J. Harrison**

Queen's University

West Lafayette, Indiana

**Barbara Hauser**

Bay de Noc Community College  
Escanaba, Michigan

**Daxin He**

University of Pittsburgh  
Pittsburgh, Pennsylvania

**Chris Hendrickson**

Carnegie Mellon University  
Pittsburgh, Pennsylvania

**Ronald A. Hess**

University of California  
Davis, California

**Fredrick J. Hill**

University of Arizona  
Tucson, Arizona

**Joe D. Hoffman**

Purdue University  
West Lafayette, Indiana

**Stephen Horan**

New Mexico State University  
Las Cruces, New Mexico

**Paul J. Hurst**

University of California  
Davis, California

**Robert B. Jacko**

Purdue University  
West Lafayette, Indiana

**Raymond G. Jacquot**

University of Wyoming  
Laramie, Wyoming

**Rolf Johansson**

Department of Automatic Control  
Lund Institute of Technology  
Lund, Sweden

Kingston, Ontario, Canada

**Mary Sue Haydt**

Santa Clara University  
Santa Clara, California

**Roger S. Hecklinger**

Roy F. Weston, Inc.  
Valhalla, New York

**John B. Herbich**

Texas A&M University  
College Station, Texas

**Russell C. Hibbeler**

University of Southwestern Louisiana  
Lafayette, Louisiana

**David J. Hills**

University of California  
Davis, California

**David Holten**

University of California  
Lawrence Livermore National Laboratory  
Livermore, California

**T. C. Hsia**

University of California  
Davis, California

**Daniel J. Inman**

Virginia Polytechnic Institute & State University  
Blacksburg, Virginia

**Thomas N. Jackson**

Pennsylvania State University  
University Park, Pennsylvania

**Bill Jamaldin**

University of Louisville  
Louisville, Kentucky

**Steven D. Johnson**

Purdue University  
West Lafayette, Indiana

**S. Casey Jones**  
Georgia Institute of Technology  
Atlanta, Georgia

**Anthony J. Kalinowski**  
Naval Undersea Warfare Center  
New London, Connecticut

**Waldemar Karwowski**  
University of Louisville  
Louisville, Kentucky

**Ralph W. Kiefer**  
University of Wisconsin  
Madison, Wisconsin

**L. Kitis**  
University of Virginia  
Charlottesville, Virginia

**Edward M. Knod, Jr.**  
Western Illinois University  
Macomb, Illinois

**William J. Koros**  
University of Texas  
Austin, Texas

**Allan D. Kraus**  
Naval Postgraduate School  
Pacific Grove, California

**Andrew Kusiak**  
University of Iowa  
Iowa City, Iowa

**Richard T. Lahey, Jr.**  
Rensselaer Polytechnic Institute  
Troy, New York

**Alan O. Lebeck**  
Mechanical Seal Technology, Inc.  
Albuquerque, New Mexico

**Arthur W. Leissa**  
Ohio State University  
Columbus, Ohio

**Reid R. June**  
The Boeing Company  
Bellevue, Washington

**Bruce Karnopp**  
University of Michigan  
Ann Arbor, Michigan

**Tawfik B. Khalil**  
General Motors Corporation  
Transportation Department  
Bloomfield Hills, Michigan

**Bang Mo Kim**  
General Electric Corporate Research and Development  
Schenectady, New York

**Joseph F. Kmec**  
Purdue University  
West Lafayette, Indiana

**Alan A. Kornhauser**  
Virginia Polytechnic Institute & State University  
Blacksburg, Virginia

**William B. Krantz**  
University of Colorado  
Boulder, Colorado

**Frank Kreith**  
University of Colorado (Retired)  
Boulder, Colorado

**Benjamin G. Kyle**  
Kansas State University  
Manhattan, Kansas

**Lee S. Langston**  
University of Connecticut  
Storrs, Connecticut

**Robert E. Lee**  
Pennsylvania Power & Light  
Allentown, Pennsylvania

**John Leonard II**  
Georgia Institute of Technology  
Atlanta, Georgia

**John B. Ligon**

Michigan Technological University  
Houghton, Michigan

**K. H. Lin**

Oak Ridge National Laboratory  
Knoxville, Tennessee

**Earl Livingston**

Babcock and Wilcox Company, Retired  
Barberton, Ohio

**Thomas E. Marlin**

McMaster University  
Hamilton, Ontario, Canada

**R. Bruce Martin**

University of California  
Davis, California

**J. Michael McCarthy**

University of California  
Irvine, California

**Alan T. McDonald**

Purdue University  
West Lafayette, Indiana

**Sue McNeil**

Carnegie Mellon University  
Pittsburgh, Pennsylvania

**J. L. Meriam**

University of California (Retired)  
Santa Barbara, California

**Karsten Meyer-Waarden**

University of Karlsruhe  
Karlsruhe, Germany

**David Mindell**

Massachusetts Institute of Technology  
Cambridge, Massachusetts

**Emil Moroz**

University of Texas  
El Paso, Texas

**Thomas M. Lillesand**

University of Wisconsin  
Madison, Wisconsin

**Noam Lior**

University of Pennsylvania  
Philadelphia, Pennsylvania

**Gregory L. Long**

University of California  
Irvine, California

**Harold E. Marshall**

National Institute of Standards and Technology  
Gaithersburg, Maryland

**Eric F. Matthys**

University of California  
Santa Barbara, California

**Jack McCormac**

Clemson University  
Clemson, South Carolina

**Ross E. McKinney**

Consulting Engineer  
Lawrence, Kansas

**Daniel A. Mendelsohn**

Ohio State University  
Columbus, Ohio

**Michael D. Meyer**

Georgia Institute of Technology  
Atlanta, Georgia

**Scott L. Miller**

University of Florida  
Gainesville, Florida

**Jan C. Monk**

National Aeronautics and Space Administration  
Marshall Space Flight Center, Alabama

**Samiha Mourad**

Santa Clara University  
Santa Clara, California

**Safwat M. A. Moustafa**

California Polytechnic University  
San Luis Obispo, California

**Bruce R. Munson**

Iowa State University  
Ames, Iowa

**Paul Neudorfer**

Seattle University  
Seattle, Washington

**M. M. Ohadi**

University of Maryland  
College Park, Maryland

**James Y. Oldshue**

Oldshue Technologies International, Inc.  
Fairport, New York

**Hasan Orbey**

University of Delaware  
Newark, Delaware

**Bulent A. Ovunc**

University of Southwestern Louisiana  
Lafayette, Louisiana

**Joseph C. Palais**

Arizona State University  
Tempe, Arizona

**Gordon R. Pennock**

Purdue University  
West Lafayette, Indiana

**Bruce E. Poling**

University of Toledo  
Toledo, Ohio

**Alexander D. Poularikas**

University of Alabama  
Huntsville, Alabama

**Kaushik S. Rajashekara**

Delphi Energy & Engine Management Systems  
Indianapolis, Indiana

**Rias Muhamed**

Virginia Polytechnic Institute & State University  
Blacksburg, Virginia

**Edward G. Nawy**

Rutgers University  
East Brunswick, New Jersey

**Norman S. Nise**

California State Polytechnic University  
Pomona, California

**Vojin G. Oklobdzija**

University of California  
Davis, California

**George Opdyke, Jr.**

Dykewood Enterprises  
Stratford, Connecticut

**Terry P. Orlando**

Massachusetts Institute of Technology  
Cambridge, Massachusetts

**William Owens**

Johnson Yokogawa Corporation  
Newnan, Georgia

**Howard S. Peavy**

University of Idaho  
Moscow, Idaho

**Walter D. Pilkey**

University of Virginia  
Charlottesville, Virginia

**H. Vincent Poor**

Princeton University  
Princeton, New Jersey

**Mansour Rahimi**

University of Southern California  
Los Angeles, California

**Rama Ramakumar**

Oklahoma State University  
Stillwater, Oklahoma

**Theodore S. Rappaport**

Virginia Polytechnic Institute & State University  
Blacksburg, Virginia

**Bahram Ravani**

University of California  
Davis, California

**M. K. Ravindra**

EQE International  
Irvine, California

**Elizabeth M. Richards**

Sandia National Laboratories  
Albuquerque, New Mexico

**E. V. Richardson**

Ayres Associates  
Fort Collins, Colorado

**Albert J. Rosa**

University of Denver  
Denver, Colorado

**Andrew P. Sage**

George Mason University  
Fairfax, Virginia

**Stanley I. Sandler**

University of Delaware  
Newark, Delaware

**Udaya B. Sathuvalli**

Rice University  
Houston, Texas

**Boyd D. Schimel**

Washington State University  
Pullman, Washington

**Paul Schonfeld**

University of Maryland  
College Park, Maryland

**Robert G. Sexsmith**

University of British Columbia  
Vancouver, British Columbia, Canada

**Muhammad H. Rashid**

Purdue University  
Fort Wayne, Indiana

**Francis H. Raven**

University of Notre Dame  
South Bend, Indiana

**Timothy A. Reinhold**

Clemson University  
Clemson, South Carolina

**John L. Richards**

University of Pittsburgh  
Pittsburgh, Pennsylvania

**Subhash H. Risbud**

University of California  
Davis, California

**Rosalie T. Ruegg**

National Institute of Standards and Technology  
Gaithersburg, Maryland

**Richard S. Sandige**

University of Wyoming  
Laramie, Wyoming

**Albert Sargent**

Arkansas Power & Light  
Hot Springs, Arkansas

**Rudolph J. Scavuzzo**

University of Akron  
Akron, Ohio

**Richard J. Schonberger**

University of Washington  
Seattle, Washington

**William T. Segui**

University of Memphis  
Memphis, Tennessee

**James F. Shackelford**

University of California  
Davis, California

**Andrew P. Shapiro**

General Electric Corporate Research and Development  
Schenectady, New York

**George Shibayama**

Doyen & Associates, Inc.  
Chicago, Illinois

**J. G. Shipp**

EQE International  
Irvine, California

**Paul W. Shuldiner**

University of Massachusetts  
Amherst, Massachusetts

**R. Paul Singh**

University of California  
Davis, California

**Kumares C. Sinha**

Purdue University  
West Lafayette, Indiana

**L. Montgomery Smith**

University of Tennessee  
Space Institute  
Tullahoma, Tennessee

**Sidney Soclof**

California State University  
Los Angeles, California

**Richard E. Sonntag**

University of Michigan  
Ann Arbor, Michigan

**E. Keith Stanek**

University of Missouri  
Rolla, Missouri

**Raymond T. Stefani**

California State University  
Long Beach, California

**P. K. Subramanyan**

Glacier Clevite Heavywall Bearings

**S. A. Sherif**

University of Florida  
Gainesville, Florida

**Yung C. Shin**

Purdue University  
West Lafayette, Indiana

**Walter D. Short**

National Renewable Energy Laboratory  
Golden, Colorado

**Ben L. Sill**

Clemson University  
Clemson, South Carolina

**Vijay P. Singh**

Louisiana State University  
Baton Rouge, Louisiana

**Shivaji Sircar**

Air Products and Chemicals, Inc.  
Allentown, Pennsylvania

**Rosemary L. Smith**

University of California  
Davis, California

**Michael A. Soderstrand**

University of California  
Davis, California

**Cary R. Spitzer**

AvioniCon, Inc.  
Williamsburg, Virginia

**John Steadman**

University of Wyoming  
Laramie, Wyoming

**Matthew P. Stephens**

Purdue University  
West Lafayette, Indiana

**James A. Svoboda**

Clarkson University

McConnelsville, Ohio

**Larry W. Swanson**  
Heat Transfer Research Institute  
College Station, Texas

**Hans J. Thamhain**  
Bentley College  
Waltham, Massachusetts

**Klaus Timmerhaus**  
University of Colorado  
Boulder, Colorado

**Matt Traini**  
University of California  
Lawrence Livermore National Laboratory  
Livermore, California

**Anne VanArsdall**  
Sandia National Laboratories  
Albuquerque, New Mexico

**Boudewijn H. W. van Gelder**  
Purdue University  
West Lafayette, Indiana

**Dennis M. Volpano**  
Naval Postgraduate School  
Monterey, California

**Curtis J. Wahlberg**  
Purdue University  
West Lafayette, Indiana

**David Wallenstein**  
Woodward-Clyde Consultants  
Oakland, California

**Barry Wilkinson**  
University of North Carolina  
Charlotte, North Carolina

**David M. Woodall**  
University of Idaho  
Moscow, Idaho

**William W. Wu**

Potsdam, New York

**Andrew Swift**  
University of Texas  
El Paso, Texas

**James F. Thompson**  
Thompson Professional Group, Inc.  
Houston, Texas

**Y. L. Tong**  
Georgia Institute of Technology  
Atlanta, Georgia

**J. Paul Tullis**  
Utah State University  
Logan, Utah

**Vincent Van Brunt**  
University of South Carolina  
Columbia, South Carolina

**Thomas L. Vincent**  
University of Arizona  
Tucson, Arizona

**Wolf W. von Maltzahn**  
The Whitaker Foundation  
Rosslyn, Virginia

**David R. B. Walker**  
University of Texas  
Austin, Texas

**Pao-lien Wang**  
University of North Carolina  
Charlotte, North Carolina

**William L. Wood**  
Purdue University  
West Lafayette, Indiana

**David G. Woods**  
University of Texas  
Austin, Texas

**Loren W. Zachary**



Consultare Technology Group  
Rockville, Maryland

**Ashraf A. Zeid**  
Army High Performance Computing Research Center  
and Computer Sciences Corporation  
Warren, Michigan

Iowa State University  
Ames, Iowa

**Rodger E. Ziemer**  
University of Colorado  
Colorado Springs, Colorado

# Contents

---

## SECTION I Statics

---

### **Introduction** *Russell C. Hibbeler*

#### **1 Force-System Resultants and Equilibrium** *R. C. Hibbeler*

Force-System · Resultants · Equilibrium

#### **2 Centroids and Distributed Forces** *W. D. Pilkey and L. Kitis*

Centroid of a Plane Area · Centroid of a Volume Surface · Forces · Line Forces · Calculation of Surface Area and Volume of a Body with Rotational Symmetry · Determination of Centroids

#### **3 Moments of Inertia** *J. L. Meriam*

Area Moments of Inertia · Mass Moments of Inertia

## SECTION II Mechanics of Materials

---

### **Introduction** *Ted L. Anderson*

#### **4 Reactions** *T. Anagnos*

Types of Supports · Actual versus Idealized Support Conditions · Static Determinacy and Indeterminacy · Computation of Reactions

#### **5 Bending Stresses in Beams** *J. M. Gere*

Longitudinal Strains in Beams · Normal Stresses in Beams (Linearly Elastic Materials)

#### **6 Shear Stresses in Beams** *J. M. Gere*

Shear Stresses in Rectangular Beams · Shear Stresses in Circular Beams · Shear Stresses in the Webs of Beams with Flanges

#### **7 Shear and Moment Diagrams** *G. R. Buchanan*

Sign Convention · Shear and Moment Diagrams · Shear and Moment Equations

#### **8 Columns** *L. W. Zachary and J. B. Ligon*

Fundamentals · Examples · Other Forms of Instability

#### **9 Pressure Vessels** *E. Livingston and R. J. Scavuzzo*

Design Criteria · Design Formulas · Opening Reinforcement

#### **10 Axial Loads and Torsion** *N. R. Bauld, Jr.*

Axially Loaded Bars · Torsion

#### **11 Fracture Mechanics** *T. L. Anderson*

Fundamental Concepts · The Energy Criterion · The Stress Intensity Approach · Time-Dependent Crack Growth and Damage Tolerance · Effect of Material Properties on Fracture

## SECTION III Dynamics and Vibrations

---

### **Introduction** *Bruce Karnopp*

#### **12 Dynamics of Particles: Kinematics and Kinetics** *B. Karnopp*

Dynamics of Particles · Newton's Second Law · Moment of Momentum Relations · Momentum · Integrals of Newton's Second Law · Work-Energy Integral of Newton's Second Law · Conclusion

#### **13 Dynamics of Rigid Bodies: Kinematics and Kinetics** *A. A. Zeid and R. R. Beck*

Kinematics of Rigid Bodies · Kinetics of Rigid Bodies

#### **14 Free Vibration, Natural Frequencies, and Mode Shapes** *D. A. Mendelsohn*

Basic Principles · Single-Degree-of-Freedom Systems · Multiple-Degree-of-Freedom Systems · Continuous Systems (Infinite DOF)

#### **15 Forced Vibrations** *A. W. Leissa*

Single-Degree-of-Freedom Systems · Multiple-Degree-of-Freedom Systems

#### **16 Lumped versus Distributed Parameter Systems** *B. A. Ovunc*

Procedure of Analysis · Continuous Mass Matrix Method · Consistent Mass Matrix Method · Lumped Mass Matrix Method · Free Vibration of Frames · Forced Vibration · Practical Applications · Structures without Additional Effects · Structures with Additional Effects

#### **17 Applications of Structural and Dynamic Principles** *A. J. Kalinowski*

Base Configuration Loaded Applications · Structural Configuration Loaded Applications · Additional Information

#### **18 Computer Simulation and Nomographs** *D. J. Inman*

Nomograph Fundamentals · Models for Numerical Simulation · Numerical Integration · Vibration Response by Computer Simulation · Commercial Software for Simulation

#### **19 Test Equipment and Measuring Instruments** *T. W. Baird*

Vibration and Shock Test Machines · Transducers and Signal Conditioners · Digital Instrumentation and Computer Control

## SECTION IV Kinematics and Mechanisms

---

### **Introduction** *Bahram Ravani*

#### **20 Linkages and Cams** *J. M. McCarthy and G. L. Long*

Linkages · Spatial Linkages · Displacement Analysis · Cam Design · Classification of Cams and Followers · Displacement Diagrams

#### **21 Tribology: Friction, Wear, and Lubrication** *B. Bhushan*

History of Tribology and Its Significance to Industry · Origins and Significance of Micro/nanotribology · Friction · Wear · Lubrication · Micro/nanotribology

#### **22 Machine Elements** *G. R. Pennock*

Threaded Fasteners · Clutches and Brakes

#### **23 Crankshaft Journal Bearings** *P. K. Subramanyan*

Role of the Journal Bearings in the Internal Combustion Engine · Construction of Modern Journal Bearings · The Function of the Different Material Layers in Crankshaft Journal Bearings · The Bearing Materials · Basics of Hydrodynamic Journal Bearing Theory · The Bearing Assembly · The Design Aspects of Journal Bearings · Derivations of the Reynolds and Harrison Equations for Oil Film Pressure

## **24 Fluid Sealing in Machines, Mechanical Devices, and Apparatus** *A. O. Lebeck*

Fundamentals of Sealing · Static Seals · Dynamic Seals · Gasket Practice · O-Ring Practice · Mechanical Face Seal Practice

## **SECTION V Structures**

---

### **Introduction** *Jack McCormac*

### **25 Loads** *P. Gergely*

Dead Loads · Live Loads · Impact Loads · Snow Loads

### **26 Wind Effects** *T. A. Reinhold and B. L. Sill*

Wind Climate · Local Wind Exposure · Mean Wind Speed Profile · Turbulence · Pressure Coefficients and Load Factors

### **27 Earthquake Effects** *M. K. Ravindra and J. G. Shipp*

Why Do Earthquakes Occur? · Characteristics of Earthquakes · Damage Mechanisms · Seismic Hazard Analysis · Earthquake Damage Surveys · Earthquake-Resistant Design

### **28 Structural Analysis** *R. G. Sexsmith and T. M. Cigic*

Beams · Trusses · Frames · Computer-Aided Analysis

### **29 Structural Steel** *W. T. Segui*

Members · Connections · Composite Members · Computer Applications

### **30 Concrete** *E. G. Nawy*

Structural Concrete · Flexural Design of Reinforced Concrete Members · Shear and Torsion Design of Reinforced Concrete Members · Prestressed Concrete · Serviceability Checks · Computer Applications for Concrete Structures

### **31 Timber** *D. E. Breyer*

Durability of Wood · Wood Products · Member Design · Connections · Lateral Force Design

### **32 Masonry Design** *J. E. Amrhein*

Basis of Design · Masonry Materials · Masonry Units · Concrete Masonry · Mortar · Grout · Unreinforced Masonry · Strength of Masonry · Design of Reinforced Masonry Members · Design of Structural Members-Strength Design

## **SECTION VI Fluid Mechanics**

---

### **Introduction** *Frank Kreith*

### **33 Incompressible Fluids** *A. T. Mc Donald*

Fundamentals of Incompressible Fluid Flow · Fluids without Relative Motion · Basic Equations in Integral Form for Control Volumes · Differential Analysis of Fluid Motion · Incompressible Inviscid Flow · Dimensional Analysis and Similitude · Internal Incompressible Viscous Flow · External Incompressible Viscous Flow

### **34 Compressible Fluids** *J. D. Hoffman*

General Features · Basic Equations · Steady Quasi-One-Dimensional Flow · Equations of State and Thermodynamics · Stagnation Properties · Isentropic Flow · Nozzles · Shock Waves · Friction and Heat Transfer

### **35 The Rheology of Non-Newtonian Fluids** *D. Doraiswamy*

Kinematics, Flow Classification, and Material Functions · Fluids · Constitutive Equations · Some Useful Correlations for Material Functions

### **36 Airfoils/Wings** *B. R. Munson and D. J. Cronin*

Nomenclature · Airfoil Shapes · Lift and Drag Characteristics for Airfoils · Lift and Drag of Wings  
[37 Boundary Layers](#) *E. R. Braun and P.-l. Wang*  
 Theoretical Boundary Layers · Reynolds Similarity in Test Data · Friction in Pipes · Noncircular Channel · Example Solutions  
[38 Valves](#) *J. P. Tullis*  
 Control Valves · Air Valves · Check Valves  
[39 Pumps and Fans](#) *R. F. Boehm*  
 Pumps · Fans  
[40 Two-Phase Flow](#) *R. T. Lahey, Jr.*  
 Notation · Conservation Equations · Closure · Two-Phase Instabilities · Conclusion  
[41 Basic Mixing Principles for Various Types of Fluid Mixing Applications](#) *J. Y. Oldshue*  
 Scaleup/Scaledown · Effect of the Circulation Time Spectrum and the Spectrum of Shear Rates on Ten Different Mixing Technologies · Computational Fluid Dynamics  
[42 Fluid Measurements](#) *S. A. Sherif*  
 Fundamental Principles · Basic Equations

## **SECTION VII Thermodynamics and Heat Transfer**

---

### **Introduction** *Frank Kreith*

[43 The First Law of Thermodynamics](#) *R. E. Sonntag*  
 System Analysis · Control Volume Analysis  
[44 Second Law of Thermodynamics and Entropy](#) *N. Lior*  
 Reversibility · Entropy · The Second Law for Bulk Flow · Applications  
[45 The Thermodynamics of Solutions](#) *S. I. Sandler and H. Orbey*  
 Fundamentals · Applications  
[46 Thermodynamics of Surfaces](#) *W. B. Krantz*  
 Basic Concepts · First Law of Thermodynamics · Effects of Curved Interfaces · Adsorption at Interfaces · Wettability and Adhesion  
[47 Phase Equilibrium](#) *B. G. Kyle*  
 Pure-Component Phase Equilibrium · Phase Equilibrium in Mixtures · Perspective  
[48 Thermodynamic Cycles](#) *W. J. Cook*  
 Power Cycles · Refrigeration Cycles  
[49 Heat Transfer](#) *Y. Bayazitoglu and U. B. Sathuvalli*  
 Conduction · Convection · Radiation · Phase Change  
[50 Heat Exchangers](#) *M. M. Ohadi*  
 Heat Exchanger Types · Shell-and-Tube Heat Exchangers · Compact Heat Exchangers · Design of Heat Exchangers  
[51 Combustion](#) *R. S. Hecklinger*  
 Fundamentals of Combustion · Combustion Calculations  
[52 Air Conditioning](#) *V. W. Goldschmidt and C. J. Wahlberg*  
 Historical Sketch · Comfort · Air Conditioning Process · Representative Cycles  
[53 Refrigeration and Cryogenics](#) *R. F. Barron*  
 Desiccant Cooling · Heat Pumps · Cryogenics  
[54 Heat Transfer to Non-Newtonian Fluids](#) *E. F. Matthys*  
 The Fluids · Friction · Heat Transfer · Instrumentation and Equipment  
[55 Heat Pipes](#) *L. W. Swanson*  
 Heat Pipe Container, Working Fluid, and Wick Structures · Heat Transfer Limitations · Effective Thermal Conductivity and Heat Pipe Temperature Difference · Application of Heat Pipes

## SECTION VIII Separation Processes

---

### Introduction *Klaus Timmerhaus*

#### 56 Distillation *J. R. Fair*

Separation Specification · Required Basic Data · Index of Separation Difficulty · Required Stages · Column Dimensions · Column Auxiliaries · Batch Distillation

#### 57 Absorption and Stripping *W. M. Edwards and J. R. Fair*

Absorber-Stripper Systems · Absorber-Stripper Design Diagrams · Key Design Assumptions · Physical Data Requirements · Absorber and Stripper Design Equations · Absorption and Stripping Efficiency

#### 58 Extraction *V. Van Brunt*

Representative Extraction Processes · Solvent Characteristics and Solvent Screening · Extraction Equilibria · Extraction Staging and Equipment

#### 59 Adsorption *S. Sircar*

Adsorbent Materials · Adsorption Equilibria · Heat of Adsorption · Thermodynamic Selectivity of Adsorption · Adsorption Kinetics · Adsorption Column Dynamics · Adsorptive Separation Processes and Design

#### 60 Crystallization and Evaporation *R. C. Bennett*

Methods of Creating Supersaturation · Reasons for the Use of Crystallization · Solubility Relations · Product Characteristics · Impurities Influencing the Product · Kinds of Crystallization Processes · Calculation of Yield in a Crystallization Process · Mathematical Models of Continuous Crystallization · Equipment Designs · Evaporation

#### 61 Membrane Separation *D. G. Woods, D. R. B. Walker, and W. J. Koros*

Dialysis · Reverse Osmosis · Gas and Vapor Separations · Asymmetric Membranes · Membrane Stability and Fouling · Module Design Considerations

#### 62 Fluid-Particle Separation *S.-H. Chiang and D. He*

Equipment · Fundamental Concept · Design Principles · Economics

#### 63 Other Separation Processes *W. C. Corder and S. P. Hanson*

Sublimation · Diffusional Separations · Adsorptive Bubble Separation · Dielectrophoresis · Electrodialysis

## SECTION IX Fuels and Energy Conversion

---

### Introduction *Frank Kreith*

#### 64 Fuels *S. M. A. Moustafa*

Coal · Oil · Natural Gas · Important Products of Crude Oil Refining

#### 65 Solar Power Systems *J. M. Chavez, E. M. Richards, and A. Van Arsdall*

Solar Thermal Systems · Photovoltaic Systems · Biomass Systems

#### 66 Internal Combustion Engines *A. A. Kornhauser*

Basics of Operation · Engine Classifications · Spark Ignition Engines · Compression Ignition Engines · Gas Exchange Systems · Design Details · Design and Performance Data for Typical Engines

#### 67 Gas Turbines *L. S. Langston and G. Opdyke, Jr.*

Gas Turbine Usage · Gas Turbine Cycles · Gas Turbine Components

#### 68 Nuclear Power Systems *D. M. Woodall*

Nuclear Power Applications · Nuclear Power Fundamentals · Economics of Nuclear Power Systems

#### 69 Power Plants *M. M. El-Wakil*

The Rankine Cycle · The Turbine · The Condenser · The Condenser Cooling System · The Feedwater

System · The Steam Generator · Cycle and Plant Efficiencies and Heat Rates

**70 Wind Turbines** *A. Swift and E. Moroz*

Fundamentals · Power Regulation and Control · Energy Capture Estimation · Stand-Alone Applications · Cost of Energy Calculations · Environmental and Social Cost Issues · Summary

**71 Hydraulic Turbines** *R. E. A. Arndt*

General Description · Principles of Operation · Factors Involved in Selecting a Turbine

**72 Steam Turbines** *G. Shibayama and R. D. Franceschinis*

Types of Steam Turbines · Impulse versus Reaction Turbines · Thermodynamics · Stop and Control Valves · Water Induction Protection · Generators · Turbine Generator Auxiliaries

**73 Cogeneration** *D. H. Cooke*

Cogeneration Fundamentals · Examples of Cogeneration

**74 Electric Machines** *E. K. Stanek*

Induction Machines · Synchronous Machines · DC Machines

---

## **SECTION X Kinetics and Reaction Engineering**

---

**Introduction** *John L. Falconer*

**75 Reaction Kinetics** *K. H. Lin*

Fundamentals · Analysis of Kinetic Data

**76 Chemical Reaction Engineering** *H. S. Fogler*

The Algorithm · Pressure Drop in Reactors · Multiple Reactions · Heat Effects · Summary

**77 The Scaleup of Chemical Reaction Systems from Laboratory to Plant** *J. B. Cropley*

General Considerations in the Rational Design of Chemical Reactors · Protocol for the Rational Design of Chemical Reactors

---

## **SECTION XI Geotechnical**

---

**Introduction** *Milton E. Harr*

**78 Soil Mechanics** *B. M. Das*

Weight-Volume Relationship Hydraulic Conductivity Effective Stress Consolidation Shear Strength

---

## **SECTION XII Transportation**

---

**Introduction** *Kumares C. Sinha*

**79 Transportation Planning** *M. D. Meyer*

Basic Framework of Transportation Planning · Transportation Modeling

**80 Design of Transportation Facilities** *J. Leonard II and M. D. Meyer*

Components of the Project Development Process · Basic Concepts of Project Design · Intermodal Transportation Terminals or Transfer Facilities · Advanced Technology Projects

**81 Operations and Environmental Impacts** *P. W. Shuldiner and K. B. Black*

Fundamental Equations · Flow, Speed, and Density Relationships · Traffic Measurements · Level of Service (LOS) · Highway Capacity · Intersection Capacity · Traffic Control Devices · Stop Sign Warrants · Traffic Signal Warrants · Air Quality Effects

**82 Transportation Systems** *P. Schonfeld*

Transportation System Components · Evaluation Measures · Air Transportation · Railroad Transportation ·

Highway Transportation · Water Transportation · Public Transportation

**83 Safety Analysis** *T. B. Khalil*

Mathematical Models · Summary

## **SECTION XIII Ocean and Coastal Engineering**

---

**Introduction** *William L. Wood*

**84 Shallow Water and Deep Water Engineering** *J. B. Herbich*

Wave Phenomena · Sediment Processes · Beach Profile · Longshore Sediment Transport · Coastal Structures · Navigational Channels · Marine Foundations · Oil Spills · Offshore Structures

## **SECTION XIV Environmental Systems and Management**

---

**Introduction** *Robert B. Jacko*

**85 Drinking Water Treatment** *A. Amirtharajah and S. C. Jones*

Water Quality · Drinking Water Regulations · Water Treatment Processes

**86 Air Pollution** *F. C. Alley and C. D. Cooper*

Control of Particulate Matter · Control of Gaseous Pollutants

**87 Wastewater Treatment and Disposal** *H. S. Peavy*

Wastewater Characteristics · Terminology in Wastewater Treatment · Sludge Advanced Wastewater Treatment · Wastewater Disposal and Reuse · Future of Wastewater Treatment and Reuse

**88 Solid Wastes** *R. E. McKinney*

Regulations · Characteristics · Generation · Collection · Transfer and Transport · Processing and Resource Recovery (Recycling) · Final Disposal

**89 Hazardous Waste Management** *H. M. Cota and D. Wallenstein*

Regulatory Overview · Definition of Hazardous · Waste Management of Hazardous Wastes · Hazardous Waste Treatment · Infectious Waste Management · Radioactive Waste Management · Mixed Wastes · Corrective Action Waste Minimization · Right-to-Know Laws · Computer Usage in Hazardous Waste Management

**90 Soil Remediation** *B. M. Kim and A. P. Shapiro*

Regulations · Treatment · Technologies · Landfilling and Containment · Soil Vapor Extraction · Thermal Treatments · Stabilization · Bioremediation · Soil Washing · Emerging Technologies

## **SECTION XV Water Resources Engineering**

---

**Introduction** *Jacques W. Delleur*

**91 Hydraulics** *B. Hauser*

Flow Characteristics · Equation of Continuity · Pressure Characteristics · Effects of Pressure<sup>3</sup>/4Dynamic Systems · Pressure Loss · Open Channel Flow · Flow Measurement · Centrifugal Pump

**92 Hydrology** *V. P. Singh*

Classification of Hydrology · Hydrologic Cycle · Laws of Science · Approaches to Hydrologic Problems · Tools for Hydrologic Analyses · Components of Hydrology Cycle<sup>3</sup>/4Deterministic Hydrology · Statistical Hydrology · Hydrologic Design

**93 Sedimentation** *E. V. Richardson*

Fluvial Geomorphology · Sediment Properties · Beginning of Motion · Sediment Yield · Bed Forms · Sediment Transport · Reservoir Sedimentation



## SECTION XVI Linear Systems and Models

---

### Introduction *Rodger E. Ziemer*

#### 94 Transfer Functions and Laplace Transforms *C. N. Dorny*

Transfer Functions · The Laplace Transformation · Transform Properties · Transformation and Solution of a System Equation

#### 95 Block Diagrams *A. P. Sage*

Elements of the Block Diagram · Block Diagram Reduction · Summary

#### 96 Signal Flow Analysis *A. D. Kraus*

The Signal Flow Graph · Transmission Gain · Signal Flow Graph Algebra · The Mason Gain Rule

#### 97 Linear State-Space Models *B. D. Schimel and W. J. Grantham*

State-Space Models · Linearization · Linear System Representations · Transforming System Representations

#### 98 Frequency Response *P. Neudorfer and P. Gehlen*

Frequency Response Plotting · A Comparison of Methods

#### 99 Convolution Integral *R. E. Ziemer*

Fundamentals · Properties of the Convolution Operation · Applications of the Convolution Integral · Two-Dimensional Convolution · Time-Varying System Analysis

#### 100 Stability Analysis *R. T. Stefani*

Response Components · Internal (Asymptotic) and External (BIBO) Stability · Unstable and Marginally Stable Responses · Structural Integrity

#### 101 $z$ Transform and Digital Systems *R. Johansson*

The  $z$  Transform · Digital Systems and Discretized Data · The Transfer Function · Digital Systems Described by Difference Equations (ARMAX Models) · Prediction and Reconstruction · The Kalman Filter

## SECTION XVII Circuits

---

### Introduction *Wai-Kai Chen*

*University of Illinois, Chicago*

#### 102 Passive Components *H. Domingos*

Resistors · Capacitors · Inductors

#### 103 RL, RC, and RLC Circuits *M. D. Ciletti*

RL Circuits · RC Circuits · RLC Circuits

#### 104 Node Equations and Mesh Equations *J. A. Svoboda*

Node Equations · Mesh Equations

#### 105 Sinusoidal Excitation and Phasors *M. H. Rashid*

Sinusoidal Source · Phasor · Passive Circuit Elements in Phasors Domain

#### 106 Three-Phase Circuits *N. Balabanian*

Relationships between Voltages and Currents · Line Voltages · Power Relationship · Balanced Source and Balanced Load · Other Types of Interconnections

#### 107 Filters (Passive) *A. J. Rosa*

Fundamentals · Applications

#### 108 Power Distribution *R. Broadwater, A. Sargent, and R. E. Lee*

Equipment · System Divisions and Types · Electrical Analysis, Planning, and Design · System Control · Operations

## 109 Grounding and Shielding *W. G. Duff*

Grounding · Shielding

## SECTION XVIII Electronics

---

### Introduction *Thomas N. Jackson*

#### 110 Operational Amplifiers *P. J. Hurst*

The Ideal Op Amp · Feedback Circuit Analysis · Input and Output Impedances · Practical Limitations and Considerations

#### 111 Active RC Filters *M. A. Soderstrand*

History of Active Filters · Active Filter Design Techniques · Filter Specifications and Approximations · Filter Design

#### 112 Diodes and Transistors *S. Soclof*

Semiconductors · Bipolar Junction Transistors · Junction Field-Effect Transistors · Metal-Oxide Silicon Field-Effect Transistors

#### 113 Analog Integrated Circuits *S. Soclof*

Operational Amplifiers · Voltage Comparators · Voltage Regulators · Power Amplifiers · Wide-Bandwidth (Video) Amplifiers · Modulators, Demodulators, and Phase Detectors · Voltage-Controlled Oscillators · Waveform Generators · Phase-Locked Loops · Digital-to-Analog and Analog-to-Digital Converters · Radio-Frequency Amplifiers · Integrated Circuit Transducers

#### 114 Optoelectronic Devices *P. Bhattacharya*

Light-Emitting Diodes · Lasers · Photodetectors · Conclusion

#### 115 Power Electronics *K. S. Rajashekara*

Power Semiconductor Devices · Power Conversion

#### 116 A/D and D/A Converters *J. C. Hamann*

The Fundamentals of D/A Converters · The Fundamentals of A/D Converters

#### 117 Superconductivity *K. A. Delin and T. P. Orlando*

Introduction · General Electromagnetic Properties · Superconducting Electronics · Types of Superconductors

## SECTION XIX Digital Systems

---

### Introduction *Vojin G. Oklobdzija*

#### 118 Logic Devices *R. S. Sandige*

AND Gates · OR Gates · INVERTER Circuit · NAND Gates · NOR Gates

#### 119 Counters and State Machines (Sequencers) *B. Wilkinson*

Binary Counters · Arbitrary Code Counters · Counter Design · State Machines · State Diagrams · State Diagrams Using Transition Expressions

#### 120 Microprocessors and Microcontrollers *F. J. Hill*

Digital Hardware Systems · Processors · Assembly Language · Some Real Microprocessors and Microcontrollers

#### 121 Memory Systems *R. S. Sandige*

CPU, Memory, and I/O Interface Connections · CPU Memory Systems Overview · Common and Separate I/O Data Buses · Single-Port RAM Devices · Additional Types of Memory Devices · Design Examples

#### 122 Computer-Aided Design and Simulation *M. D. Ciletti*

Design Flow · Schematic Entry · Hardware Description Languages · Trade-offs between HDLs and Schematic Entry · HDLs and Synthesis · Transistor-Level Design and Simulation

### **123 Logic Analyzers** *S. Mourad and M. S. Haydt*

Nature of Digital Signals · Signal Sampling · Timing Analysis · State Analysis · Components of a Logic Analyzer · Advanced Features of Logic Analyzers · Applications of Logic Analyzers

## **SECTION XX Communications and Signal Processing**

---

### **Introduction** *H. Vincent Poor*

#### **124 Transforms and Fast Algorithms** *A. D. Poularikas*

Fourier Transforms · Walsh-Hadamard Transform

#### **125 Digital Filters** *B. W. Bomar and L. M. Smith*

Finite Impulse Response Filter Design · Infinite Impulse Response Filter Design · Digital Filter Implementation

#### **126 Modulation and Detection** *H. V. Poor*

Analog Modulation and Detection · Digital Modulation and Detection · Further Issues

#### **127 Coding** *S. L. Miller and L. W. Couch II*

Block Codes · Convolutional Codes · Trellis-Coded Modulation

#### **128 Computer Communication Networks** *J. N. Daigle*

General Networking Concepts · Computer Communication Network Architecture · Local-Area Networks and Internets · Some Additional Recent Developments

#### **129 Satellites and Aerospace** *S. W. Fordyce and W. W. Wu*

Communications Satellite Services and Frequency Allocation · Information Transfer and Link Margins<sup>3/4</sup>Ground to Space (Up-Link) · Communication Satellite Orbits · Launch Vehicles · Spacecraft Design · Propagation · Earth Stations

#### **130 Mobile and Cellular Radio Communications** *T. S. Rappaport, R. Muhamed, M. Buehrer, and A. Doradla*

Paging Systems · Cordless Telephone Systems · Cellular Telephone Systems · Personal Communications System (PCS) · The Cellular Concept and System Fundamentals · System Capacity and Performance of Cellular Systems · Mobile Radio Systems Around the World

#### **131 Optical Communications** *J. C. Palais*

Optical Communications Systems · Topologies · Fibers · Other Components · Signal Quality

## **SECTION XXI Computers**

---

### **Introduction** *Vojin G. Oklobdzija*

#### **132 Computer Organization: Architecture** *V. G. Oklobdzija*

Instruction Set · RISC Architecture

#### **133 Operating Systems** *L.-F. Cabrera*

Typical Services · General Flow of Control · Structure · Communication · Advanced Data Management Services

#### **134 Programming Languages** *D. M. Volpano*

Principles of Programming Languages · Program Verification · Programming Language Paradigms

#### **135 Input/Output Devices** *R. Freitas*

Input/Output Subsystem · I/O Devices

#### **136 Memory and Mass Storage Systems** *P. M. Chen*

Aspects of Storage Devices · Types of Storage Devices · Storage Organization

## SECTION XXII Measurement and Instrumentation

---

### **Introduction** *Wolf W. von Maltzahn*

#### **137 Sensors and Transducers** *R. L. Smith*

Physical Sensors · Chemical Sensors · Biosensors · Microsensors

#### **138 Measurement Errors and Accuracy** *S. J. Harrison*

Measurement Errors, Accuracy, and Precision · Estimating Measurement Uncertainty · Propagation of Measurement Uncertainty · Uncertainty Analysis in Experimental Design

#### **139 Signal Conditioning** *S. A. Dyer*

Linear Operations · Nonlinear Operations

#### **140 Telemetry** *S. Horan*

Telemetry Systems · Frame Telemetry · Packet Telemetry

#### **141 Recording Instruments** *W. Owens*

Types of Recording Instruments · Methods of Data Recording · Future Directions: Distributed Data Acquisition and Recording

#### **142 Bioinstrumentation** *W. W. von Maltzahn and K. Meyer-Waarden*

Basic Bioinstrumentation Systems · Applications and Examples · Summary

## SECTION XXIII Surveying

---

### **Introduction** *Jack McCormac*

#### **143 Quality Control** *B. H. W. van Gelder*

Errors · Precision · Law of Propagation of Errors · Statistical Testing · Accuracy · Reliability · Actuality

#### **144 Elevation** *S. D. Johnson*

Measures of Elevation and Height · Deflection of the Vertical · Vertical Datums · Elevation Measurement · Systematic Errors

#### **145 Distance Measurements** *R. B. Buckner*

Fundamentals of Distance Measurement · Applications and Calculations

#### **146 Directions** *B. A. Dewitt*

Angles · Meridians · Direction · Back Bearing and Back Azimuth · Applications

#### **147 Photogrammetry and Topographic Mapping** *J. Bethel*

Basic Concepts · Orientation and Model Setup · Data Collection for Topography · Data Processing for Topography · Data Presentation

#### **148 Surveying Computations** *B. H. W. van Gelder*

Principles of Multivariate Calculus · Principles of Linear Algebra · Model of Two Sets of Variables, Observations, and Parameters: The Mixed Model · Observations as a Function of Parameters Only: The Model of Observation Equations · All Parameters Eliminated: The Model of Condition Equations · An Example: Traversing · Dynamical Systems

#### **149 Satellite Surveying** *B. H. W. van Gelder*

A Satellite Orbiting the Earth · The Orbital Ellipse · Relationship between Cartesian and Keplerian Orbital Elements · Orbit of a Satellite in a Noncentral Force Field · The Global Positioning System (GPS) · Gravity Field and Related Issues

#### **150 Surveying Applications for Geographic Information Systems** *J. F. Thompson*

GIS Fundamentals · Monumentation or Control Surveying · Topographic Surveying · Future GIS Surveying Applications

#### **151 Remote Sensing** *R. W. Kiefer and T. M. Lillesand*

Electromagnetic Energy · Atmospheric Effects · Remote Sensing Systems · Remote Sensing from Earth Orbit · Digital Image Processing

## SECTION XXIV Control Systems

---

### Introduction *Francis H. Raven*

#### 152 Feedback *W. L. Brogan*

Characteristics of a Feedback Control System • Fundamentals of Feedback for Time-invariant Linear Systems • Illustrative Applications • Feedback in Nonlinear Systems • Summary of the Typical Steps in the Design of Feedback Systems

#### 153 Root Locus *W. L. Brogan*

The Root Locus Concept • Root Locus Details • Generation of Root Locus Plots • Examples • Summary and Conclusions

#### 154 Nyquist Criterion and Stability *N. S. Nise*

Concept and Definition of Frequency Response • Plotting Frequency Response • Stability • Nyquist Criterion for Stability • Gain Design for Stability via the Nyquist Criterion • Stability via Separate Magnitude and Phase Plots (Bode Plots)

#### 155 System Compensation *F. H. Raven*

Correlation between Transient and Frequency Response • Determining  $K$  to Yield a Desired  $\zeta$  • Gain Margin and Phase Margin • Series Compensation • Internal Feedback • Compensation on the  $S$  Plane

#### 156 Process Control *T. E. Marlin*

Control Performance and Design Decisions

#### 157 Digital Control *R. G. Jacquot*

Feedback Control • Digital Control • Microcontroller Architecture • Linear Digital Control • Digital Control Stability Analysis and Design • Computer-Aided Design

#### 158 Robots and Control *R. G. Bonitz and T. C. Hsia*

Independent Joint Control • Method of Computed Torque • Cartesian-Space Control

#### 159 State Variable Feedback *T. L. Vincent*

Linear State Space Control Systems • Controllability and Observability • Eigenvalue Placement • Observer Design

## SECTION XXV Manufacturing

---

### Introduction *Frank Kreith*

#### 160 Types of Manufacturing *R. J. Schonberger*

Job-Shop and Batch Production • Mass Production • Continuous Production • Mixtures and Gray Areas • Capital Investment, Automation, Advanced Technology, Skills, and Layout

#### 161 Quality *M. P. Stephens and J. F. Kmec*

Measurement • Statistical Quality Control • Tolerances and Capability

#### 162 Flexible Manufacturing *A. Kusiak and C.-X. Feng*

Flexible Machining • Flexible Assembly • The Economic Justification of Flexibility

#### 163 Management and Scheduling *E. M. Knod, Jr.*

Management: Definition and Innovations • Scheduling

#### 164 Design, Modeling, and Prototyping *W. L. Chapman and A. T. Bahill*

The System Design Process • Rapid Prototyping • When to Use Modeling and Prototyping

#### 165 Materials Processing and Manufacturing Methods *S. H. Risbud*

Processing Metals and Alloys • Ceramics, Glasses, and Polymers • Joining of Materials

#### **166 Machine Tools and Processes** *Y. C. Shin*

Economic Impact · Types of Machine Tools · Control of Machine Tools · Machine Tool Accuracy

#### **167 Human Factors and Ergonomics** *W. Karwowski and B. Jamaldin*

The Concept of Human-Machine Systems · Ergonomics in Industry · The Role of Ergonomics in Prevention of Occupational Musculoskeletal Injury · Fitting the Work Environment to the Workers

#### **168 Pressure and Vacuum** *P. Biltoft, C. Borzileri, D. Holten, and M. Traini*

Pressure · The Vacuum Environment

#### **169 Food Engineering** *R. P. Singh*

Liquid Transport Systems · Heat Transfer

#### **170 Agricultural Engineering** *D. J. Hills*

Equipment Sizing Criteria · Equipment Selection

#### **171 System Reliability** *R. Ramakumar*

Catastrophic Failure Models · The Bathtub Curve · Mean Time to Failure · Average Failure Rate · A Posteriori Failure Probability · Units for Failure Rates · Application of the Binomial Distribution · Application of the Poisson Distribution · The Exponential Distribution · The Weibull Distribution · Combinatorial Aspects · Modeling Maintenance · Markov Models · Binary Model for a Repairable Component · Two Dissimilar Repairable Components · Two Identical Repairable Components · Frequency and Duration Techniques · Applications of Markov Process · Some Useful Approximations

### **SECTION XXVI Aeronautical and Aerospace**

---

#### **Introduction** *Cary R. Spitzer*

#### **172 Aerodynamics** *J. F. Donovan*

Background · Flow about a Body · Two-Dimensional Airfoils · Finite Wing Effects · Effects of Compressibility

#### **173 Stability and Turbulence** *R. A. Hess*

Descriptions of Atmospheric Movement · Turbulence and Aircraft Dynamics · An Example · Other Applications

#### **174 Computational Fluid Dynamics** *R. K. Agarwal*

Geometry Modeling and Grid Generation · Flow Simulation Algorithms · Turbulence Modeling · Flow Simulation Examples · Future Directions and Challenges

#### **175 Aerospace Materials** *R. R. June*

System Requirements and Materials Selection · Design Considerations · Nonstructural Materials · The Future

#### **176 Propulsion Systems** *J. C. Monk*

Performance Characteristics · Liquid Rocket Engine Cycles · Major Components · System Preliminary Design Process · Conclusion

#### **177 Aircraft Performance and Design** *F. J. Hale*

Aircraft Forces and Subsystems · Level Flight · Climbing Flight · Turning Flight

#### **178 Spacecraft and Mission Design** *W. T. Fowler*

Spacecraft Environments · Fundamental Principles · Spacecraft/Mission Categories · Spacecraft Subsystems · Spacecraft/Mission Design Process

### **SECTION XXVII Safety**

---

#### **Introduction** *A. Keith Furr*

#### **179 Hazard Identification and Control** *M. Rahimi*

Hazard Identification Physical Hazards Chemical Hazards Airborne Contaminants Noise Fire Hazards An Engineering Approach to Hazard Control Hazard Analysis and Quantification

[180 Regulations and Standards](#) *A. K. Furr*

Engineering Practices Summary

## **SECTION XXVIII Engineering Economics and Management**

---

**Introduction** *Tung Au*

[181 Present Worth Analysis](#) *W. D. Short*

Calculation · Application · Other Considerations

[182 Project Analysis Using Rate-of-Return Criteria](#) *R. G. Beaves*

Net Present Value · Internal Rate of Return · Overall Rate of Return · Project Investment Base ·

Scale-Adjusted ORR · Project Life Differences · Conclusion

[183 Project Selection from Alternatives](#) *C. Hendrickson and S. McNeil*

Problem Statement for Project Selection · Steps in Carrying Out Project Selection · Selection Criteria ·

Applications · Conclusion

[184 Depreciation and Corporate Taxes](#) *T. Au*

Depreciation as Tax Deduction · Tax Laws and Tax Planning · Decision Criteria for Project Selection ·

Inflation Consideration · After-Tax Cash Flows · Evaluation of After-Tax Cash Flows · Effects of Various Factors

[185 Financing and Leasing](#) *C. Fazzi*

Debt Financing · Equity Financing · Leasing

[186 Risk Assessment](#) *R. T. Ruegg*

Expected Value (EV) Analysis · Mean-Variance Criterion (MVC) and Coefficient of Variation (CV) ·

Risk-Adjusted Discount Rate (RADR) Technique · Certainty Equivalent (CE) Technique · Simulation

Technique · Decision Analysis

[187 Sensitivity Analysis](#) *H. E. Marshall*

Sensitivity Analysis Applications · Advantages and Disadvantages

[188 Life-Cycle Costing](#) *W. J. Fabrycky and B. S. Blanchard*

The Life-Cycle Costing Situation · Cost Generated over the Life Cycle · The Cost Breakdown Structure ·

Life-Cycle Cost Analysis · Cost Treatment over the Life Cycle · Summary

[189 Project Evaluation and Selection](#) *H. J. Thamhain*

Quantitative Approaches to Project Evaluation and Selection · Qualitative Approaches to Project

Evaluation and Selection · Recommendations for Effective Project Evaluation and Selection

[190 Critical Path Method](#) *J. L. Richards*

Planning the Project · Scheduling the Project · Controlling the Project · Modifying the Project Schedule ·

Project Management Using CPM

[191 Patents, Copyrights, Trademarks, and Licenses](#) *P. A. Beck and C. I. Bordas*

Patents · Copyrights · Trademarks · Licenses

## **SECTION XXIX Materials Engineering**

---

**Introduction** *James F. Shackelford*

[192 Properties of Solids](#) *J. F. Shackelford*

Structure · Composition · Physical Properties · Mechanical Properties · Thermal Properties · Chemical

Properties · Electrical and Optical Properties

[193 Failure Analysis](#) *J. F. Shackelford*

Types of Failures · Failure Analysis Methodology · Fracture Mechanics · Nondestructive Testing ·

Engineering Design for Failure Prevention

[194 Liquids and Gases](#) *B. E. Poling*

Viscosity · Thermal Conductivity · Heat Capacity · Vapor Pressure

[195 Biomaterials](#) *R. B. Martin*

History · Problems Associated with Implanted Devices · Immunology and Biocompatibility · Commonly Used Implant Materials · Metals · Polymers · Ceramics · Carbon Materials

## **SECTION XXX Mathematics**

---

**Introduction** *William F. Ames*

[196 General Mathematics](#)

Trigonometry · Series · Differential Calculus · Integral Calculus · Special Functions

[197 Linear Algebra Matrices](#) *G. Cain*

Basic Definitions · Algebra of Matrices · Systems of Equations · Vector Spaces · Rank and Nullity · Orthogonality and Length · Determinants · Eigenvalues and Eigenvectors

[198 Vector Algebra and Calculus](#) *G. Cain*

Basic Definitions · Coordinate Systems · Vector Functions · Gradient, Curl, and Divergence · Integration · Integral Theorems

[199 Complex Variables](#) *G. Cain*

Basic Definitions and Arithmetic · Complex Functions · Analytic Functions · Integration · Series · Singularities · Conformal Mapping

[200 Difference Equations](#) *W. F. Ames*

First-Order Equations · Second-Order Equations · Linear Equations with Constant Coefficients · Generating Function ( $z$  Transform)

[201 Differential Equations](#) *W. F. Ames*

Ordinary Differential Equations · Partial Differential Equations

[202 Integral Equations](#) *W. F. Ames*

Classification and Notation · Relation to Differential Equations · Methods of Solution

[203 Approximation Methods](#) *W. F. Ames*

Perturbation · Iterative Methods

[204 Integral Transforms](#) *W. F. Ames*

Laplace Transform · Convolution Integral · Fourier Transform · Fourier Cosine Transform

[205 Chaos, Fractals, and Julia Sets](#) *A. Deliu*

Chaos · Fractals · Julia Sets

[206 Calculus of Variations](#) *W. F. Ames*

The Euler Equation · The Variation · Constraints

[207 Probability and Statistics](#) *Y. L. Tong*

Elementary Probability · Random Sample and Sampling Distributions · Normal Distribution-Related Sampling Distributions · Confidence Intervals · Testing Statistical Hypotheses · A Numerical Example

[208 Optimization](#) *G. Cain*

Linear Programming · Unconstrained Nonlinear Programming · Constrained Nonlinear Programming

[209 Numerical Methods](#) *W. F. Ames*

Linear Algebra Equations · Nonlinear Equations in One Variable · General Methods for Nonlinear Equations in One Variable · Numerical Solution of Simultaneous Nonlinear Equations · Interpolation and Finite Differences · Numerical Differentiation · Numerical Integration · Numerical Solution of Ordinary Differential Equations · Numerical Solution of Integral Equations · Numerical Methods for Partial Differential Equations · Discrete and Fast Fourier Transform · Software

[210 Dimensional Analysis](#) *W. F. Ames*

Units and Variables · Method of Dimensions



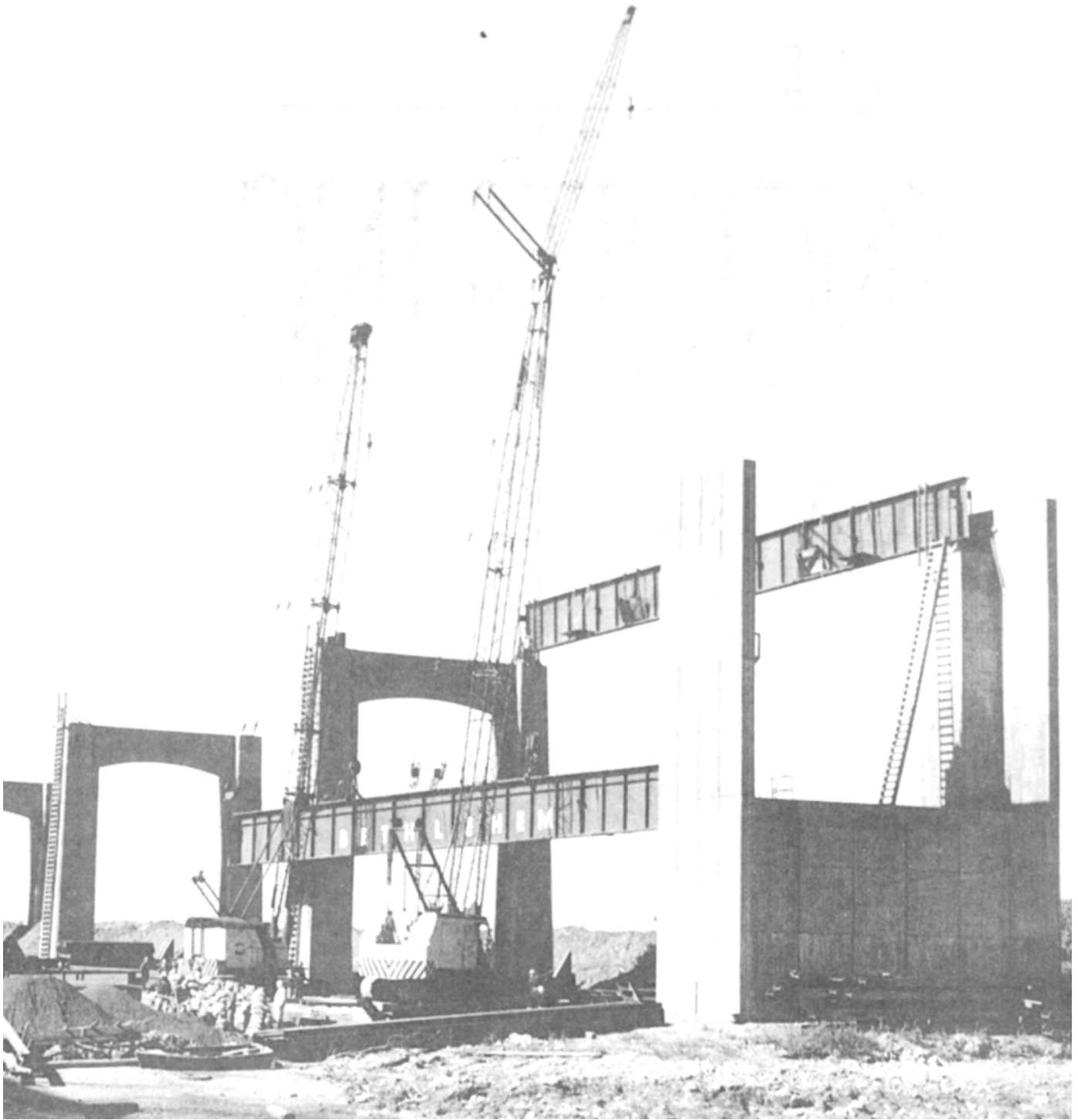
## 211 Computer Graphics Visualization *R. S. Gallagher*

The Display of Objects in 3-D · Scalar Display Techniques · Vector and Tensor Field Display · Continuum  
Volume Visualization · Animation over Time · Summary

## **Appendix: Mathematical Tables and Formulae**

## **Associations and Societies**

Hibbeler R. C. "Statics"  
*The Engineering Handbook.*  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



An understanding of the basic principles of statics is necessary for both the design and construction phases of the Delaware Memorial Bridge, shown here. The first steel for the nearly half-mile-long plate girder portion of the west approach is being hoisted into place and set atop 80-foot-high land piers. Each plate girder is 118 feet long, 9 feet deep, weighing 35 tons. (Photo courtesy of Bethlehem Steel.)

# I

## Statics

---

**Russell C. Hibbeler**

*University of Southwestern Louisiana*

**1 Force-System Resultants and Equilibrium** *R. C. Hibbeler*

Force-System • Resultants • Equilibrium

**2 Centroids and Distributed Forces** *W. D. Pilkey and L. Kitis*

Centroid of a Plane Area • Centroid of a Volume Surface • Forces • Line Forces • Calculation of Surface Area and Volume of a Body with Rotational Symmetry • Determination of Centroids

**3 Moments of Inertia** *J. L. Meriam*

Area Moments of Inertia • Mass Moments of Inertia

STATICS IS THE STUDY of the resultants of force systems and is concerned with problems that involve the equilibrium of a body. It is a very practical subject of vital importance in the design and analysis of all structural and mechanical components. For this reason, a fundamental understanding of statics is imperative if one is to build any structure or perform a force analysis of linkage, gearing, or the framework for a machine.

The subject of statics is the oldest branch of mechanics, with its beginnings at the time of the Babylonians and Egyptians. Archimedes recorded how forces act on levers; however, the main principles of statics were developed by the 17th century, notably from the work of Varignon, Stevinus, and Newton. These principles are few in number and have been established through experience and verified by experiment.

The chapters of this section provide a comprehensive review of the many topics covered in statics, including simplification of concentrated and distributed force systems, the definition of the moment of a force and couple, the necessary and sufficient conditions for equilibrium, frictional effects, and a discussion of the geometric properties of an area, namely, the centroid and moment of inertia. The latter topics are of vital importance in the development of many formulas used in mechanics of materials and hydrostatics, as will be shown in later sections of this handbook.

Hibbeler R. C. "Force-System Resultants and Equilibrium"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Force-System Resultants and Equilibrium

---

## 1.1 Force-System Resultants

Concurrent Force Systems • Moment of a Force • Couple • Resultants of a Force and Couple System • Distributed Loadings

## 1.2 Equilibrium

Equations of Equilibrium • Free-Body Diagram • Support Reactions • Friction • Constraints • Internal Loadings • Numerical Applications

**Russell C. Hibbeler**

*University of Louisiana*

Statics is a branch of mechanics that deals with the equilibrium of bodies, that is, those that are either at rest or move with constant velocity. In order to be able to apply the laws of statics, it is first necessary to understand how to simplify force systems and compute the moment of a force. In this chapter these topics will be discussed, and some examples will be presented to show how the laws of statics are applied.

## 1.1 Force-System Resultants

---

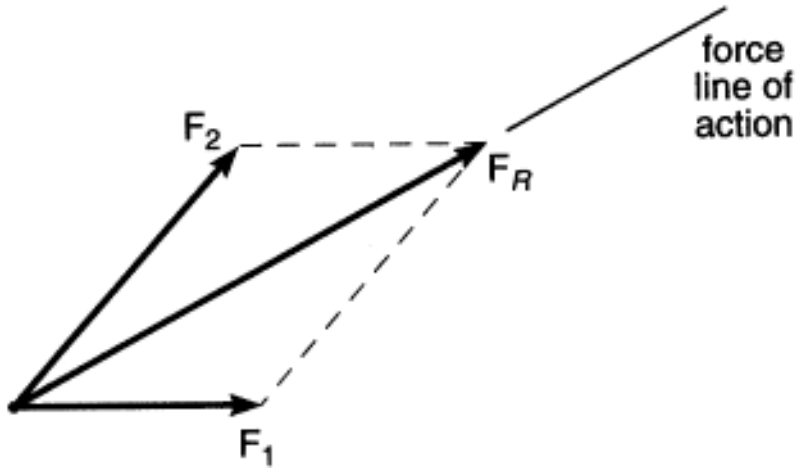
### Concurrent Force Systems

Force is a vector quantity that is completely characterized by its magnitude, direction, and point of application. When two forces  $F_1$  and  $F_2$  are **concurrent** they can be added together to form a resultant  $F_R = F_1 + F_2$  using the **parallelogram law**, Fig. 1.1. Here  $F_1$  and  $F_2$  are referred to as components of  $F_R$ . Successive applications of the parallelogram law can also be applied when several concurrent forces are to be added; however, it is perhaps simplest first to determine the two components of each force along the axes of a coordinate system and then add the respective components. For example the  $x$ ,  $y$ , and  $z$  (or Cartesian) components of  $\mathbf{F}$  are shown in Fig. 1.2. Here,  $\mathbf{i}$ ,  $\mathbf{j}$ ,  $\mathbf{k}$  are unit vectors used to define the direction of the positive  $x$ ,  $y$ , and  $z$  axes, and  $F_x$ ,  $F_y$ , and  $F_z$  are the magnitudes of each component. By vector addition,  $\mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k}$ . When each force in a concurrent system of forces is expressed by its Cartesian components, the resultant force is therefore

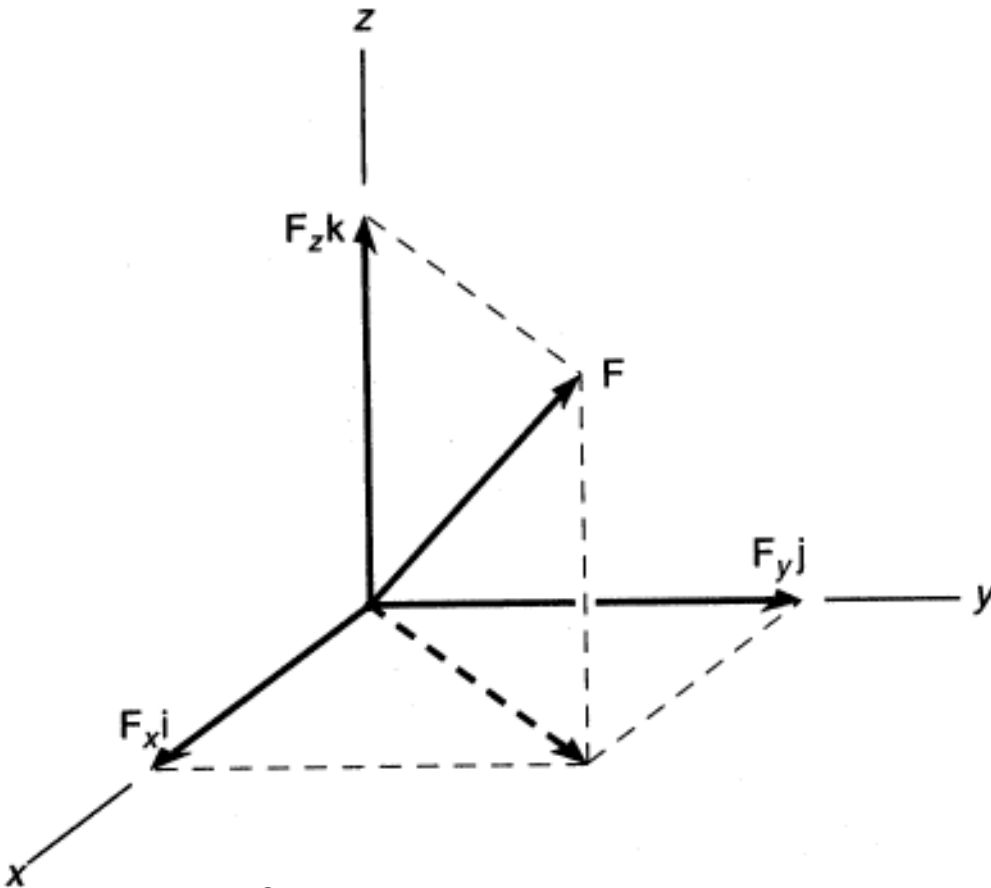
$$\mathbf{F}_R = \Sigma F_x \mathbf{i} + \Sigma F_y \mathbf{j} + \Sigma F_z \mathbf{k} \quad (1.1)$$

where  $\Sigma F_x$ ,  $\Sigma F_y$ ,  $\Sigma F_z$  represent the scalar additions of the  $x$ ,  $y$ , and  $z$  components.

**Figure 1.1** Addition of forces by parallelogram law.



**Figure 1.2** Resolution of a vector into its  $x, y, z$  components.

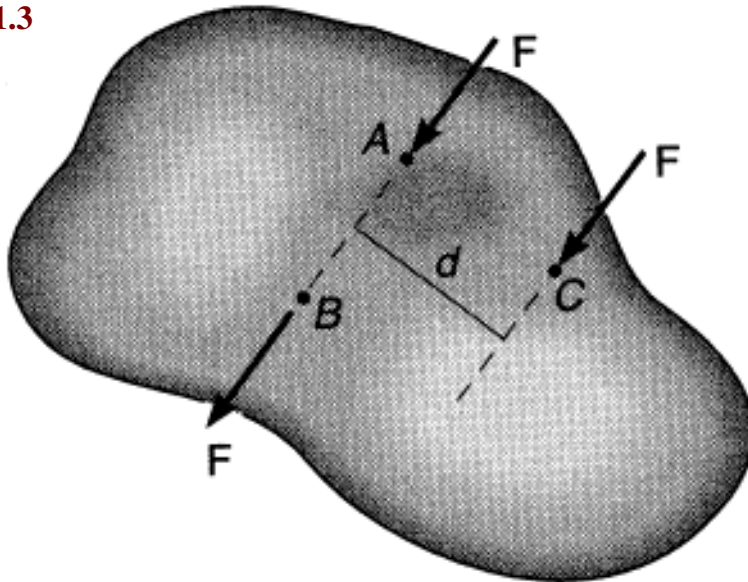


## Moment of a Force

When a force  $\mathbf{F}$  acts on a body, it will cause both external and internal effects on the body. These effects depend upon where the force is located. For example, if  $\mathbf{F}$  acts at point A on the body in [Fig. 1.3](#), it will cause a specific translation and rotation of the body. If instead  $\mathbf{F}$  is applied to some

other point,  $B$ , which lies along the line of action of  $\mathbf{F}$ , the external effects regarding the motion of the body remain unchanged, although the body's internal effects will be different. This is referred to as the **principle of transmissibility**. However, if the force acts at point  $C$ , which is not along the line of action  $AB$ , then both the external and internal effects on the body will change. The difference in external effects—notably the difference in the rotation of the body—occurs because of the distance  $d$  that separates the lines of action of the two positions of the force.

**Figure 1.3**



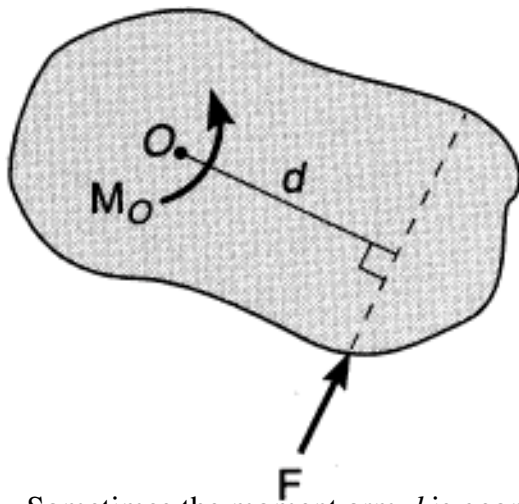
This tendency for the body to rotate about a specified point  $O$  or axis as caused by a force is a vector quantity called a *moment*. By definition, the magnitude of the moment is

$$M_O = Fd \quad (1.2)$$

where  $d$  is the moment arm or perpendicular distance from the point to the line of action of the force, as in Fig. 1.4. The direction of the moment is defined by the right-hand rule, whereby the curl of the right-hand fingers follows the tendency for rotation caused by the force, and the thumb specifies the directional sense of the moment. In this case,  $M_O$  is directed out of the page, since  $\mathbf{F}$  produces counterclockwise rotation about  $O$ . It should be noted that the force can act at any point along its line of action and still produce the same moment about  $O$ .

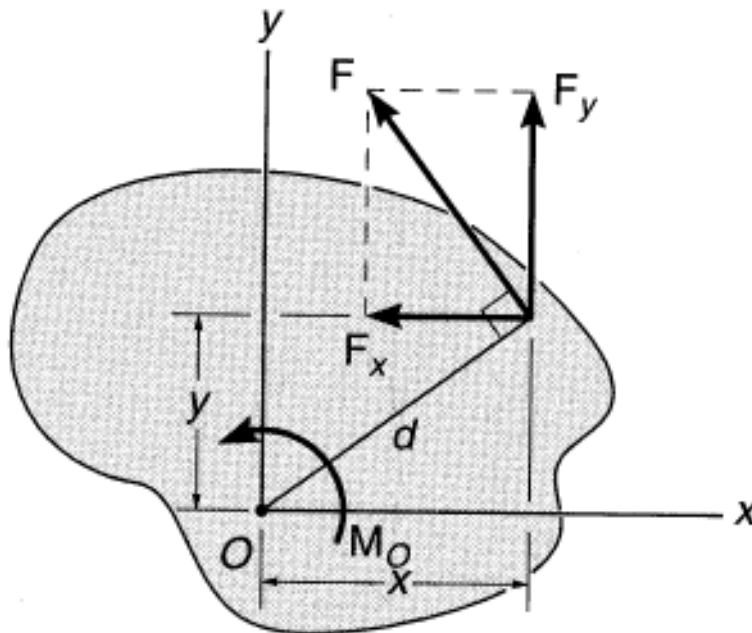


**Figure 1.4 Moment of a force.**



Sometimes the moment arm  $d$  is geometrically hard to determine. To make the calculation easier, the force is first resolved into its Cartesian components and then the moment about point  $O$  is determined using the **principle of moments**, which states that the moment of the force about  $O$  is equal to the sum of the moments of the force's components about  $O$ . Thus, as shown in Fig. 1.5, we have  $M_O = Fd = F_x y + F_y x$ .

**Figure 1.5**



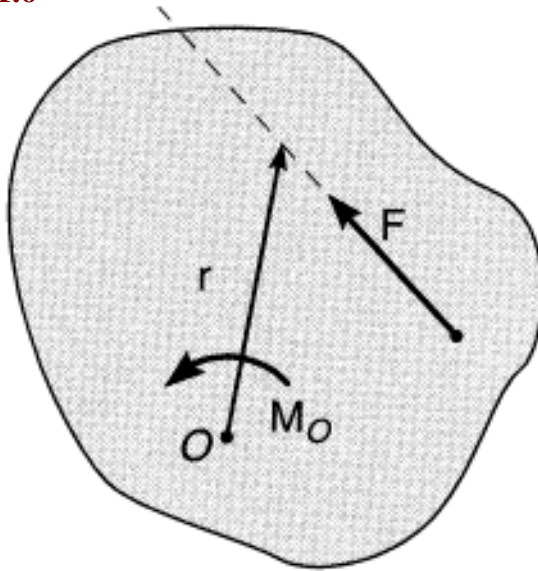
The moment about point  $O$  can also be expressed as a vector cross product of the position vector  $\mathbf{r}$  directed from  $O$  to any point on the line of action of the force  $\mathbf{F}$ , as shown in Fig. 1.6. Here,

$$\mathbf{M}_O = \mathbf{r} \times \mathbf{F} \quad (1.3)$$

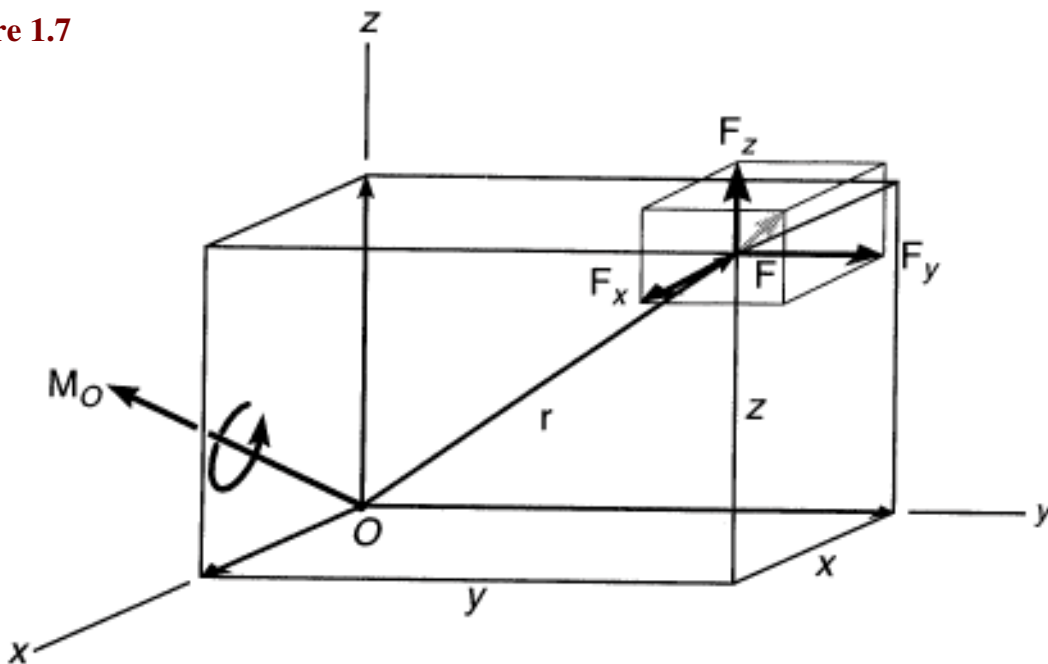
If  $\mathbf{r}$  and  $\mathbf{F}$  are expressed in terms of their Cartesian components, then as in Fig. 1.7 the Cartesian components for the moment about  $O$  are

$$\begin{aligned}
 \mathbf{M}_O &= \mathbf{r} \times \mathbf{F} = (x\mathbf{i} + y\mathbf{j} + z\mathbf{k}) \times (F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k}) \\
 &= (yF_z - zF_y)\mathbf{i} + (zF_x - xF_z)\mathbf{j} + (xF_y - yF_x)\mathbf{k} \\
 &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x & y & z \\ F_x & F_y & F_z \end{vmatrix}
 \end{aligned} \tag{1.4}$$

**Figure 1.6**



**Figure 1.7**



## Couple

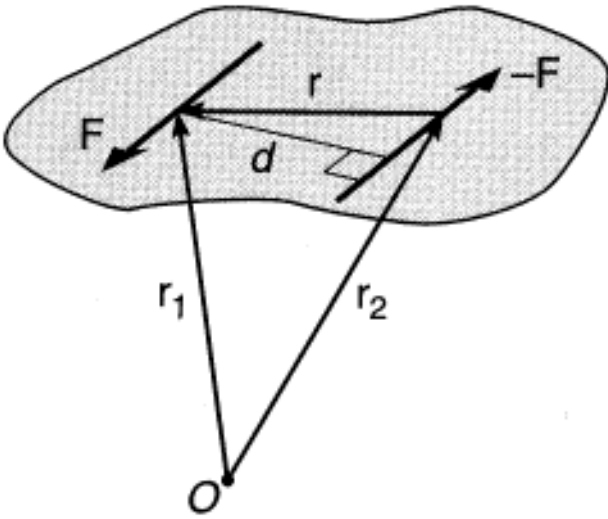
A **couple** is defined as two parallel forces that have the same magnitude and opposite directions,

and are separated by a perpendicular distance  $d$ , as in Fig. 1.8. The moment of a couple about the arbitrary point  $O$  is

$$\begin{aligned} \mathbf{M}_C &= \mathbf{r}_1 \times \mathbf{F} + \mathbf{r}_2 \times (-\mathbf{F}) = (\mathbf{r}_1 - \mathbf{r}_2) \times \mathbf{F} \\ &= \mathbf{r} \times \mathbf{F} \end{aligned} \quad (1.5)$$

Here the couple moment  $\mathbf{M}_C$  is independent of the location of the moment point  $O$ . Instead, it depends only on the distance between the forces; that is,  $\mathbf{r}$  in the above equation is directed from any point on the line of action of one of the forces ( $-\mathbf{F}$ ) to any point on the line of action of the other force  $\mathbf{F}$ . The external effect of a couple causes rotation of the body with no translation, since the resultant force of a couple is zero.

**Figure 1.8**



## Resultants of a Force and Couple System

A general force and couple-moment system can always be replaced by a single resultant force and couple moment acting at any point  $O$ . As shown in Figs. 1.9(a,b) these resultants are

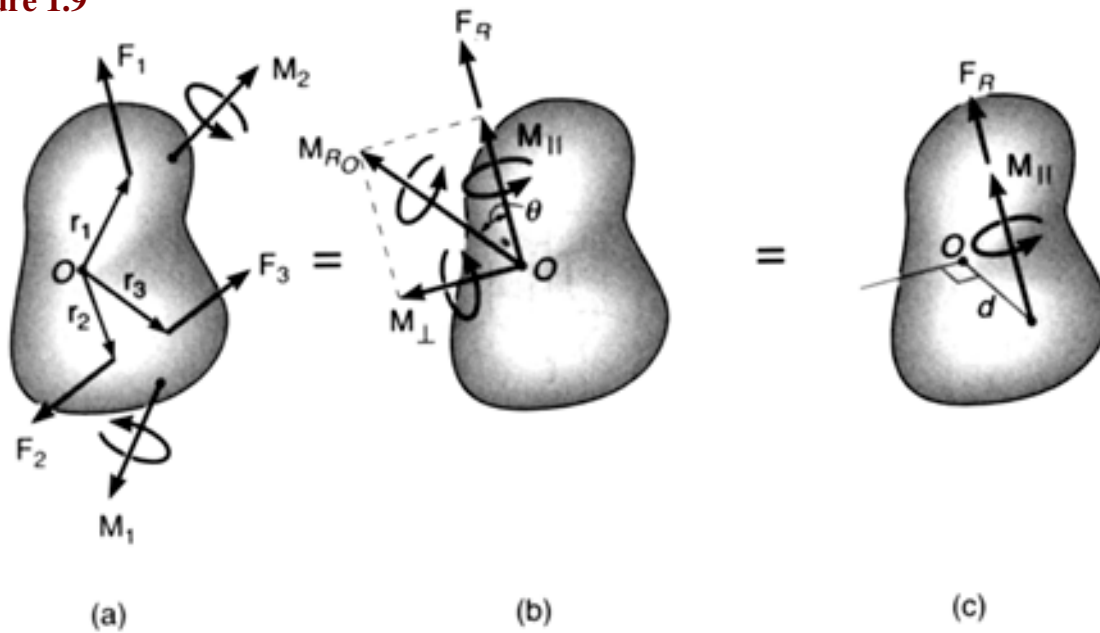
$$\mathbf{F}_R = \Sigma \mathbf{F} \quad (1.6)$$

$$\mathbf{M}_{R_O} = \Sigma \mathbf{M}_O \quad (1.7)$$

where  $\Sigma \mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2 + \mathbf{F}_3$  is the vector addition of all the forces in the system, and  $\Sigma \mathbf{M}_O = (\mathbf{r}_1 \times \mathbf{F}_1) + (\mathbf{r}_2 \times \mathbf{F}_2) + (\mathbf{r}_3 \times \mathbf{F}_3) + \mathbf{M}_1 + \mathbf{M}_2$  is the vector sum of the moments of all the forces about point  $O$  plus the sum of all the couple moments. This system may be further simplified by first resolving the couple moment  $\mathbf{M}_{R_O}$  into two components—one parallel and the other perpendicular to the force  $\mathbf{F}_R$ , as in Fig. 1.9(b). By moving the line of action of  $\mathbf{F}_R$  in the plane perpendicular to  $\mathbf{M}_\perp$  a distance  $d = M_\perp / F_R$ , so that  $\mathbf{F}_R$  creates the moment  $\mathbf{M}_\perp$  about  $O$ , the system can then be represented by a **wrench**, that is, a single force  $\mathbf{F}_R$  and collinear moment

$M_{\parallel}$  , Fig. 1.9(c).

**Figure 1.9**



Note that in the special case of  $\theta = 90^\circ$  , Fig. 1.9(b),  $M_{\parallel} = 0$  and the system only reduces to a single resultant force  $F_R$  having a specified line of action. This will always be the case if the force system is either concurrent, parallel, or coplanar.

## Distributed Loadings

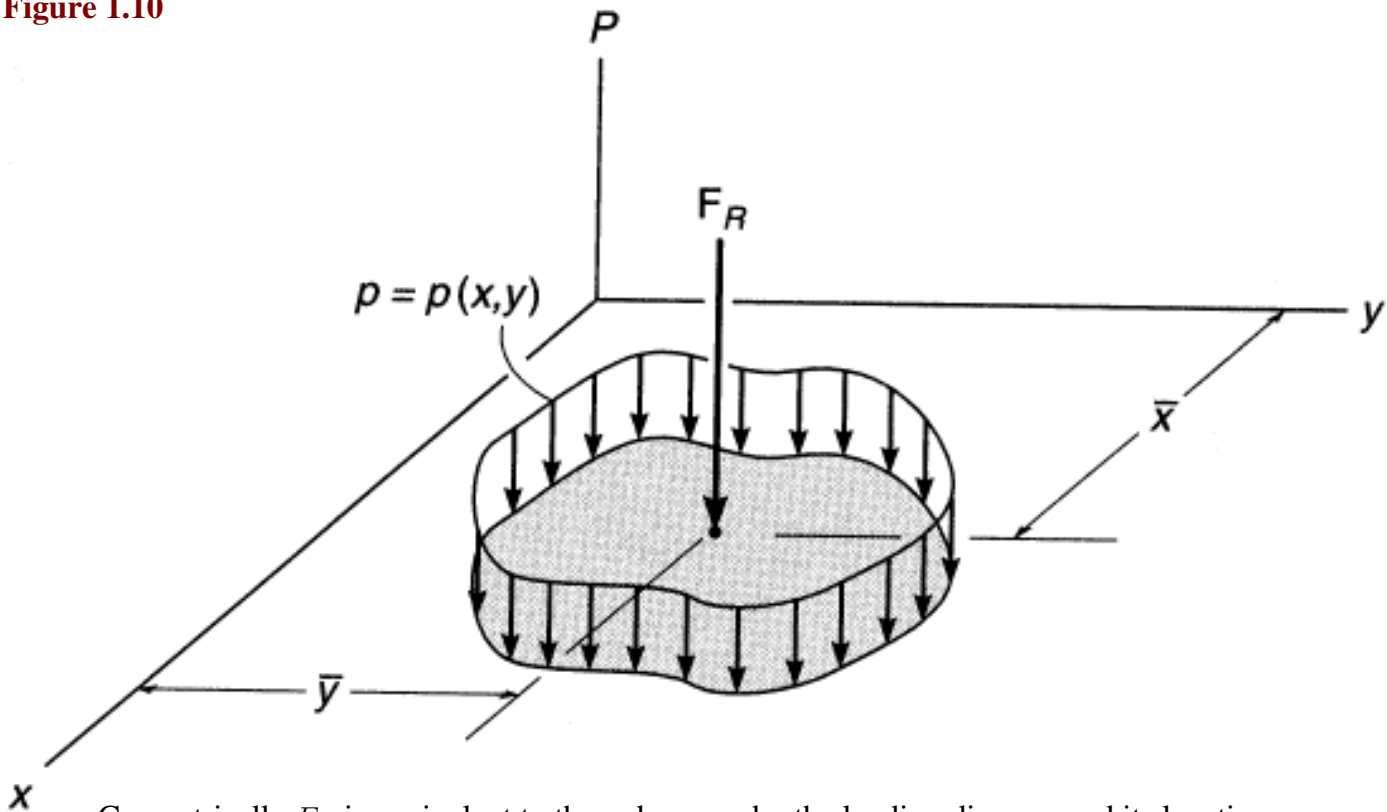
When a body contacts another body, the loads produced are always distributed over the area of each body. If the area on one of the bodies is small compared to the entire surface area of the body, the loading can be represented by a single concentrated force acting at a point on the body.

However, if the loading occurs over a large surface area—such as that caused by wind or a fluid—the distribution of load must be taken into account. The intensity of this surface loading at each point is defined as a pressure and its variation is defined by a load-intensity diagram. On a flat surface the load intensity diagram is described by the loading function  $p = p(x, y)$  , which consists of an infinite number of parallel forces, as in Fig. 1.10. Applying Eqs. (1.6) and (1.7), the resultant of this loading and its point of application  $(\bar{x}, \bar{y})$  can be determined from

$$F_R = \int p(x, y) \, dA \quad (1.8)$$

$$\bar{x} = \frac{\int x \, p(x, y) \, dA}{\int p(x, y) \, dA} \quad \bar{y} = \frac{\int y \, p(x, y) \, dA}{\int p(x, y) \, dA} \quad (1.9)$$

**Figure 1.10**



Geometrically  $F_R$  is equivalent to the volume under the loading diagram and its location passes through the centroid or geometric center of this volume. Often in engineering practice, the surface loading is symmetric about an axis, in which case the loading is a function of only one coordinate,  $w = w(x)$ . Here the resultant is geometrically equivalent to the area under the loading curve, and the line of action of the resultant passes through the centroid of this area.

Besides surface forces as discussed above, loadings can be transmitted to another body without direct physical contact. These body forces are distributed throughout the volume of the body. A common example is the force of gravity. The resultant of this force is termed the **weight**; it acts through the body's center of gravity and is directed toward the center of the earth.

## 1.2 Equilibrium

---

### Equations of Equilibrium

A body is said to be in equilibrium when it is either at rest or moves with constant velocity. For purposes of analysis, it is assumed that the body is perfectly rigid, meaning that the particles composing the body remain at fixed distances from one another both before and after applying the load. Most engineering materials deform only slightly under load, so that moment arms and the orientation of the loading remain essentially constant. For these cases, therefore, the rigid-body model is appropriate for analysis. The necessary and sufficient conditions to maintain equilibrium of a rigid body require the resultant external force and moment acting on the body to be equal to zero. From Eqs. (1.6) and (1.7) this can be expressed mathematically as

$$\Sigma \mathbf{F} = \mathbf{0} \quad (1.10)$$

$$\Sigma \mathbf{M}_O = \mathbf{0} \quad (1.11)$$

If the forces acting on the body are resolved into their  $x$ ,  $y$ , and  $z$  components, these equations can be written in the form of six scalar equations, namely,

$$\begin{aligned} \Sigma F_x &= 0 & \Sigma M_{Ox} &= 0 \\ \Sigma F_y &= 0 & \Sigma M_{Oy} &= 0 \\ \Sigma F_z &= 0 & \Sigma M_{Oz} &= 0 \end{aligned} \quad (1.12)$$

Actually, any set of three nonorthogonal, nonparallel axes will be suitable references for either of these force or moment summations.

If the forces on the body can be represented by a system of coplanar forces, then only three equations of equilibrium must be satisfied, namely,

$$\begin{aligned} \Sigma F_x &= 0 \\ \Sigma F_y &= 0 \\ \Sigma M_O &= 0 \end{aligned} \quad (1.13)$$

Here the  $x$  and  $y$  axes lie in the plane of the forces and point  $O$  can be located either on or off the body.

## Free-Body Diagram

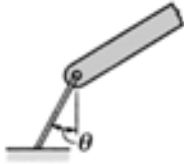
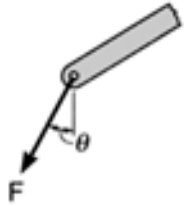

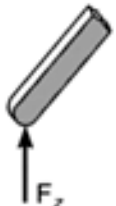

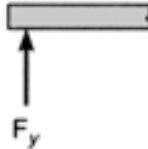

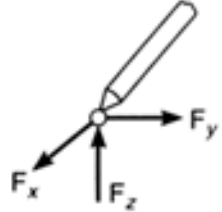

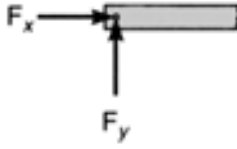

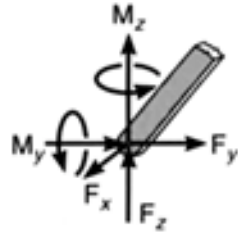

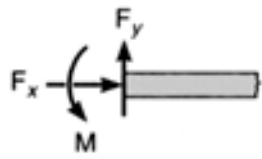

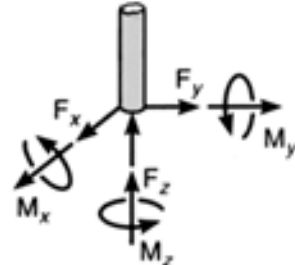
Application of the equations of equilibrium requires accountability for *all* the forces that act on the body. The best way to do this is to draw the body's **free-body diagram**. This diagram is a sketch showing an outlined shape of the body and so represents it as being isolated or "free" from its surroundings. On this sketch it is necessary to show all the forces and couples that act on the body. Those generally encountered are due to applied loadings, reactions that occur at the supports and at points of contact with other bodies, and the weight of the body. Also one should indicate the dimensions of the body necessary for computing the moments of forces and label the known and unknown magnitudes as well as directions of the forces and couple moments. Once the free-body diagram has been drawn and the coordinate axes established, application of the equations of equilibrium becomes a straightforward procedure.

## Support Reactions

Various types of supports can be used to prevent a body from moving. [Table 1.1](#) shows some of the most common types, along with the reactions each exerts on the body at the connection. As a

general rule, if a support prevents translation in a given direction, then a force is developed on the body in that direction, whereas if rotation is prevented, a couple moment is exerted on the body.

**Table 1.1** Force Systems

Connection	Reaction	Connection	Reaction
 cable		 smooth surface	
 roller		 ball and socket	
 pin		 single pin	
 fixed support		 fixed support	

## Friction

When a body is in contact with a rough surface, a force of resistance called **friction** is exerted on the body by the surface in order to prevent or retard slipping of the body. This force always acts tangent to the surface at points of contact with the surface and is directed so as to oppose the possible or existing motion of the body. If the surface is dry, the frictional force acting on the body must satisfy the equation

$$F < \mu_s N \quad (1.14)$$

The equality  $F = \mu_s N$  applies only when motion is impending. Here  $N$  is the resultant normal force on the body at the surface of contact, and  $\mu_s$  is the coefficient of static friction, a

dimensionless number that depends on the characteristics of the contacting surfaces. Typical values of  $\mu_s$  are shown in Table 1.2. If the body is sliding, then  $F = \mu_k N$ , where  $\mu_k$  is the coefficient of kinetic motion, a number which is approximately 25% smaller than those listed in Table 1.2.

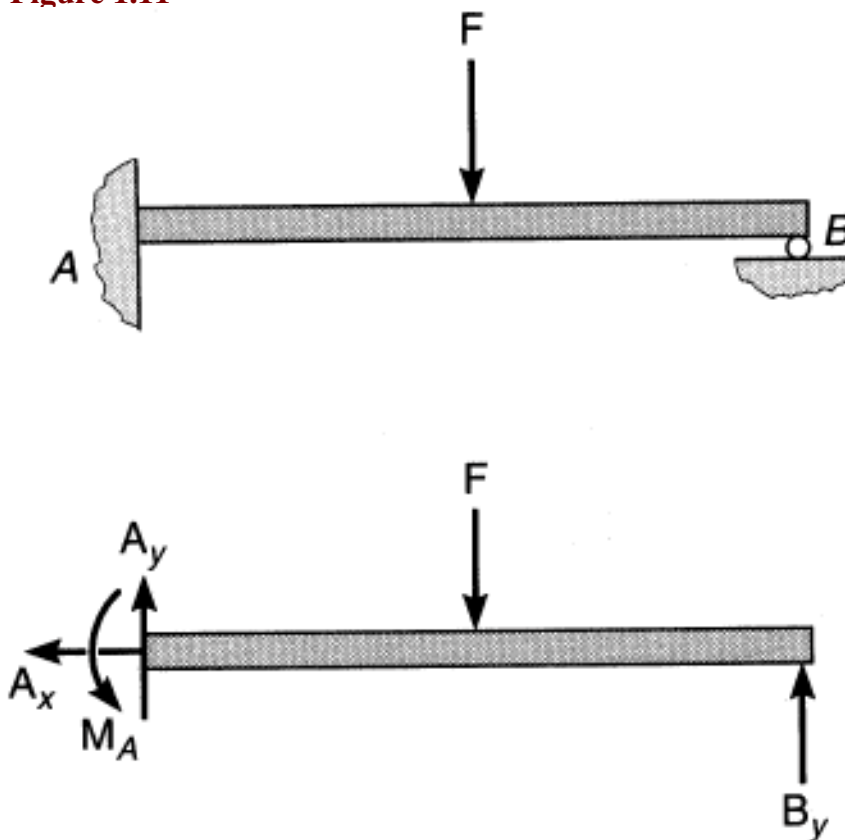
**Table 1.2** Typical Values for Coefficients of Static Friction

Materials	$\mu_s$
Metal on ice	0.03–0.05
Wood on wood	0.30–0.70
Leather on wood	0.20–0.50
Leather on metal	0.30–0.60
Aluminum on aluminum	1.10–1.70

## Constraints

Equilibrium of a body is ensured not only by satisfying the equations of equilibrium, but also by its being properly held or constrained by its supports. If a body has more supports than are needed for equilibrium, it is referred to as *statically indeterminate*, since there will be more unknowns than equations of equilibrium. For example, the free-body diagram of the beam in Fig. 1.11 shows there are four unknown support reactions,  $A_x$ ,  $A_y$ ,  $M_A$ , and  $B_y$ , but only three equations of equilibrium are available for solution [Eq. (1.13)]. The additional equation needed requires knowledge of the physical properties of the body and deals with the mechanics of deformation, which is discussed in subjects such as mechanics of materials.

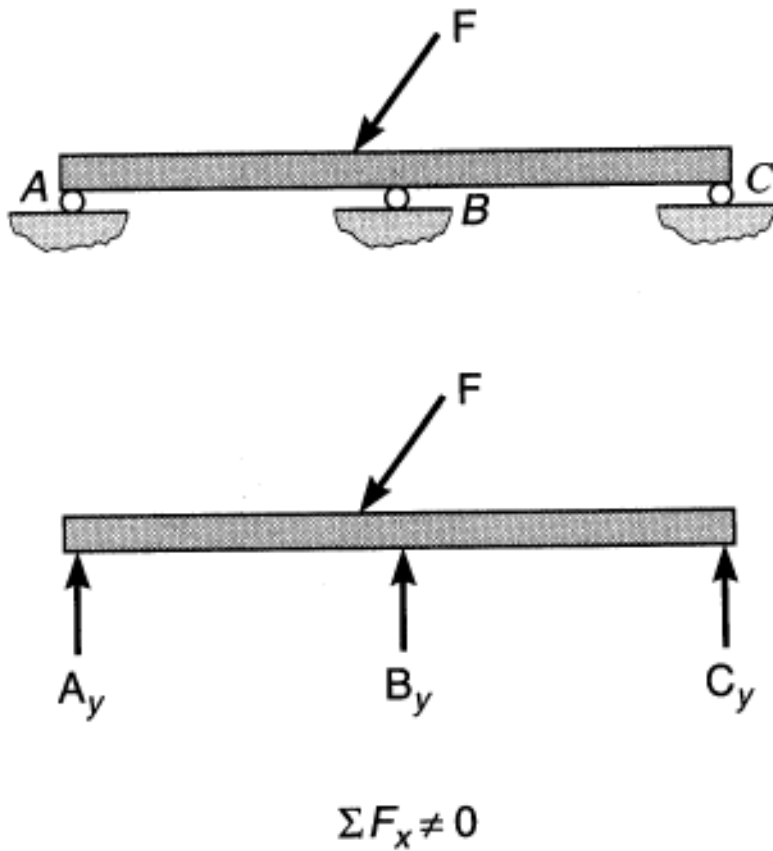
**Figure 1.11**



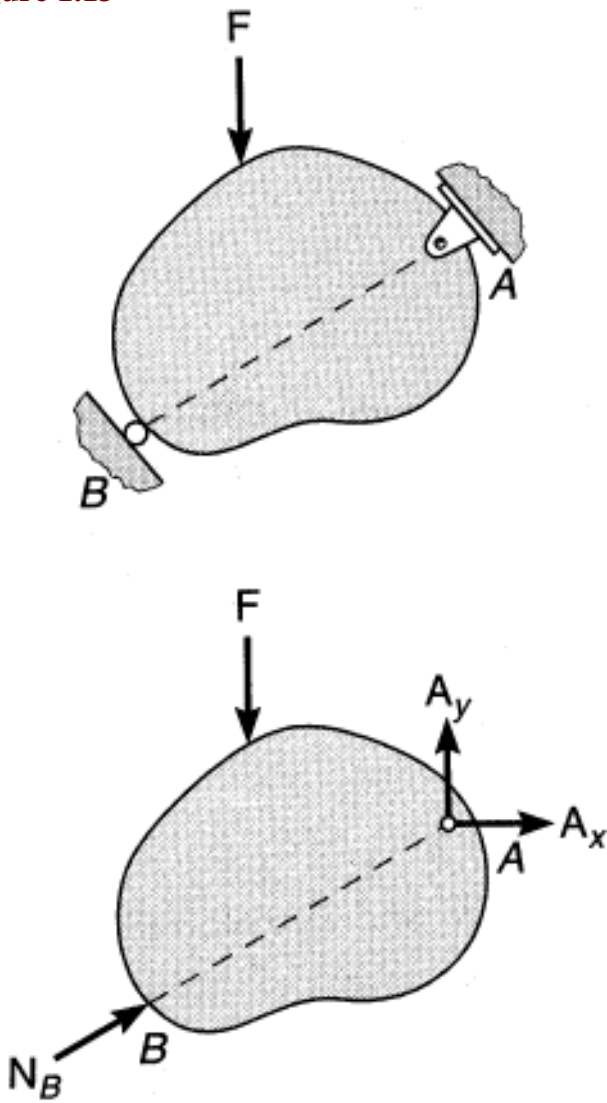


A body may be improperly constrained by its supports. When this occurs, the body becomes unstable and equilibrium cannot be maintained. Either of two conditions may cause this to occur—namely, when the reactive forces are all parallel (Fig. 1.12) or when they are concurrent (Fig. 1.13).

**Figure 1.12**



**Figure 1.13**

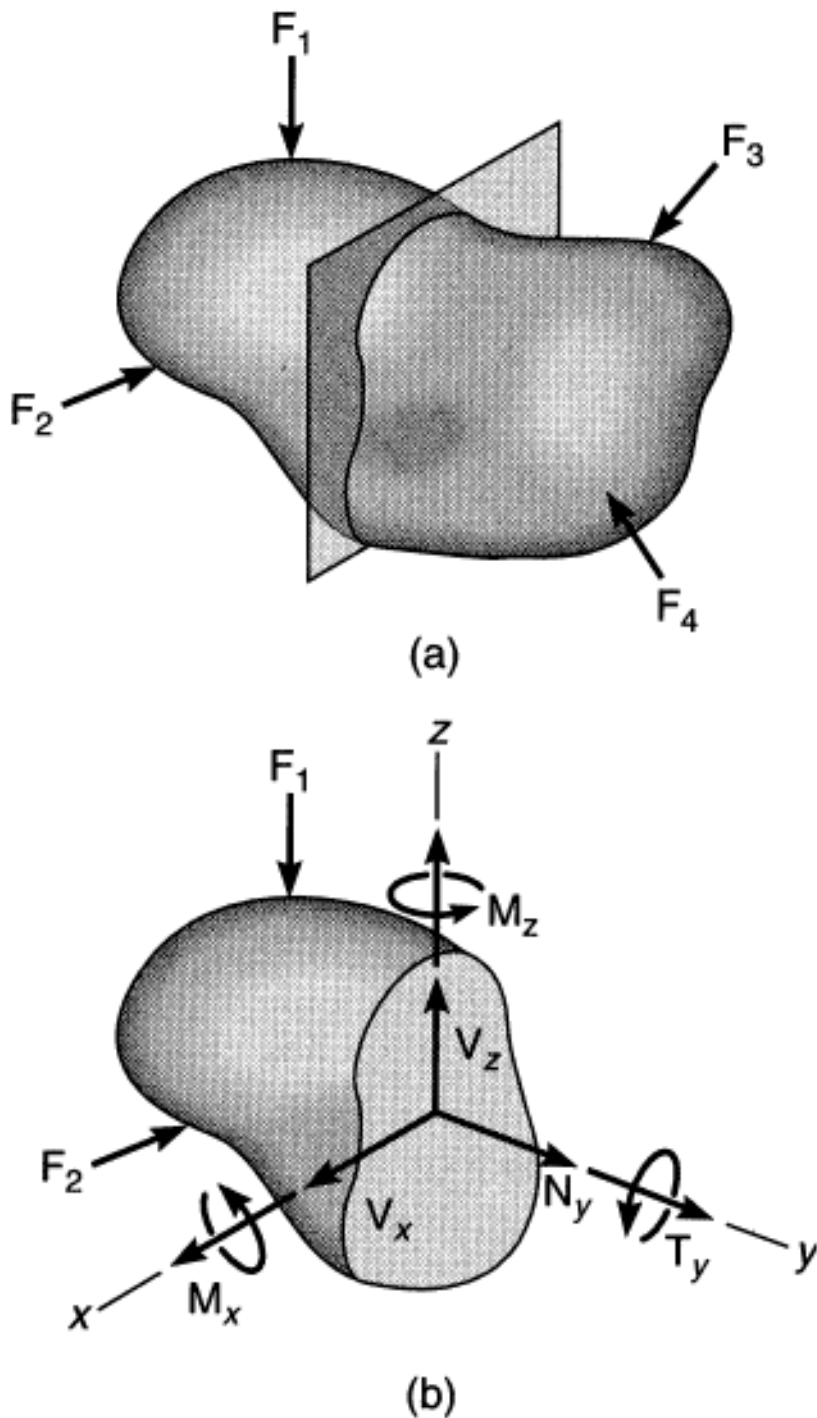


In summary, then, if the number of reactive forces that restrain the body is a minimum—and these forces are not parallel or concurrent—the problem is statically determinate and the equations of equilibrium are sufficient to determine all the reactive forces.

## Internal Loadings

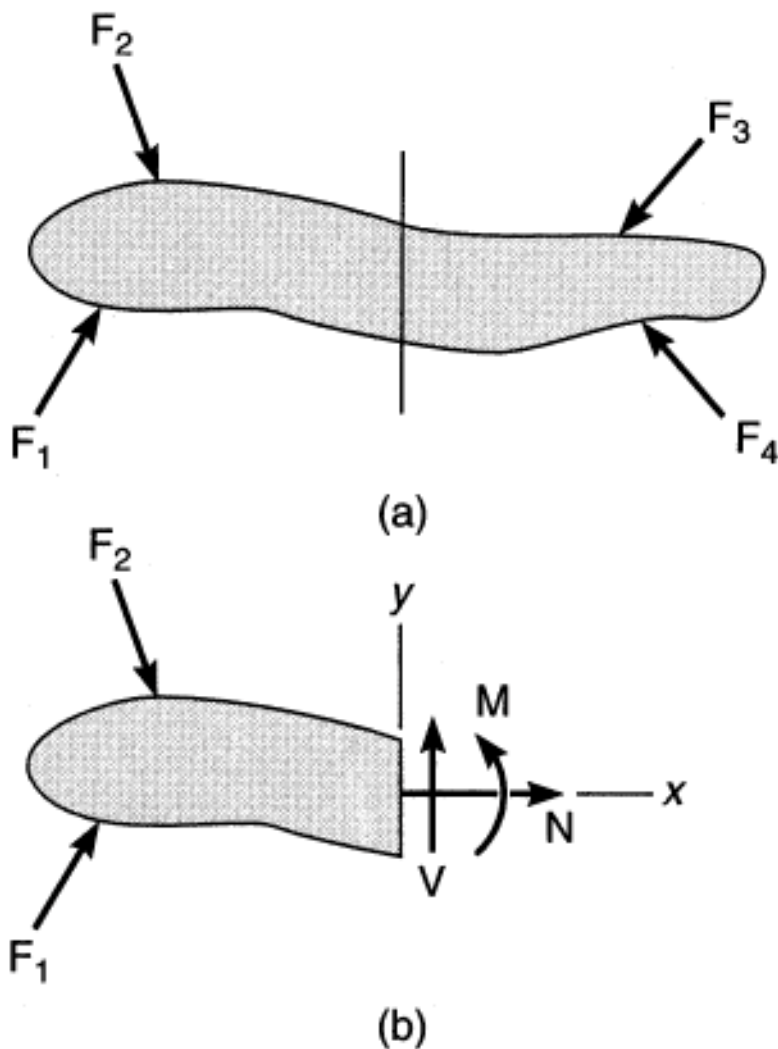
The equations of equilibrium can also be used to determine the internal resultant loadings in a member, provided the external loads are known. The calculation is performed using the **method of sections**, which states that if a body is in equilibrium, then so is any segment of the body. For example, if an imaginary section is passed through the body in [Fig. 1.14\(a\)](#), separating it into two parts, the free-body diagram of the left part is shown in [Fig. 1.14\(b\)](#). Here the six internal resultant components are "exposed" and can be determined from the six equations of equilibrium given by Eq. (1.12). These six components are referred to as the normal force,  $N_y$ , the shear-force components,  $V_x$  and  $V_z$ , the torque or twisting moment,  $T_y$ , and the bending-moment components,  $M_x$  and  $M_z$ .

**Figure 1.14**



If only coplanar loads act on the body [Fig. 1.15(a)], then only three internal resultant loads occur [Fig. 1.15(b)], namely, the normal force,  $N$ , the shear force,  $V$ , and the bending moment,  $M$ . Each of these loadings can be determined from Eq. (1.13). Once these internal resultants have been computed, the actual load distribution over the sectioned surface involves application of the theory related to mechanics of materials.

**Figure 1.15**

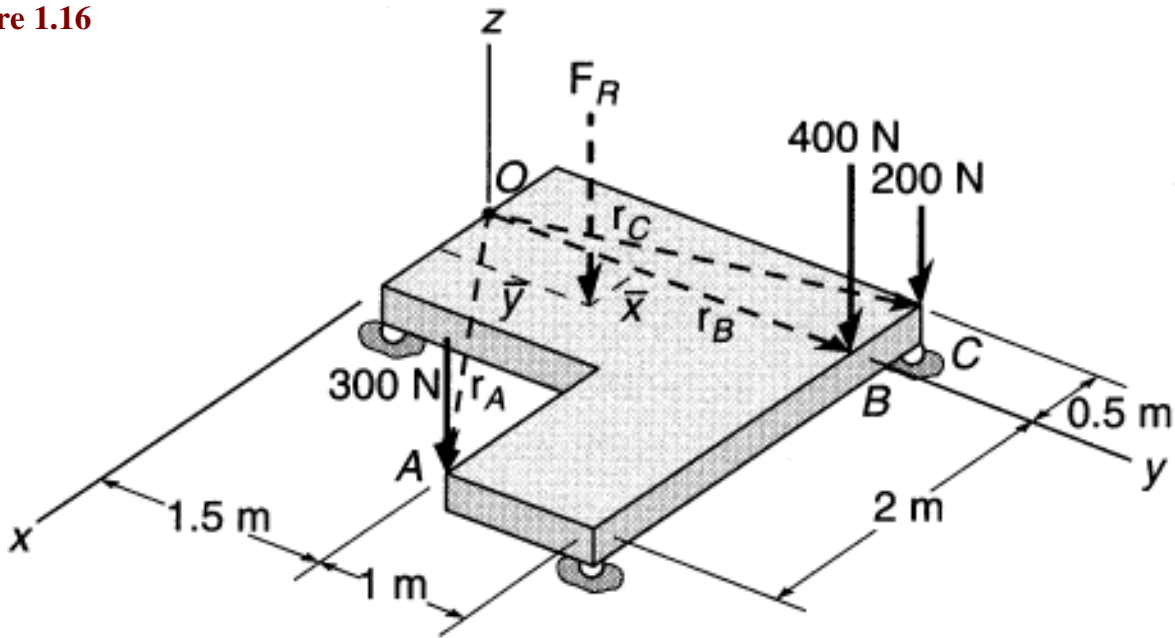


## Numerical Applications

The following examples illustrate application of most of the principles discussed above. Solution of any problem generally requires first establishing a coordinate system, then representing the data on a diagram, and finally applying the necessary equations for solution.

**Example 1.1.** Simplify the system of three parallel forces acting on the plate in [Fig. 1.16](#) to a single resultant force and specify where the force acts on the plate.

**Figure 1.16**



**Solution.** First Eqs. (1.6) and (1.7) are applied in order to replace the force system by a single resultant force and couple moment at point  $O$ .

$$\begin{aligned}
 \mathbf{F}_R &= \Sigma \mathbf{F} & \mathbf{F}_R &= -300\mathbf{k} - 400\mathbf{k} - 200\mathbf{k} = \{-900\mathbf{k}\} \text{ N} & \text{Ans.} \\
 \mathbf{M}_{R_O} &= \Sigma \mathbf{M}_O & \mathbf{M}_{R_O} &= \mathbf{r}_A \times (-300\mathbf{k}) + \mathbf{r}_B \times (-400\mathbf{k}) + \mathbf{r}_C \times (-200\mathbf{k}) \\
 & & &= (2\mathbf{i} + 1.5\mathbf{j}) \times (-300\mathbf{j}) + (2.5\mathbf{j}) \times (-400\mathbf{k}) \\
 & & &\quad + (-0.5\mathbf{i} + 2.5\mathbf{j}) \times (-200\mathbf{k}) \\
 & & &= \{-1950\mathbf{i} + 500\mathbf{j}\} \text{ N} \cdot \text{m}
 \end{aligned}$$

Since the forces are parallel, note that as expected  $\mathbf{F}_R$  is perpendicular to  $\mathbf{M}_{R_O}$ .

The two components of  $\mathbf{M}_{R_O}$  can be eliminated by moving  $\mathbf{F}_R$  along the respective  $y$  and  $x$  axes an amount:

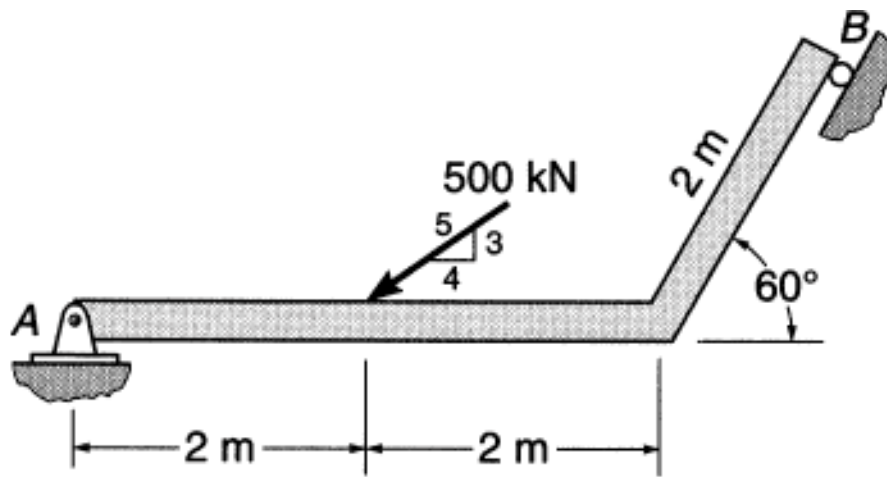
$$\bar{x} = M_{Oy} / F_R = 500 \text{ N} \cdot \text{m} / 900 \text{ N} = 0.556 \text{ m} \quad \text{Ans.}$$

$$\bar{y} = M_{Ox} / F_R = 1900 \text{ N} \cdot \text{m} / 900 \text{ N} = 2.17 \text{ m} \quad \text{Ans.}$$

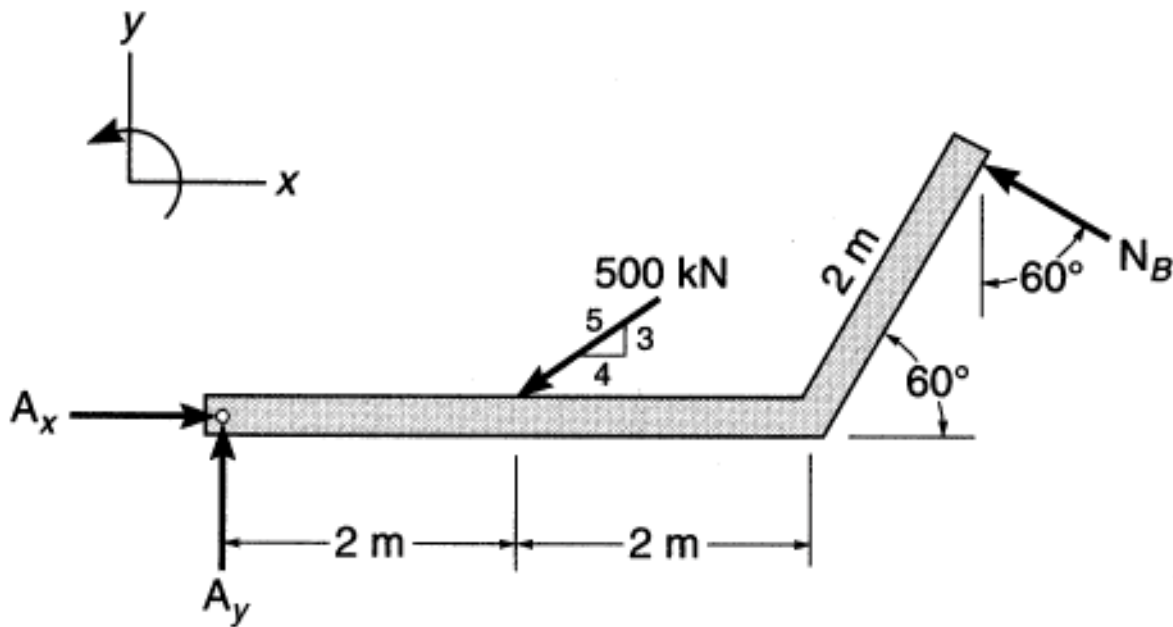
Both coordinates are positive since  $\mathbf{F}_R$  acting at  $\mathbf{r} = \{0.556\mathbf{i} + 2.17\mathbf{j}\} \text{ m}$  will produce the required moment  $\mathbf{M}_{R_O} = \mathbf{r} \times \mathbf{F}_R$ .

**Example 1.2.** Determine the reactions at the supports for the beam shown in Fig. 1.17(a).

**Figure 1.17**



(a)



(b)

**Solution.** Using Table 1.1, the free-body diagram for the beam is shown in Fig. 1.17(b). The problem is statically determinate. The reaction  $N_B$  can be found by using the principle of moments and summing moments about point A to eliminate  $A_x$  and  $A_y$ . Applying Eq. (1.13) with reference to the coordinate system shown gives

$$\begin{aligned}\Sigma M_A = 0 \quad & - 500 \text{ N}(3/5)(2 \text{ m}) + N_B \cos 60^\circ(4 \text{ m} + 2 \cos 60^\circ \text{ m}) \\ & + N_B \sin 60^\circ(2 \sin 60^\circ \text{ m}) = 0\end{aligned}$$

$$N_B = 150 \text{ N} \quad \text{Ans.}$$

$$\Sigma F_x = 0 \quad A_x - 500 \text{ N}(4/5) - 150 \sin 60^\circ \text{ N} = 0$$

$$A_x = 530 \text{ N} \quad \text{Ans.}$$

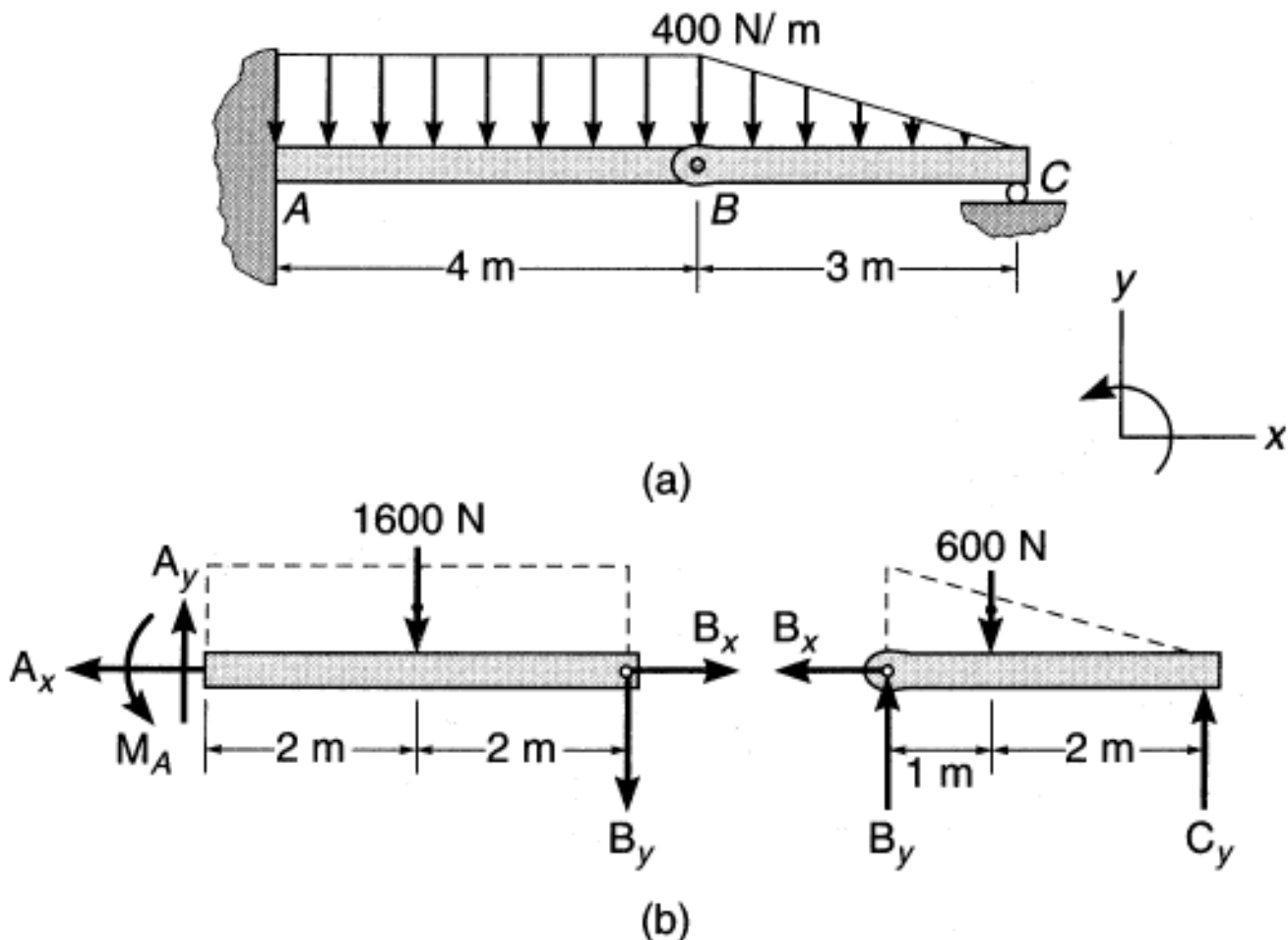
$$\Sigma F_y = 0 \quad A_y - 500 \text{ N}(3/5) + 150 \cos 60^\circ \text{ N} = 0$$

$$A_y = 225 \text{ N} \quad \text{Ans.}$$

Since the answers are all positive, the assumed sense of direction of the once unknown reactive forces are shown correctly on the free-body diagram.

**Example 1.3.** The compound beam shown in Fig. 1.18(a) consists of two segments,  $AB$  and  $BC$ , which are pinned together at  $B$ . Determine the reactions on the beam at the supports.

**Figure 1.18**



**Solution.** The free-body diagrams of both segments of the beam are shown in Fig. 1.18(b). Notice how the principle of action—equal but opposite reaction, Newton's third law—applies to the two force components at  $B$ . Also, the distributed loading has been simplified to resultant forces, determined from the area under each loading diagram and passing through the centroid or geometric center of each area.

The six unknowns are determined by applying Eq. (1.13) to each segment. For segment  $BC$ :

$$\Sigma F_x = 0 \quad B_x = 0 \quad Ans.$$

$$\Sigma M_B = 0 \quad -600 \text{ N}(1 \text{ m}) + C_y(3 \text{ m}) = 0$$

$$C_y = 200 \text{ N} \quad Ans.$$

$$\Sigma F_y = 0 \quad B_y - 600 \text{ N} + 200 \text{ N} = 0$$

$$B_y = 400 \text{ N} \quad Ans.$$

For segment  $AB$ :

$$\Sigma F_x = 0 \quad A_x = 0 \quad Ans.$$

$$\Sigma F_y = 0 \quad A_y - 1600 \text{ N} - 400 \text{ N} = 0$$

$$A_y = 2000 \text{ N} \quad Ans.$$

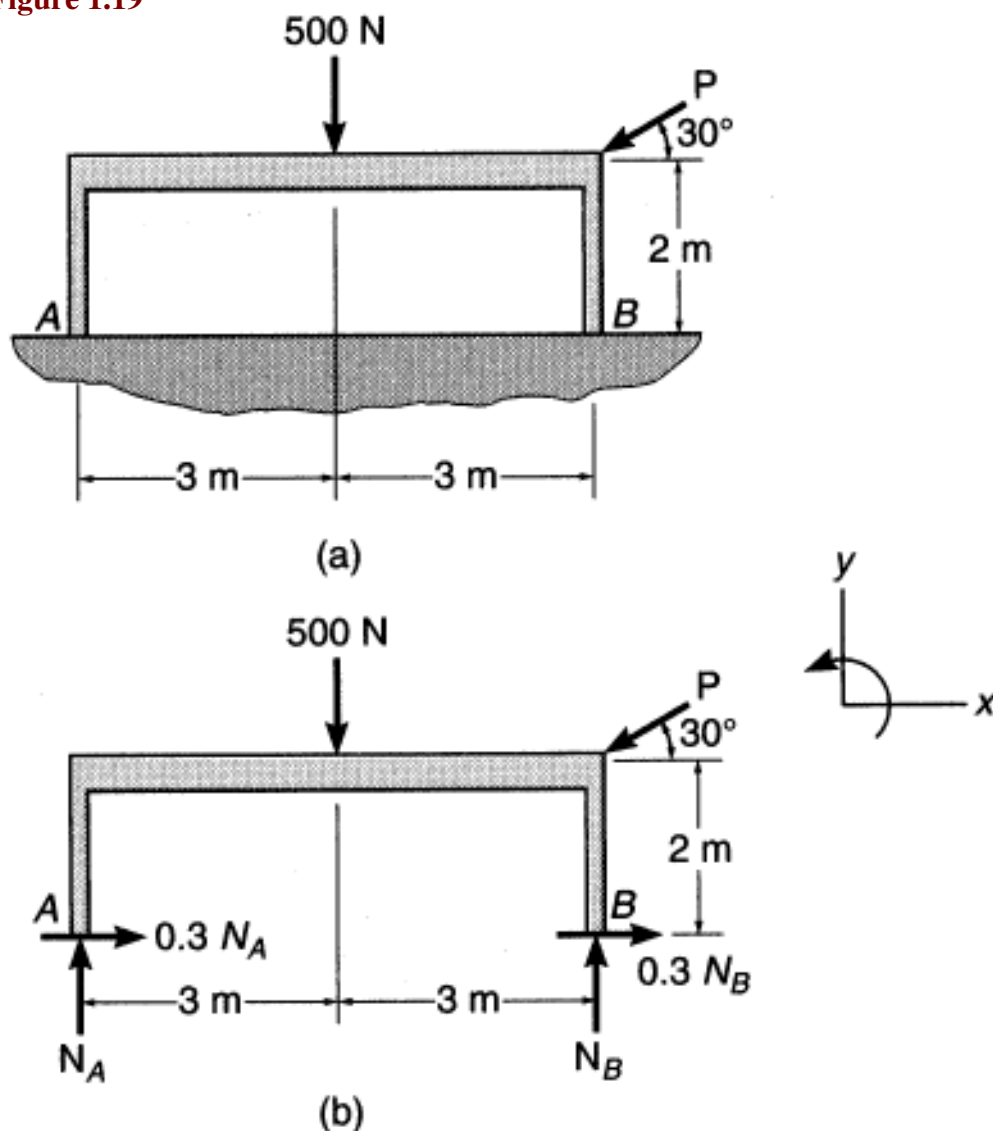
$$\Sigma M_A = 0 \quad M_A = 1600 \text{ N}(2 \text{ m}) - 400 \text{ N}(4 \text{ m}) = 0$$

$$M_A = 4800 \text{ N} \cdot \text{m} \quad Ans.$$

**Example 1.4.** The table in [Fig. 1.19\(a\)](#) rests on a rough surface for which  $\mu_s = 0.3$ . If it supports a load of 500 N, determine the largest magnitude of force **P** that can be applied before it begins to move.



**Figure 1.19**



**Solution.** The free-body diagram is shown in Fig. 1.19(b). Since the maximum force  $P$  is to be determined, slipping must impend at both  $A$  and  $B$ . Therefore, the friction equation  $F = \mu_s N$  applies at these points. There are three unknowns. Applying the equations of equilibrium yields

$$\Sigma M_B = 0 \quad -N_A(6 \text{ m}) + 500 \text{ N}(3 \text{ m}) + P \cos 30^\circ(2 \text{ m}) = 0$$

$$\Sigma F_x = 0 \quad 0.3N_A + 0.3N_B - P \cos 30^\circ = 0$$

$$\Sigma F_y = 0 \quad N_A + N_B - 500 \text{ N} - P \sin 30^\circ = 0$$

Solving,

$$P = 209 \text{ N}$$

$$N_A = 310 \text{ N} \quad \text{Ans.}$$

$$N_B = 294 \text{ N}$$

Since  $N_A$  and  $N_B$  are both positive, the forces of the floor push up on the table as shown on the free-body diagram, and the table remains in contact with the floor.

## Defining Terms

**Concurrent forces:** Forces that act through the same point.

**Couple:** Two forces that have the same magnitude and opposite directions, and do not have the same line of action. A couple produces rotation with no translation.

**Free-body diagram:** A diagram that shows the body "free" from its surroundings. All possible loads and relevant dimensions are labeled on it.

**Friction:** A force of resistance caused by one surface on another.

**Method of sections:** This method states that if a body is in equilibrium, any sectioned part of it is also in equilibrium. It is used for drawing the free-body diagram to determine the internal loadings in any region of a body.

**Parallelogram law:** The method of vector addition whereby two vectors, called *components*, are joined at their tails; parallel lines are then drawn from the head of each vector so that they intersect at a common point forming the adjacent sides of a parallelogram. The resultant vector is the diagonal that extends from the tails of the component vectors to the intersection of the lines.

**Principle of moments:** This concept states that the moment of the force about a point is equal to the sum of the moments of the force's components about the point.

**Principle of transmissibility:** A property of a force stating that the force can act at any point along its line of action and produce the same external effects on a body.

**Weight:** The gravitational attraction of the earth on the mass of a body, usually measured at sea level and 45° latitude.

**Wrench:** A force and collinear moment. The effect is to produce both a push and simultaneous twist.

## Reference

Hibbeler, R. C. 1995. *Engineering Mechanics: Statics*, 7th ed. Prentice Hall, Englewood Cliffs, NJ.

## Further Information

Many textbooks are available for the study of statics and they can be found in any engineering library.

Pilkey W. D., Kitis L. "Centroids and Distributed Forces"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Centroids and Distributed Forces

---

- 2.1 Centroid of a Plane Area
- 2.2 Centroid of a Volume
- 2.3 Surface Forces
- 2.4 Line Forces
- 2.5 Calculation of Surface Area and Volume of a Body with Rotational Symmetry
- 2.6 Determination of Centroids

**Walter D. Pilkey**

*University of Virginia*

**L. Kitis**

*University of Virginia*

## 2.1 Centroid of a Plane Area

---

Any set of forces acting on a rigid body is reducible to an equivalent force-couple system at any selected point  $O$ . This force-couple system, which consists of a force  $\mathbf{R}$  equal to the vector sum of the set of forces and a couple of moment equal to the moment  $M_O$  of the forces about the point  $O$ , is equivalent to the original set of forces as far as the statics and dynamics of the entire rigid body are concerned. In particular, concurrent, coplanar, or parallel forces can always be reduced to a single equivalent force by an appropriate choice of the point  $O$ .

Consider, for example, a distributed load  $p(x)$  acting on a straight beam (Fig. 2.1). If  $p(x)$  has units of force per length, the differential increment of force is  $dR = p(x) dx$  and the total load  $R$  is found by integration:

$$R = \int p(x) dx \quad (2.1)$$

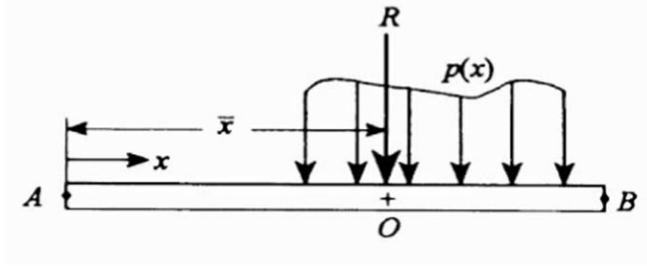
A point  $O$  is to be chosen such that the distributed load  $p(x)$  is equivalent to a single resultant force of magnitude  $R$  acting at  $O$ . This requires that the moment of  $R$  about any point—say, point  $A$ —be the same as the moment of the load  $p(x)$  about that point. Therefore the distance  $\bar{x}$  between  $A$  and  $O$  is given by

$$\bar{x} R = \int x p(x) dx = \int x dR \quad (2.2)$$

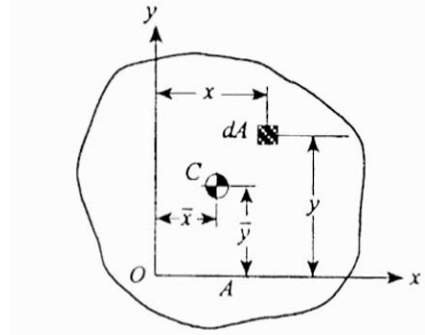
The line of action of the resultant force  $R$  is a vertical line drawn at a distance  $\bar{x}$  from point  $A$ . At

any point on the line of action of  $R$ , the force-couple equivalent of the load  $p(x)$  reduces to a single force. This is a consequence of the *principle of transmissibility*, which states that moving a force along its line of action leaves the conditions of equilibrium or of motion of a rigid body unchanged.

**Figure 2.1** Distributed load on a beam.



**Figure 2.2** Centroid of an area.



Given a plane area  $A$ , the point with coordinates  $\bar{x}$  and  $\bar{y}$  defined by

$$\bar{x}A = \int_A x \, dA \quad \bar{y}A = \int_A y \, dA \quad (2.3)$$

is known as the **centroid** of the area  $A$ . In the calculation of the resultant force  $R$  on a beam, the increment of force  $dR = p(x) \, dx$  is a differential element of area under the load curve  $p(x)$  and the length  $\bar{x}$  given by Eq. (2.2) is the  $x$  coordinate of the centroid of the area under the load curve. Thus, a distributed load on a beam can be replaced by a single force whose magnitude is equal to the area under the load curve  $p(x)$  and whose line of action passes through the centroid of the area under  $p(x)$ .

Suppose that the homogeneous plate of uniform thickness  $t$  shown in Fig. 2.2 is in a uniform and parallel field of force due to the earth's gravitational attraction in the negative  $y$  direction. The resultant of the gravitational forces is a single force in the same direction. The magnitude of this force, called the *weight*, is

$$W = \gamma t A \quad (2.4)$$

where  $\gamma$  is the weight per unit volume of the material and  $A$  is the area of the plate. Since the gravitational forces and their single force equivalent must have the same moment about any point

$$\int_A x \gamma t \, dA = M_O = \bar{x}W = \bar{x} \gamma t A \quad (2.5)$$

or, canceling the constant common factor  $\gamma t$  on both sides,

$$\bar{x} A = \int_A \bar{x} dA \quad (2.6)$$

Similarly, if the earth's attraction is assumed to be in the  $\bar{x}$  direction, the line of action of the weight  $W$ , acting in the same direction, is specified by the coordinate  $\bar{y}$ , given by

$$\bar{y} A = \int_A \bar{y} dA \quad (2.7)$$

The coordinates  $\bar{x}$  and  $\bar{y}$  locate the centroid of the area  $A$  of the plate. Thus, the centroid and the **center of gravity** of a homogeneous plate in a parallel and uniform gravitational field are coincident.

It can be shown that the coordinates  $\bar{x}$  and  $\bar{y}$  given by Eq. (2.3) define the same geometric point regardless of the choice of coordinate axes. Thus, the location of the centroid is independent of any particular choice of orientations for the axes and of the choice of origin. As a consequence, the centroid is a well-defined intrinsic geometric property of any object having a rigid shape. [Table 3.1](#) in **Chapter 3** lists the centroids of some common shapes.

## 2.2 Centroid of a Volume

If a rigid body of volume  $V$  is subjected to a distributed force of intensity  $\mathbf{f}$  (force per unit volume), the force-couple equivalent of this distributed force at an arbitrarily selected origin  $O$  of coordinates can be determined from

$$\mathbf{R} = \int_V \mathbf{f} dV \quad (2.8)$$

$$\mathbf{M}_O = \int_V \mathbf{r} \times \mathbf{f} dV \quad (2.9)$$

where  $\mathbf{r}$  is the position vector of volume elements  $dV$  measured from point  $O$ . In particular, if only one component of  $\mathbf{f}$  is nonzero, the resulting parallel system of distributed body forces acting on the body is reducible to a single equivalent force. For example, with  $f_x = f_y = 0$  and  $f_z$  nonzero, the force-couple equivalent at  $O$  has the components

$$R_x = 0 \quad R_y = 0 \quad R_z = \int_V f_z dV \quad (2.10)$$

and

$$M_x = \int_V y f_z dV \quad M_y = - \int_V x f_z dV \quad M_z = 0 \quad (2.11)$$

The single force equivalent of this force-couple system is found by moving the force  $R_z$  from point  $O$  to a point in the  $xy$  plane such that the moment of  $R_z$  about  $O$  is equal to the moment of the force-couple system about  $O$ . The coordinates  $\bar{x}$  and  $\bar{y}$  of this point are therefore given by

$$\bar{x}R_z = \int_V x f_z dV \quad \bar{y}R_z = \int_V y f_z dV \quad (2.12)$$

The arbitrary choice of  $z = 0$  for the point of application of the resultant  $R_z$  is permissible because  $R_z$  can be slid along its line of action according to the principle of transmissibility.

If the rigid body is homogeneous and  $f_z$  is its specific weight in a uniform gravitational field,  $f_z dV$  is an incremental weight  $dW$  and  $R_z$  is the total weight  $W$  of the body. Equation (2.12) becomes

$$\bar{x}W = \int_V x dW \quad \bar{y}W = \int_V y dW \quad (2.13)$$

However, since the specific weight  $f_z$  is constant and

$$dW = f_z dV \quad W = f_z V \quad (2.14)$$

Eq. (2.13) can be written as

$$\bar{x}V = \int_V x dV \quad \bar{y}V = \int_V y dV \quad (2.15)$$

Similarly, if the body is placed in a uniform gravitational field that exerts a force in the  $x$  direction only, then the line of action of the single equivalent force penetrates the  $yz$  plane at  $\bar{y}$  defined in Eq. (2.15) and  $\bar{z}$  given by

$$\bar{z}V = \int_V z dV \quad (2.16)$$

The point with coordinates  $\bar{x}$ ,  $\bar{y}$ ,  $\bar{z}$  defined in Eqs. (2.15) and (2.16) is the centroid  $C$  of the volume  $V$  of the body. The centroid and the center of gravity of a homogeneous body in a uniform and parallel gravitational field are coincident points. If the body is not homogeneous, Eqs. (2.15) and (2.16) cannot be used to find the center of gravity; they still define the centroid, which is an intrinsic geometric property of any rigid shape. The **center of mass** of a rigid body is defined by

$$\bar{\mathbf{r}}_m = \int \mathbf{r} dm \quad (2.17)$$

where  $m$  is the mass. The center of mass depends solely on the mass distribution and is independent of the properties of the gravitational field in which the body may be placed. In a uniform and parallel gravitational field, however,  $dW = g dm$  where  $g$  is the gravitational

constant, so that the center of mass coincides with the center of gravity. If, in addition, the body is homogeneous, the  $x, y, z$  components of the vector  $\bar{\mathbf{r}}$  given by Eq. (2.17) are  $\bar{x}, \bar{y}, \bar{z}$  defined in Eqs. (2.15) and (2.16). In this case, the centroid, the center of mass, and the center of gravity are coincident. The mass centers of some homogeneous solids are listed in [Table 3.2, Chapter 3](#).

## 2.3 Surface Forces

---

Suppose the distributed load is a force per unit area  $\mathbf{p}$  and let  $\mathbf{p}$  be a function of the position vector  $\mathbf{r}$  with respect to an origin  $O$ . For a surface of area  $A$ , the resultant of the distributed surface forces is

$$\mathbf{R} = \int_A \mathbf{p} \, dA \quad (2.18)$$

where  $dA$  is a differential surface element. The resultant moment of the distributed surface forces with respect to the reference point  $O$  is given by

$$\mathbf{M}_O = \int_A \mathbf{r} \times \mathbf{p} \, dA \quad (2.19)$$

The line of action of a single force equivalent for a parallel surface force distribution is determined as in the case of volume forces. For example, with a plane surface in the  $xy$  plane and with  $p_z$  as the only nonzero force, the line of action of the resultant force intersects the  $xy$  plane at the point whose coordinates are  $\bar{x}, \bar{y}$  given by

$$\bar{x} = \frac{\int_A x p_z \, dA}{\int_A p_z \, dA} \quad \bar{y} = \frac{\int_A y p_z \, dA}{\int_A p_z \, dA} \quad (2.20)$$

## 2.4 Line Forces

---

In general the formula for body forces can be employed for a line. For the special case of a plane curve of length  $l$  in the  $xy$  plane and distributed line force  $p_z$  (force/length) acting in the  $z$  direction, the force resultant is

$$R_z = \int_l p_z(s) \, ds \quad (2.21)$$

where  $s$  is the coordinate along the curve. The moment resultant about the reference point  $O$  is

$$\mathbf{M}_O = \int_l (y p_z \mathbf{i} - x p_z \mathbf{j}) \, ds \quad (2.22)$$



where  $\mathbf{i}, \mathbf{j}$  are unit vectors in the  $x, y$  directions, respectively. The coordinates where the line of action of the single force resultant intersects the  $xy$  plane are

$$\bar{x} = \frac{\int_l x(s) p_z(s) ds}{\int_l p_z(s) ds} \quad \bar{y} = \frac{\int_l y(s) p_z(s) ds}{\int_l p_z(s) ds} \quad (2.23)$$

## 2.5 Calculation of Surface Area and Volume of a Body with Rotational Symmetry

### The two Pappus–Guldin formulas

Pappus of Alexandria was a 3rd-century Greek geometer. Of a collection of eight mathematical books, only a portion survived. This is an informative source on ancient Greek mathematics. Included in this collection is a method for the measurement of a surface area bounded by a spiral on a sphere.

Paul Habakuk Guldin (1577–1643) was a professor of mathematics in several Italian and Austrian Jesuit colleges. In 1635 he revealed a relationship between the volume and the area of a body of revolution.

take advantage of the definition of the centroid to assist in the calculation of the surface area and the volume of a body with rotational symmetry.

When a plane meridian curve (Fig. 2.3) that does not intersect the  $y$  axis is rotated through  $\phi$  radians about the  $y$  axis, a line element  $ds$  generates a surface of area

$$dS = x\phi ds \quad (2.24)$$

The total surface area generated by a meridian curve of length  $l$  is

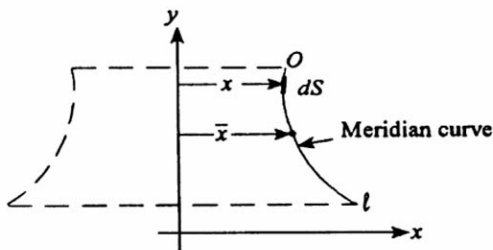
$$S = \phi \int_0^l x ds \quad (2.25)$$

Therefore

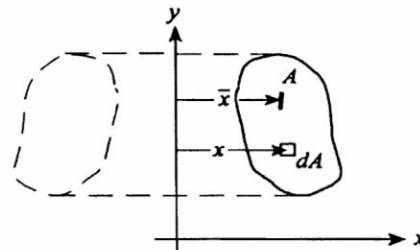
$$S = \phi \bar{x} l \quad (2.26)$$

where  $\bar{x}$  locates the centroid of the meridian curve. This is the first Pappus–Guldin formula.

**Figure 2.3** A surface with rotational symmetry.



**Figure 2.4** A volume with rotational symmetry.



If the meridian curve forms a shell of revolution,  $\phi = 2\pi$  and

$$S = 2\pi \bar{x}l \quad (2.27)$$

When a plane area  $A$  (Fig. 2.4) that is not cut by the  $y$  axis is rotated through an angle  $\phi$  radians about the  $y$  axis, a volume element

$$dV = \phi x \, dA \quad (2.28)$$

is generated and the total volume is given by

$$V = \phi \bar{x} A \quad (2.29)$$

This is the second Pappus–Guldin formula.

For a solid of revolution, with  $\phi = 2\pi$ ,

$$V = 2\pi \bar{x} A \quad (2.30)$$

If  $A$  and  $V$  are known, the centroid of  $A$  can be determined from

$$\bar{x} = \frac{V}{2\pi A} \quad (2.31)$$

## 2.6 Determination of Centroids

When an area or a line has an axis of symmetry, the centroid of the area or line is on that axis. If an area or line has two axes of symmetry, the centroid of the area or line is at the intersection of the axes of symmetry. Thus, the geometric centers of circles, ellipses, squares, rectangles, or lines in the shape of the perimeter of an equilateral triangle are also their centroids. When a volume has a plane of symmetry, the centroid of the volume lies on that plane. When a volume has two planes of symmetry, the centroid of the volume lies on the line of intersection of the planes of symmetry. When a volume has three planes of symmetry intersecting at a point, the point of intersection of the three planes is the centroid of the volume. Thus, the geometric centers of spheres, ellipsoids, cubes, or rectangular parallelepipeds are also their centroids.

Centroids of unsymmetrical areas, lines, or volumes can be determined by direct integration. The Pappus–Guldin formulas can be used to determine the centroid of a plane curve when the area of the surface generated by the curve is known, or to determine the centroid of a plane area when the volume generated by the area is known.

If a body can be divided into  $n$  parts, for which the volumes  $V_i$  and the centroids  $\bar{x}_i, \bar{y}_i, \bar{z}_i$  are known, the centroid of the entire body has coordinates  $\bar{x}, \bar{y}, \bar{z}$ , given by

$$\bar{x}V = \sum_{i=1}^n \bar{x}_i V_i \quad \bar{y}V = \sum_{i=1}^n \bar{y}_i V_i \quad \bar{z}V = \sum_{i=1}^n \bar{z}_i V_i \quad (2.32)$$

where  $V$  is the total volume of the body. The same equations are applicable for a body with cutouts, provided that the volumes of the cutouts are taken as negative numbers.

**Example 2.1.** Determine the location of the centroid for the triangle shown in Fig. 2.5. The triangle is symmetric about the  $y$  axis. Hence the centroid lies on the  $y$  axis and  $\bar{x} = 0$ . To determine  $\bar{y}$ , select an element of area  $dA$  parallel to the  $x$  axis, thus making all points in the element an equal distance from the  $x$  axis. Then  $dA = 2x dy$  and, with  $A = bh/2$ ,

$$\bar{y} = \frac{\int_0^h y(2x dy)}{bh/2} = \frac{4 \int_0^h xy dy}{bh} \quad (2.33)$$

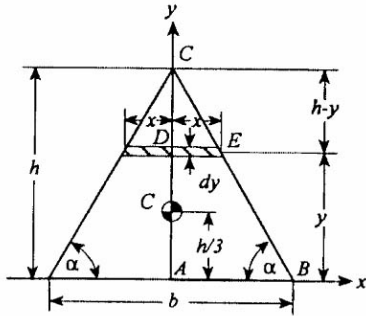
The expression for  $x$  in terms of  $y$  follows from the proportionality of the similar triangles  $ABC$  and  $DEC$ . This provides

$$\frac{x}{h-y} = \frac{b}{2h} \quad \text{or} \quad x = \frac{1}{2} \frac{b}{h}(h-y) \quad (2.34)$$

Finally,

$$\bar{y} = \frac{4 \int_0^h \frac{1}{2} (b/h)(h-y)y dy}{bh} = \frac{2}{h^2} \int_0^h (hy - y^2) dy = \frac{2}{h^2} \left( h \frac{h^2}{2} - \frac{h^3}{3} \right) = \frac{h}{3} \quad (2.35)$$

**Figure 2.5** Centroid of a triangle.



Thus the centroid of the triangular area is on the  $y$  axis at a distance of one-third the altitude from the base of the triangle.

**Example 2.2.** Figure 2.6 shows a complicated shape that is made up of a semicircle, a rectangle, and a triangle. The areas and centroids of the semicircle, rectangle, and triangle are known:

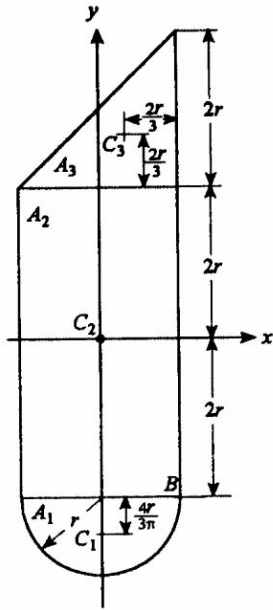
$$A_1 = \frac{\pi r^2}{2} \quad A_2 = 8r^2 \quad A_3 = 2r^2 \quad (2.36)$$

$$C_1 = \left( 0, -\frac{2(2+3\pi)r}{3\pi} \right) \quad C_2 = (0, 0) \quad C_3 = \left( \frac{r}{3}, \frac{8r}{3} \right) \quad (2.37)$$

The  $x$  coordinate  $\bar{x}$  of the centroid of the figure is given by

$$\bar{x}A = \sum_{i=1}^3 \bar{x}_i A_i = \frac{2r^3}{3} \quad (2.38)$$

**Figure 2.6** Centroid of a complicated area.



where  $A$  is the total area

$$A = A_1 + A_2 + A_3 = \frac{(\pi + 20)r^2}{2} \quad (2.39)$$

Therefore,

$$\bar{x} = \frac{4r}{3(\pi + 20)} \quad (2.40)$$

Similarly, the  $y$  coordinate  $\bar{y}$  of the centroid is calculated from

$$\bar{y}A = \sum_{i=1}^3 \bar{y}_i A_i \quad (2.41)$$

which yields

$$\bar{y} = \frac{2(14 - 3\pi)r}{3(\pi + 20)} \quad (2.42)$$

**Example 2.3.** In this example, the  $y$  coordinate of the centroid of the frustum of the right circular cone shown in Fig. 2.7 will be found by treating the frustum as a cone of height  $h$  with a cutout cone of height  $h/2$ . From Table 3.2 of Chapter 3, the  $y$  coordinates of the centroids of these two cones are

$$y_1 = \frac{3h}{4} \quad y_2 = \frac{3h}{8} \quad (2.43)$$

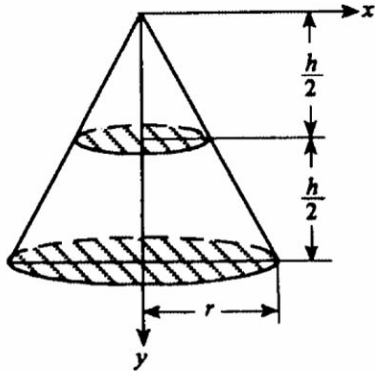
The corresponding volumes are

$$V_1 = \frac{\pi r^2 h}{3} \quad V_2 = \frac{\pi r^2 h}{24} \quad (2.44)$$

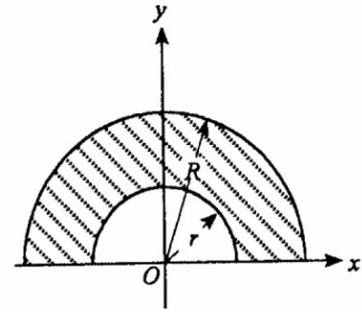
and the  $y$  coordinate of the centroid of the frustum of the cone is

$$\bar{y} = \frac{y_1 V_1 - y_2 V_2}{V_1 - V_2} = \frac{45h}{56} \quad (2.45)$$

**Figure 2.7** Frustum of a cone.



**Figure 2.8** Plane area of Example 2.4.



**Example 2.4.** The centroid of the area enclosed between the semicircles of radius  $R$  and  $r$  and part of the  $x$  axis (see Fig. 2.8) can be found by an application of the second Pappus–Guldin formula. The volume generated by rotating this area through  $2\pi$  is

$$V = \frac{4\pi}{3} (R^3 - r^3) \quad (2.46)$$

Hence

$$\bar{y} = \frac{V}{2\pi\pi(R^2 - r^2)/2} = \frac{4(r^2 + rR + R^2)}{3\pi(r + R)} \quad (2.47)$$

**Example 2.5.** Consider the problem of finding the centroid of the half torus in Table 3.2 of Chapter 3. The half torus is obtained by revolving a circle of radius  $a$  through  $\pi$  radians about the  $z$  axis, and the second Pappus–Guldin formula gives its volume  $V = \pi^2 a^2 R$ . To calculate the  $x$  coordinate  $\bar{x}$  of the centroid, a volume element obtained by taking a horizontal slice of thickness  $dz$  (see Fig. 2.9) can be used. The volume of this element is given by

$$dV = \frac{\pi}{2} (r_o^2 - r_i^2) dz \quad (2.48)$$

where

$$\begin{aligned} r_i &= R - a \cos \theta \\ r_o &= R + a \cos \theta \end{aligned} \quad (2.49)$$

Since  $dz = a \cos \theta d\theta$ , the expression for  $dV$  simplifies to

$$\begin{aligned} r_i &= R - a \cos \theta \\ r_o &= R + a \cos \theta \end{aligned} \quad (2.49)$$

Since  $dz = a \cos \theta d\theta$ , the expression for  $dV$  simplifies to

$$dV = 2\pi a^2 R \cos^2 \theta d\theta \quad (2.50)$$

It is known from Example 2.4 that the  $x$  coordinate of the centroid of the volume element is

$$x = \frac{4(r_i^2 + r_i r_o + r_o^2)}{3\pi(r_i + r_o)} = \frac{a^2(1 + \cos 2\theta) + 6R^2}{3\pi R} \quad (2.51)$$

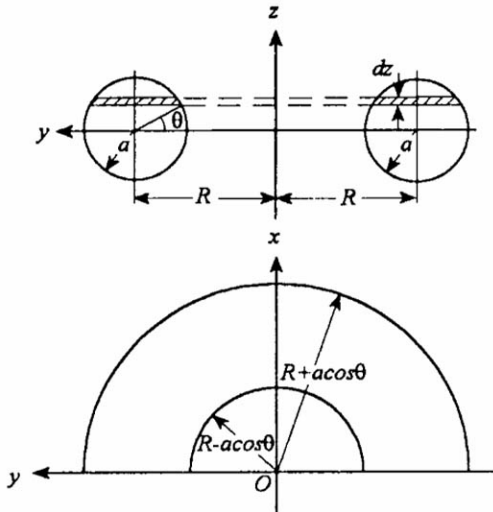
Therefore, the  $x$  coordinate of the centroid of the half torus is

$$\bar{x} = 2 \frac{\int_0^{\pi/2} x dV}{V} = \frac{4}{3\pi^2 R} \int_0^{\pi/2} (a^2(1 + \cos 2\theta) + 6R^2) \cos^2 \theta d\theta \quad (2.52)$$

in which the integration limits extend over the upper half of the volume so that the additional multiplicative factor of 2 is necessary. The result is

$$\bar{x} = \frac{a^2 + 4R^2}{2\pi R} \quad (2.53)$$

**Figure 2.9** Example 2.5.



**Example 2.6.** Figure 2.10 shows a thin homogeneous semicircular cantilever beam of specific weight  $\gamma$  in a uniform gravitational field. Suppose that the only force acting on the beam is the gravitational force of attraction perpendicular to the plane of the figure. To determine the internal forces at any section  $E$  of the beam, the free-body diagram of the circular arc from  $B$  to  $E$  is used. The force distribution on this circular arc is a line force of constant intensity so that Eqs. (2.23), which give the coordinates of a point on the line of action of the single force resultant of a line force, show that the resultant passes through the centroid of the arc. The centroid  $C$  of the arc is at a distance  $\bar{r}$  from the center  $O$ :

$$\bar{r} = \frac{r \sin \theta/2}{\theta/2} \quad (2.54)$$

The single force equivalent of the gravitational forces acting on the arc  $BE$  is the weight  $\gamma r \theta$  acting at the centroid  $C$  in the direction perpendicular to the plane of the figure. The moment arm for the twisting moment  $T$  at  $E$  is  $DE$  and the moment arm for the bending moment  $M$  at  $E$  is  $DC$ . Hence

$$T = \gamma r \theta \left( r - \bar{r} \cos \frac{\theta}{2} \right) \quad (2.55)$$

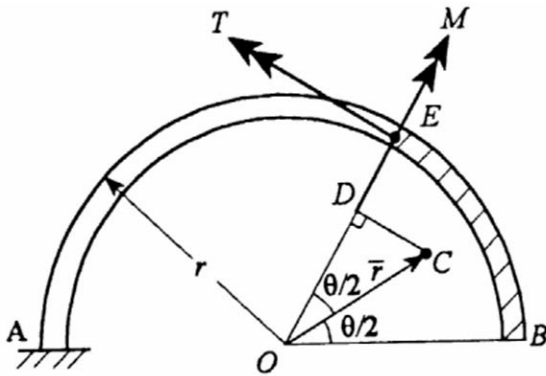
$$M = \gamma r \theta \bar{r} \sin \frac{\theta}{2} \quad (2.56)$$

These expressions may be rewritten as

$$T = \gamma r^2 (\theta - \sin \theta) \quad (2.57)$$

$$M = \gamma r^2 (1 - \cos \theta) \quad (2.58)$$

**Figure 2.10** Semicircular cantilever beam.



## Defining Terms

**Center of gravity:** The point of application of the single force resultant of the distributed gravitational forces exerted by the earth on a rigid body.

**Center of mass:** A unique point of a rigid body determined by the mass distribution. It coincides with the center of gravity if the gravity field is parallel and uniform.

**Centroid:** A unique point of a rigid geometric shape, defined as the point with coordinates  $\bar{x} = \int_V x \, dV/V$ ,  $\bar{y} = \int_V y \, dV/V$ ,  $\bar{z} = \int_V z \, dV/V$ , for a three-dimensional solid of volume  $V$ .

**Pappus–Guldin formulas:** Two formulas that use the definition of the centroid to assist in the calculation of the area of a surface of revolution and the volume of a body of revolution.

## References

- Pilkey, W. D. 1994. *Formulas for Stress, Strain and Structural Matrices*. John Wiley & Sons, New York.
- Pilkey, W. D. and Pilkey, O. H. 1986. *Mechanics of Solids*. Krieger, Malabar, FL.

## Further Information

Consult undergraduate textbooks on statics, dynamics, and mechanics of solids.



Meriam, J. L. "Moments of Inertia"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Moments of Inertia

---

## 3.1 Area Moments of Inertia

Defining Relations

## 3.2 Mass Moments of Inertia

Defining Relations

### J. L. Meriam

*University of California (Retired)*

The **mass moment of inertia**,  $I$ , of a body is a measure of the inertial resistance of the body to rotational acceleration and is expressed by the integral  $I = \int r^2 dm$ , where  $dm$  is the differential element of mass and  $r$  is the perpendicular distance from  $dm$  to the rotation axis. The **area moment of inertia** of a defined area about a given axis is expressed by the integral  $I = \int s^2 dA$ , where  $dA$  is the differential element of area and  $s$  is the perpendicular distance from  $dA$  to a defined axis either in or normal to the plane of the area. The mathematical similarity to mass moment of inertia gives rise to its name. A more fitting but less used term is the *second moment of area*.

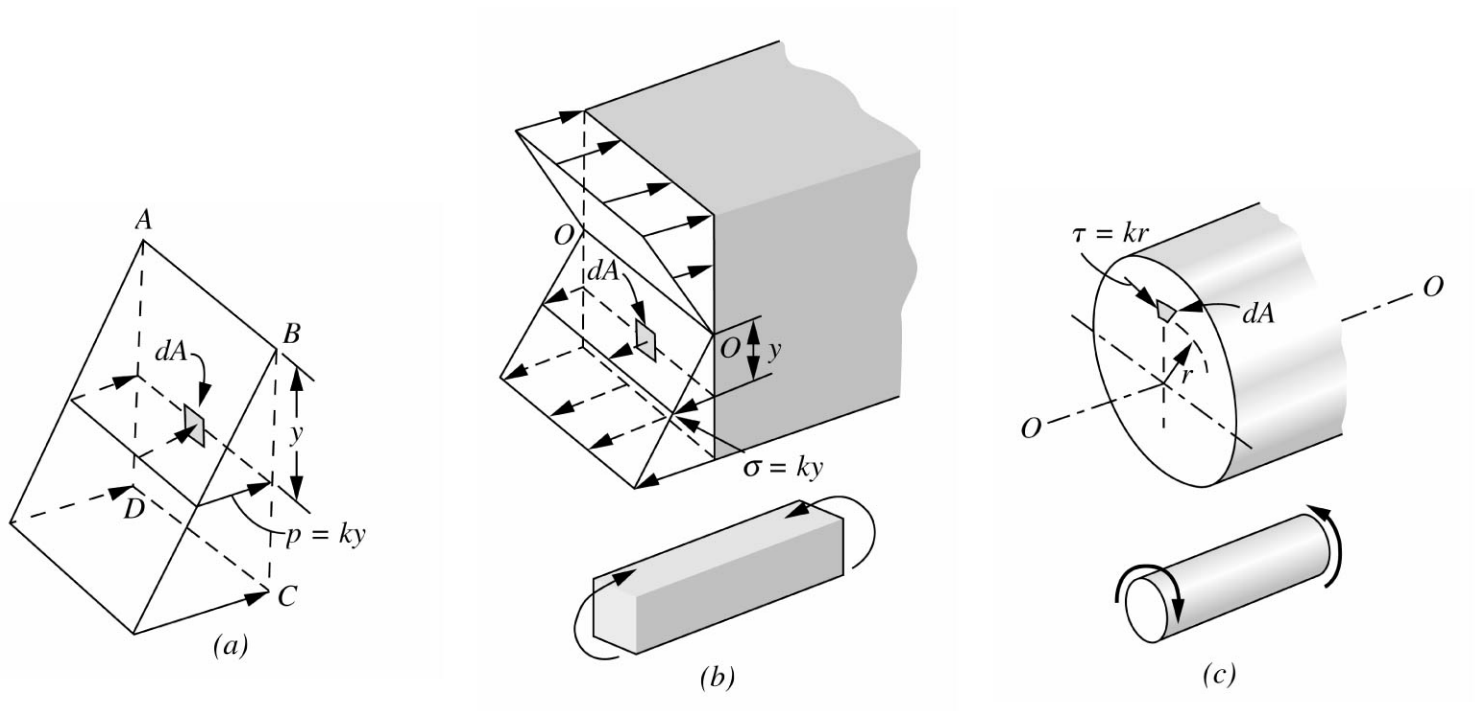
The frequent occurrence of area and mass moments of inertia in mechanics justifies establishing and tabulating their properties for commonly encountered shapes, as given in [Tables 3.1](#) and [3.2](#) at the end of each section.

## 3.1 Area Moments of Inertia

---

[Figure 3.1](#) illustrates the physical origin of the area moment-of-inertia integrals. In [Fig. 3.1\(a\)](#) the surface area  $ABCD$  is subject to a distributed pressure  $p$  whose intensity is proportional to the distance  $y$  from the axis  $AB$ . The moment about  $AB$  that is due to the pressure on the element of area  $dA$  is  $p(p dA) = ky^2 dA$ . Thus the integral in question appears when the total moment  $M = k \int y^2 dA$  is evaluated.

**Figure 3.1** (Source: Meriam, J. L. and Kraige, L. G. 1992. *Engineering Mechanics*, 3rd ed. John Wiley & Sons, New York.)



**Figure 3.2** (Source: Meriam, J. L. and Kraige, L. G. 1992. *Engineering Mechanics*, 3rd ed. John Wiley & Sons, New York.)

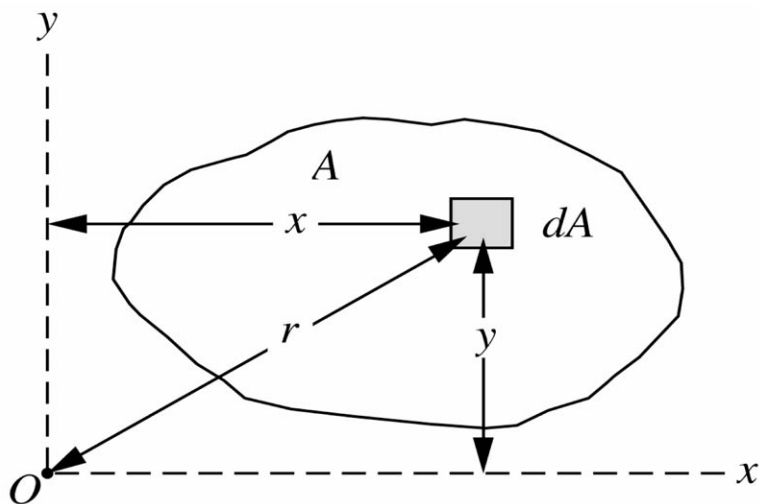


Figure 3.1(b) shows the distribution of stress acting on a transverse section of a simple linear elastic beam bent by equal and opposite couples applied one to each end. At any section of the beam a linear distribution of force intensity or stress  $\sigma$ , given by  $\sigma = ky$ , is present, the stress being positive (tensile) below the axis  $O-O$  and negative (compressive) above the axis. The elemental moment about axis  $O-O$  is  $dM = y(\sigma dA) = ky^2 dA$ . Thus the same integral appears when the total moment  $M = k \int y^2 dA$  is evaluated.

A third example is given in Fig. 3.1(c), which shows a circular shaft subjected to a twist or torsional moment. Within the elastic limit of the material this moment is resisted at each cross section of the shaft by a distribution of tangential or shear stress  $\tau$  that is proportional to the radial distance  $r$  from the center. Thus  $\tau = kr$  and the total moment about the central axis becomes  $M = \int r(\tau dA) = k \int r^2 dA$ . Here the integral differs from that in the preceding two examples in that the area is normal instead of parallel to the moment axis and in that  $r$  is a radial coordinate instead of a rectangular one.

## Defining Relations

### Rectangular and Polar Moments of Inertia

For area  $A$  in the  $xy$  plane, Fig. 3.2, the moments of inertia of the element  $dA$  about the  $x$  and  $y$  axes are, by definition,  $dI_x = y^2 dA$  and  $dI_y = x^2 dA$ , respectively. The moments of inertia of  $A$  about the same axes become

$$\begin{aligned} I_x &= \int y^2 dA \\ I_y &= \int x^2 dA \end{aligned} \quad (3.1)$$

where the integration is carried out over the entire area.

The moment of inertia of  $dA$  about the pole  $O$  ( $z$  axis) is, by definition,  $dI_z = r^2 dA$ , and the moment of inertia of the entire area about  $O$  is

$$I_z = \int r^2 dA \quad (3.2)$$

The expressions defined by Eq. (3.1) are known as *rectangular* moments of inertia, whereas the expression of Eq. (3.2) is known as the *polar* moment of inertia. (In the literature the polar moment of inertia is sometimes denoted by the symbol  $J$ .) Because  $x^2 + y^2 = r^2$ , it follows that

$$I_z = I_x + I_y \quad (3.3)$$

A polar moment of inertia for an area whose boundaries are more simply described in rectangular coordinates than in polar coordinate is easily calculated using Eq. (3.3).

Because the area moment of inertia involves distance squared, it is always a positive

quantity for a positive area. (A hole or void may be considered a negative area.) In contrast, the first moment of area  $\int y \, dA$  involves distance to the first power, so it can be positive, negative, or zero.

The dimensions of moments of inertia of areas are  $L^4$ , where  $L$  stands for the dimension of length. The SI units for moments of inertia of areas are expressed in quartic meters ( $\text{m}^4$ ) or quartic millimeters ( $\text{mm}^4$ ). The U.S. customary units are quartic feet ( $\text{ft}^4$ ) or quartic inches ( $\text{in.}^4$ ).

Rectangular coordinates should be used for shapes whose boundaries are most easily expressed in these coordinates. Polar coordinates will usually simplify problems where the boundaries are expressed in  $r$  and  $\theta$ . The choice of an element of area that simplifies the integration as much as possible is equally important.

### Radius of Gyration

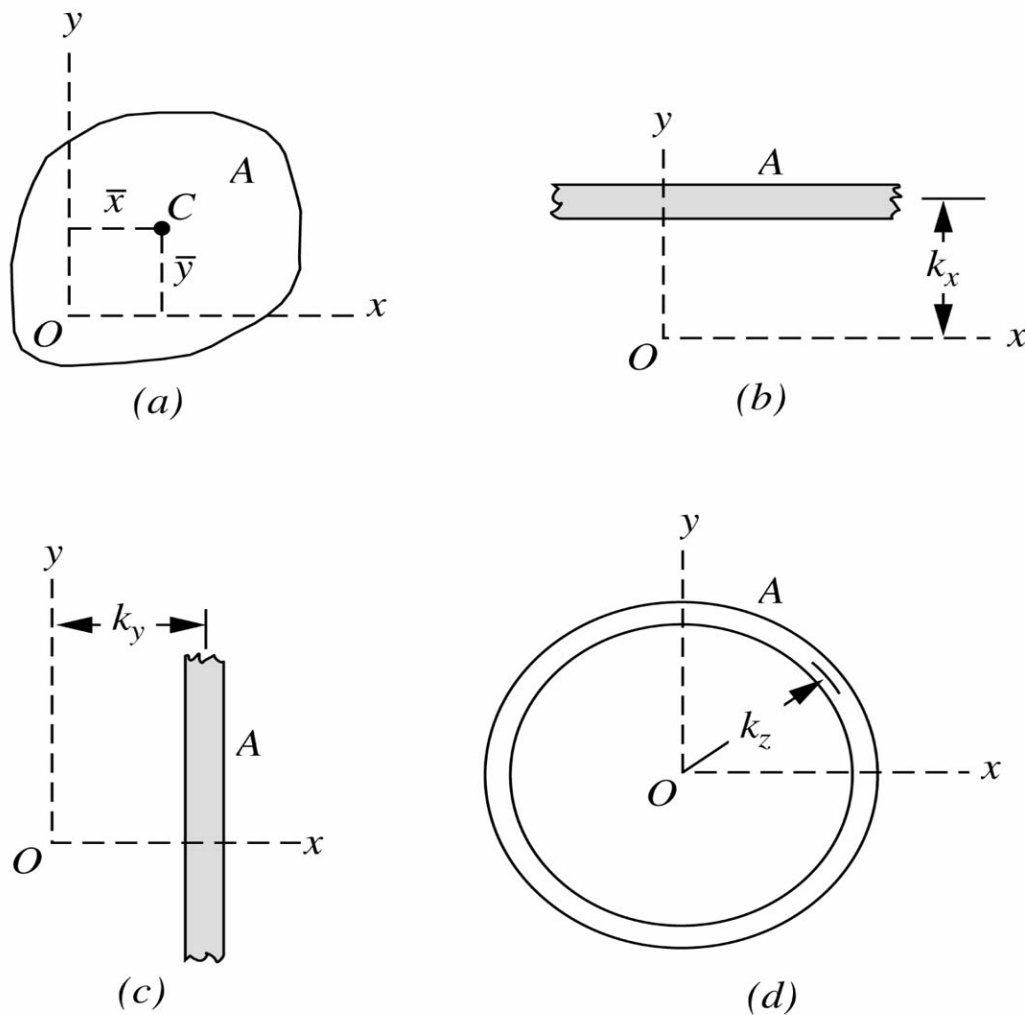
Consider the area  $A$ , Fig. 3.3(a), which has rectangular moments of inertia  $I_x$  and  $I_y$  and a polar moment of inertia  $I_z$  about  $O$ . If the area is visualized as being concentrated into a long narrow strip of area  $A$  a distance  $k_x$  from the  $x$  axis, Fig. 3.3(b), by definition the moment of inertia of the strip about the  $x$  axis will be the same as that of the original area if  $k_x^2 A = I_x$ . The distance  $k_x$  is known as the **radius of gyration** of the area about the  $x$  axis. A similar relation for the  $y$  axis is found by considering the area to be concentrated into a narrow strip parallel to the  $y$  axis as shown in Fig. 3.3(c). Also, by visualizing the area to be concentrated into a narrow ring of radius  $k_z$ , as shown in Fig. 3.3(d), the polar moment of inertia becomes  $k_z^2 A = I_z$ . Summarizing,

$$\begin{aligned} I_x &= k_x^2 A & k_x &= \sqrt{I_x/A} \\ I_y &= k_y^2 A & k_y &= \sqrt{I_y/A} \\ I_z &= k_z^2 A & k_z &= \sqrt{I_z/A} \end{aligned} \quad (3.4)$$

A rectangular or polar moment of inertia may be expressed by specifying its radius of gyration and its area. Substituting Eq. (3.4) into Eq. (3.3) gives

$$k_z^2 = k_x^2 + k_y^2 \quad (3.5)$$

**Figure 3.3** (Source: Meriam, J. L. and Kraige, L. G. 1992. *Engineering Mechanics*, 3rd ed. John Wiley & Sons, New York.)



### Parallel-Axis Theorem

The moment of inertia of an area about a noncentroidal axis may be easily expressed in terms of the moment of inertia about a parallel centroidal axis. In Fig. 3.4 the  $x_0$  and  $y_0$  axes pass through the centroid  $C$  of the area. By definition the moment of inertia of the element  $dA$  about the  $x$  axis is  $dI_x = (y_0 + d_x)^2 dA$ . Expanding and integrating give

$$I_x = \int y_0^2 dA + 2d_x \int y_0 dA + d_x^2 \int dA$$

The first integral is by definition the moment of inertia  $\bar{I}_x$  about the centroidal  $x_0$  axis. The

second integral is zero because  $\int y_0 dA = A\bar{y}_0$  where  $\bar{y}_0$  is automatically zero because the centroid for the area lies on the  $x_0$  axis. The third term is simply  $Ad_x^2$ . Thus, the expression for  $I_x$  and the similar expression for  $I_y$  become

$$\begin{aligned} I_x &= \bar{I}_x + Ad_x^2 \\ I_y &= \bar{I}_y + Ad_y^2 \end{aligned} \quad (3.6)$$

By Eq. (3.3) the sum of these two equations gives

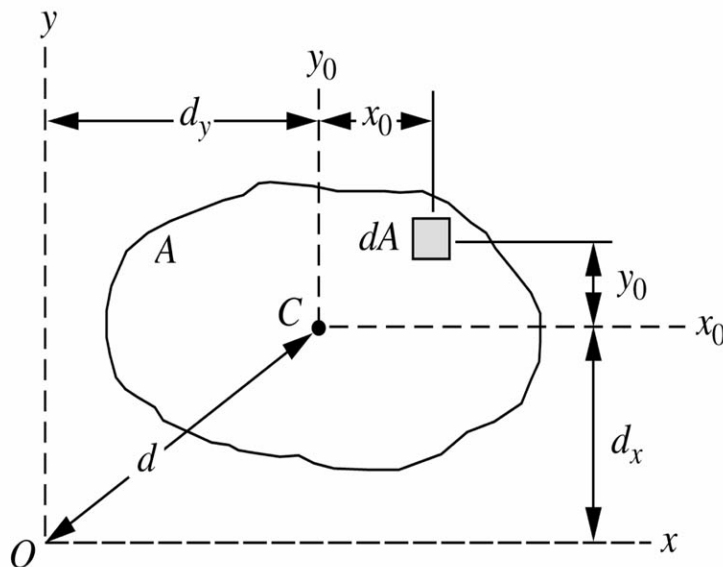
$$I_z = \bar{I}_z + Ad^2 \quad (3.6a)$$

Equations (3.6) and (3.6a) are the so-called *parallel-axis theorems*. It is noted that the axes between which transfer is made *must be parallel* and that one of the axes *must pass through the centroid* of the area. The parallel-axis theorems also hold for radii of gyration. Substitution of the definition of  $k$  into Eq. (3.6) gives

$$k^2 = \bar{k}^2 + d^2 \quad (3.6b)$$

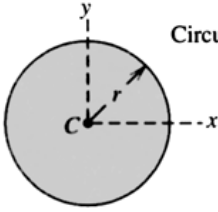
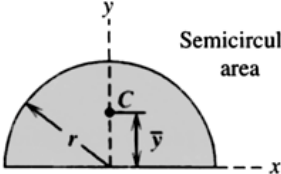
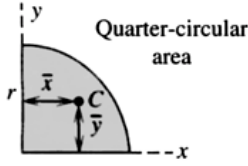
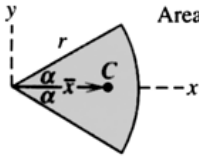
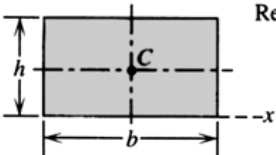
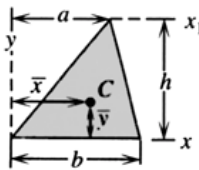
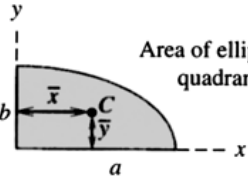
where  $\bar{k}$  is the radius of gyration about the centroidal axis parallel to the axis about which  $k$  applies and  $d$  is the distance between the two axes. The axes may be either in or normal to the plane of the area.

**Figure 3.4** (Source: Meriam, J. L. and Kraige, L. G. 1992. *Engineering Mechanics*, 3rd ed. John Wiley & Sons, New York.)



A summary of formulas for area moments of inertia for various commonly encountered plane areas is given in [Table 3.1](#).

**Table 3.1** Properties of Plane Areas

Figure	Centroid	Area Moments of Inertia
 <p>Circular area</p>	—	$I_x = I_y = \frac{\pi r^4}{4}$ $I_z = \frac{\pi r^4}{2}$
 <p>Semicircular area</p>	$\bar{y} = \frac{4r}{3\pi}$	$I_x = I_y = \frac{\pi r^4}{8}$ $\bar{I}_x = \left( \frac{\pi}{8} - \frac{8}{9\pi} \right) r^4$ $I_z = \frac{\pi r^4}{4}$
 <p>Quarter-circular area</p>	$\bar{x} = \bar{y} = \frac{4r}{3\pi}$	$I_x = I_y = \frac{\pi r^4}{16}$ $\bar{I}_x = \bar{I}_y = \left( \frac{\pi}{16} - \frac{4}{9\pi} \right) r^4$ $I_z = \frac{\pi r^4}{8}$
 <p>Area of circular sector</p>	$\bar{x} = \frac{2}{3} \frac{r \sin \alpha}{\alpha}$	$I_x = \frac{r^4}{4} \left( \alpha - \frac{1}{2} \sin 2\alpha \right)$ $I_y = \frac{r^4}{4} \left( \alpha + \frac{1}{2} \sin 2\alpha \right)$ $I_z = \frac{1}{2} r^4 \alpha$
 <p>Rectangular area</p>	—	$I_x = \frac{bh^3}{3}$ $\bar{I}_x = \frac{bh^3}{12}$ $\bar{I}_z = \frac{bh}{12} (b^2 + h^2)$
 <p>Triangular area</p>	$\bar{x} = \frac{a + b}{3}$ $\bar{y} = \frac{h}{3}$	$I_x = \frac{bh^3}{12}$ $\bar{I}_x = \frac{bh^3}{36}$ $I_{x_1} = \frac{bh^3}{4}$
 <p>Area of elliptical quadrant</p>	$\bar{x} = \frac{4a}{3\pi}$ $\bar{y} = \frac{4b}{3\pi}$	$I_x = \frac{\pi ab^3}{16}, \quad \bar{I}_x = \left( \frac{\pi}{16} - \frac{4}{9\pi} \right) ab^3$ $I_y = \frac{\pi a^3 b}{16}, \quad \bar{I}_y = \left( \frac{\pi}{16} - \frac{4}{9\pi} \right) a^3 b$ $I_z = \frac{\pi ab}{16} (a^2 + b^2)$

Source: Meriam, J. L. and Kraige, L. G. 1992. *Engineering Mechanics*, 3rd ed. John Wiley & Sons, New York.



## Composite Areas

When an area is the composite of a number of distinct parts, its moment of inertia is obtained by summing the results for each of the parts in terms of its area  $A$ , its centroidal moment of inertia  $\bar{I}$ , the perpendicular distance  $d$  from its centroidal axis to the axis about which the moment of inertia of the composite area is being computed, and the product  $Ad^2$ . The results are easily tabulated in the form

Part	Area, $A$	$\bar{I}_x$	$\bar{I}_y$	$d_x$	$d_y$	$Ad_x^2$	$Ad_y^2$
Sums		$\sum \bar{I}_x$	$\sum \bar{I}_y$			$\sum Ad_x^2$	$\sum Ad_y^2$

The final results are simply

$$I_x = \sum \bar{I}_x + \sum Ad_x^2 \quad I_y = \sum \bar{I}_y + \sum Ad_y^2$$

## Products of Inertia

In certain problems involving unsymmetrical cross sections, an expression of the form  $dI_{xy} = xy \, dA$  occurs, and its integral

$$I_{xy} = \int xy \, dA \quad (3.7)$$

is known as the **product of inertia**. Unlike moments of inertia, which are always positive for positive areas, the product of inertia may be positive, negative, or zero, depending on the signs of  $x$  and  $y$ .

## Rotation of Axes

It may be shown that the moments and products of inertia for the area of [Fig. 3.5](#) about the rotated axes  $x^0$ - $y^0$  are given by

$$\begin{aligned}
I_{x^0} &= \frac{I_x + I_y}{2} + \frac{I_x - I_y}{2} \cos 2\theta + I_{xy} \sin 2\theta \\
I_{y^0} &= \frac{I_x + I_y}{2} - \frac{I_x - I_y}{2} \cos 2\theta + I_{xy} \sin 2\theta \\
I_{x^0 y^0} &= \frac{I_x - I_y}{2} \sin 2\theta + I_{xy} \cos 2\theta
\end{aligned} \tag{3.8}$$

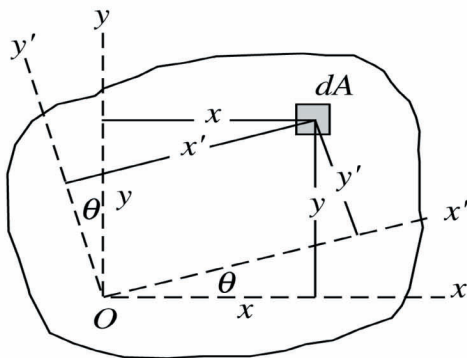
The angle that makes  $I_{x^0}$  and  $I_{y^0}$  a maximum or a minimum is determined by setting the derivative of  $I_{x^0}$  and  $I_{y^0}$  with respect to  $\theta$  equal to zero. Denoting this critical angle by  $\alpha$  gives

$$\tan 2\alpha = \frac{2I_{xy}}{I_y - I_x} \tag{3.9}$$

Substitution of Eq. (3.9) for  $2\theta$  in Eq. (3.8) gives  $I_{x^0 y^0} = 0$  and

$$\begin{aligned}
I_{\max} &= \frac{1}{2} (I_x + I_y) + \frac{1}{2} \sqrt{(I_x - I_y)^2 + 4I_{xy}^2} \\
I_{\min} &= \frac{1}{2} (I_x + I_y) - \frac{1}{2} \sqrt{(I_x - I_y)^2 + 4I_{xy}^2}
\end{aligned} \tag{3.10}$$

**Figure 3.5** (Source: Meriam, J. L. and Kraige, L. G. 1992. *Engineering Mechanics*, 3rd ed. John Wiley & Sons, New York.)



## 3.2 Mass Moments of Inertia

The dynamics of bodies that rotate with angular acceleration calls for a knowledge of mass moments of inertia and is treated in the chapter on Dynamics and Vibration.

## Defining Relations

### Fixed Axis

The moment of inertia of the body of mass  $m$  in Fig. 3.6 about the fixed axis  $O-O$  is given by

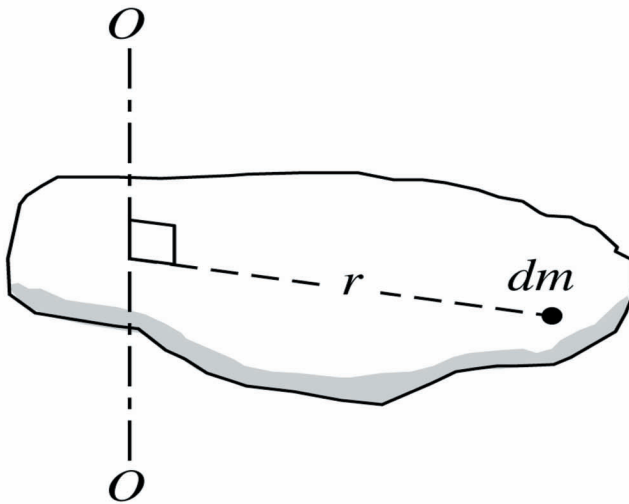
$$I_o = \int r^2 dm \quad (3.11)$$

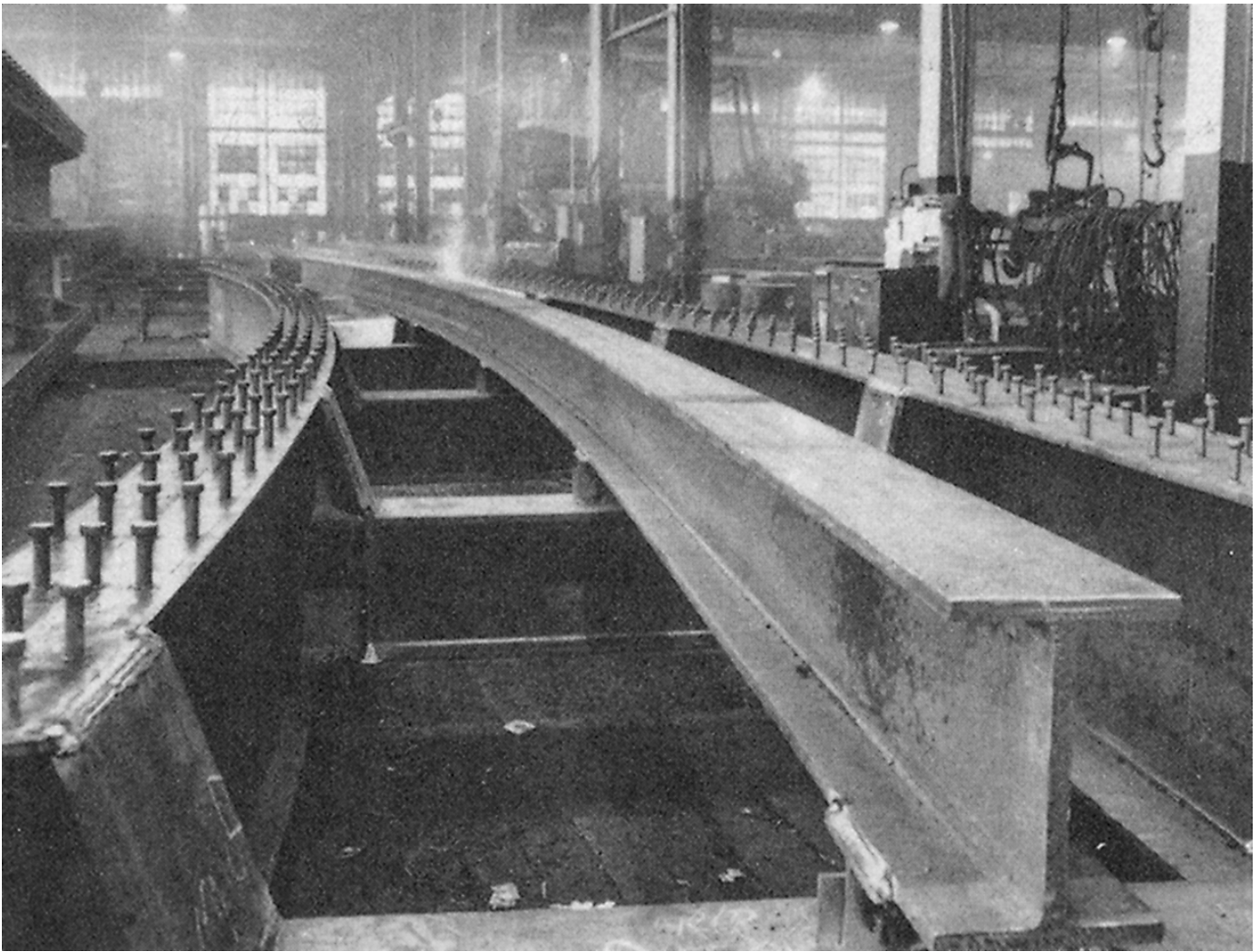
The dimensions are (mass)(length)<sup>2</sup>, which are kg·m<sup>2</sup> in SI units and lb·ft·s<sup>2</sup> in U.S. customary units. If the density  $\rho$  of the body is constant, then  $dm = \rho dV$  and the integral becomes

$$I_o = \rho \int r^2 dV \quad (3.12)$$

where  $dV$  is the differential volume of the mass element. To facilitate integration, coordinates that best suit the boundaries of the body should be utilized.

**Figure 3.6** (Source: Meriam, J. L. and Kraige, L. G. 1992. *Engineering Mechanics*, 3rd ed. John Wiley & Sons, New York.)





Knowledge of the geometric properties of the cross section of a structural member is important for its design. The wide-flange members shown here are to be used for the Transit Expressway demonstration project in Pittsburgh, Pennsylvania. It consists of a 9340-foot-long, 8000-foot-elevated experimental loop designed to test and display a new concept in urban rapid transit. (Photo courtesy of Bethlehem Steel.)

### Radius of Gyration

The **radius of gyration**  $k$  of a mass  $m$  about an axis for which the moment of inertia is  $I$  is defined as

$$k = \sqrt{I/m} \quad \text{or} \quad I = k^2 m \quad (3.13)$$

Thus  $k$  is a measure of the distribution of mass about the axis in question, and its definition is analogous to the similar definition of radius of gyration for area moments of inertia.

### Parallel-Axis Theorem

If the moment of inertia of a body of mass  $m$  is known about an axis through the mass center, it may easily be determined about any parallel axis by the expression

$$I = \bar{I} + md^2 \quad (3.14)$$

where  $\bar{I}$  is the moment of inertia about the parallel axis through the mass center and  $d$  is the perpendicular distance between the axes. This *parallel-axis theorem* is analogous to that for area moments of inertia, Eq. (3.6). It applies *only* if transfer is made to or from a parallel axis through the mass center. From the definition of Eq. (3.13) it follows that

$$k^2 = \bar{k}^2 + d^2 \quad (3.15)$$

where  $k$  is the radius of gyration about an axis a distance  $d$  from the parallel axis through the mass center for which the radius of gyration is  $\bar{k}$ .

### Flat Plates

The moments of inertia of a flat plate, [Fig. 3.7](#), about axes in the plane of and normal to the plate are frequently encountered. The elemental mass is  $\rho(t \, dA)$ , where  $\rho$  is the plate density,  $t$  is its thickness, and  $dA = dx \, dy$  is the face area of  $dm$ . If  $\rho$  and  $t$  are constant, the moment of inertia about each of the axes becomes

$$\begin{aligned} I_{xx} &= \int y^2 \, dm = \rho t \int y^2 \, dA = \rho t I_x \\ I_{yy} &= \int x^2 \, dm = \rho t \int x^2 \, dA = \rho t I_y \\ I_{zz} &= \int r^2 \, dm = \rho t \int r^2 \, dA = \rho t I_z \end{aligned} \quad (3.16)$$

where the double subscript designates mass moment of inertia and the single subscript designates the moment of inertia of the plate area. Inasmuch as  $I_z = I_x + I_y$  for area moments of inertia, it follows that

$$I_{zz} = I_{xx} + I_{yy} \quad (3.16a)$$

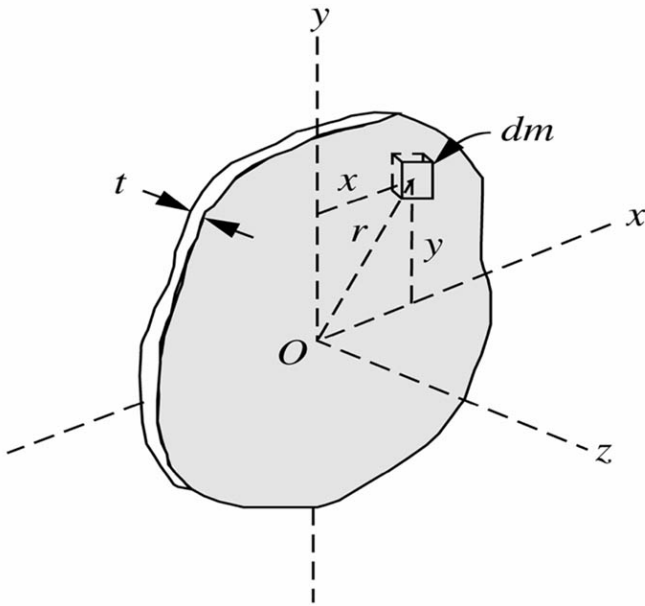
This relation holds *only* if  $t$  is small compared with the other plate dimensions.

### Composite Bodies

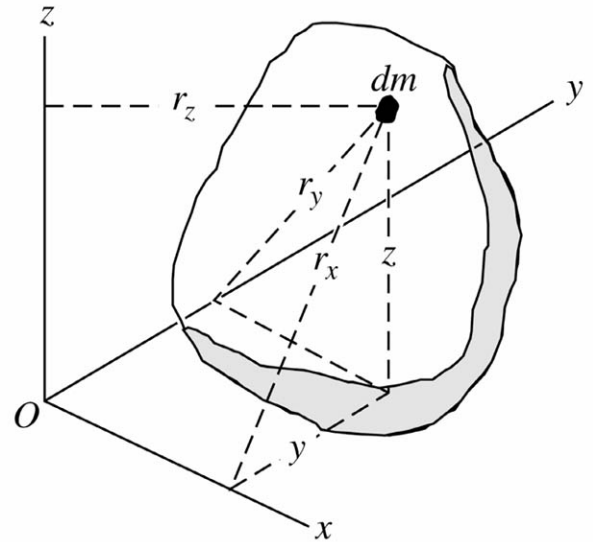
The mass moment of inertia of a composite body about a given axis is simply the sum of the moments of inertia of its individual components about the same axis.

A summary of formulas for mass moments of inertia for various bodies of common shape is given in [Table 3.2](#).

**Figure 3.7** (Source: Meriam, J. L. and Kraige, L. G. 1992. Engineering Mechanics, 3rd ed. John Wiley & Sons, New York.)



**Figure 3.8** (Source: Meriam, J. L. and Kraige, L. G. 1992. Engineering Mechanics, 3rd ed. John Wiley & Sons, New York.)



### General Rotation

For three-dimensional rotation of a rigid body the moments and products of inertia assume a more general form from Fig. 3.8, as follows:

$$\begin{aligned}
 I_{xx} &= \int r_x^2 dm = \int (y^2 + z^2) dm & I_{xy} &= I_{yx} = \int xy dm \\
 I_{yy} &= \int r_y^2 dm = \int (z^2 + x^2) dm & I_{xz} &= I_{zx} = \int xz dm \\
 I_{zz} &= \int r_z^2 dm = \int (x^2 + y^2) dm & I_{yz} &= I_{zy} = \int yz dm
 \end{aligned} \tag{3.17}$$

Whereas the moments of inertia are always positive, the products of inertia may be positive, negative, or zero. Parallel-axis theorems for products of inertia are

$$\begin{aligned}
 I_{xy} &= \bar{I}_{xy} + m\bar{x}\bar{y} \\
 I_{xz} &= \bar{I}_{xz} + m\bar{x}\bar{z} \\
 I_{yz} &= \bar{I}_{yz} + m\bar{y}\bar{z}
 \end{aligned} \tag{3.18}$$

where the bar represents the product of inertia with respect to axes through the mass center and  $\bar{x}$ ,  $\bar{y}$ , and  $\bar{z}$  represent the coordinates of the mass center.

**Table 3.2** Moments of Inertia of Homogeneous Solids ( $m$  = Mass of Body Shown) (*continues*)

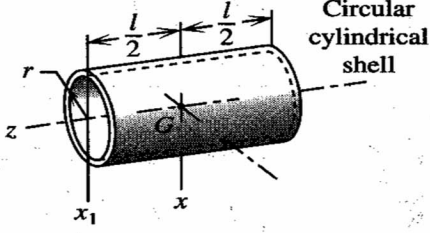
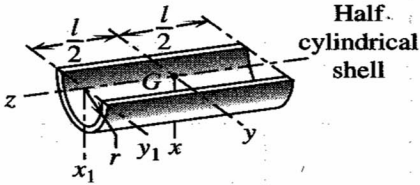
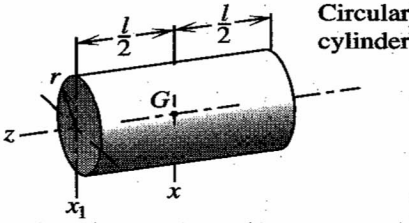
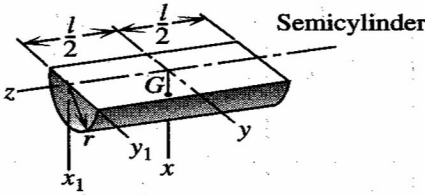
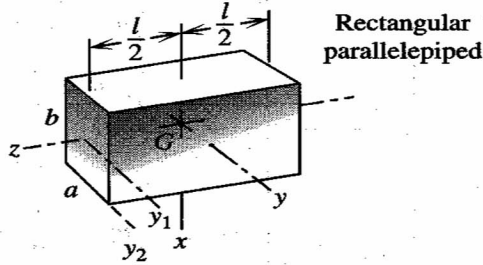
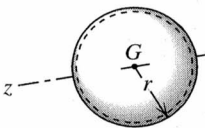
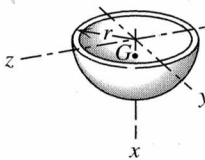
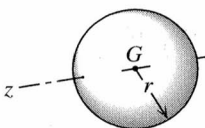
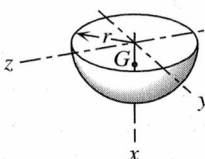
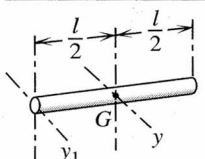
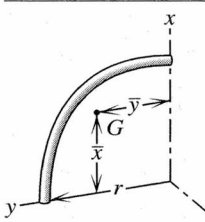
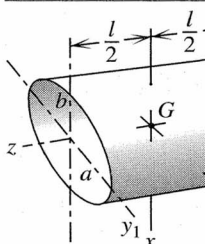
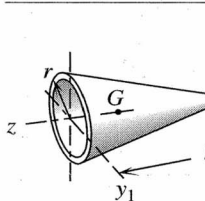
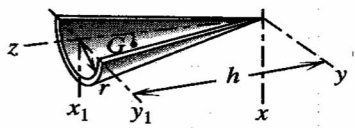
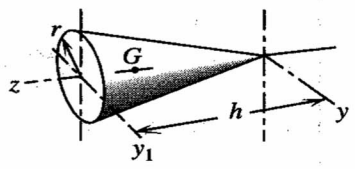
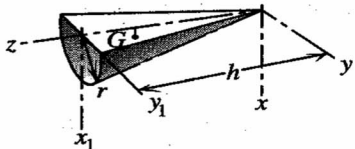
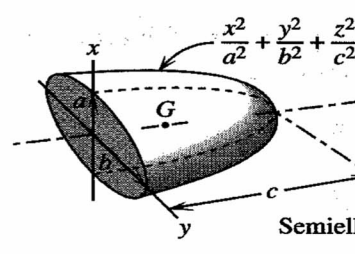
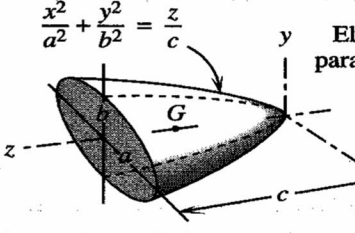
Body	Mass Center	Mass Moments of Inertia
 <p>Circular cylindrical shell</p>	—	$I_{xx} = \frac{1}{2}mr^2 + \frac{1}{12}ml^2$ $I_{x_1x_1} = \frac{1}{2}mr^2 + \frac{1}{3}ml^2$ $I_{zz} = mr^2$
 <p>Half-cylindrical shell</p>	$\bar{x} = \frac{2r}{\pi}$	$I_{xx} = I_{yy} = \frac{1}{2}mr^2 + \frac{1}{12}ml^2$ $I_{x_1x_1} = I_{y_1y_1} = \frac{1}{2}mr^2 + \frac{1}{3}ml^2$ $I_{zz} = mr^2$ $\bar{I}_{zz} = \left(1 - \frac{4}{\pi^2}\right)mr^2$
 <p>Circular cylinder</p>	—	$I_{xx} = \frac{1}{4}mr^2 + \frac{1}{12}ml^2$ $I_{x_1x_1} = \frac{1}{4}mr^2 + \frac{1}{3}ml^2$ $I_{zz} = \frac{1}{2}mr^2$
 <p>Semicylinder</p>	$\bar{x} = \frac{4r}{3\pi}$	$I_{xx} = I_{yy} = \frac{1}{4}mr^2 + \frac{1}{12}ml^2$ $I_{x_1x_1} = I_{y_1y_1} = \frac{1}{4}mr^2 + \frac{1}{3}ml^2$ $I_{zz} = \frac{1}{2}mr^2$ $\bar{I}_{zz} = \left(\frac{1}{2} - \frac{16}{9\pi^2}\right)mr^2$
 <p>Rectangular parallelepiped</p>	—	$I_{xx} = \frac{1}{12}m(a^2 + l^2)$ $I_{yy} = \frac{1}{12}m(b^2 + l^2)$ $I_{zz} = \frac{1}{12}m(a^2 + b^2)$ $I_{y_1y_1} = \frac{1}{12}mb^2 + \frac{1}{3}ml^2$ $I_{y_2y_2} = \frac{1}{3}m(b^2 + l^2)$

Table 3.2 (continued)

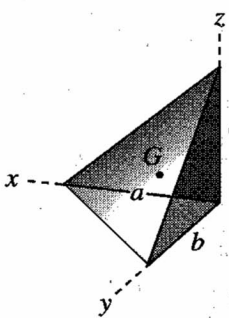
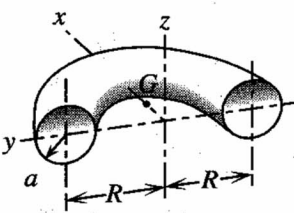
Body	Mass Center	Mass Moments of Inertia
 <p>Spherical shell</p>	—	$I_{zz} = \frac{2}{3}mr^2$
 <p>Hemispherical shell</p>	$\bar{x} = \frac{r}{2}$	$I_{xx} = I_{yy} = I_{zz} = \frac{2}{3}mr^2$ $\bar{I}_{yy} = \bar{I}_{zz} = \frac{5}{12}mr^2$
 <p>Sphere</p>	—	$I_{zz} = \frac{2}{5}mr^2$
 <p>Hemisphere</p>	$\bar{x} = \frac{3r}{8}$	$I_{xx} = I_{yy} = I_{zz} = \frac{2}{5}mr^2$ $\bar{I}_{yy} = \bar{I}_{zz} = \frac{83}{320}mr^2$
 <p>Uniform slender rod</p>	—	$I_{yy} = \frac{1}{12}ml^2$ $I_{y_1y_1} = \frac{1}{3}ml^2$
 <p>Quarter-circular rod</p>	$\bar{x} = \bar{y} = \frac{2r}{\pi}$	$I_{xx} = I_{yy} = \frac{1}{2}mr^2$ $I_{zz} = mr^2$
 <p>Elliptical cylinder</p>	—	$I_{xx} = \frac{1}{4}ma^2 + \frac{1}{12}ml^2$ $I_{yy} = \frac{1}{4}mb^2 + \frac{1}{12}ml^2$ $I_{zz} = \frac{1}{4}m(a^2 + b^2)$ $I_{y_1y_1} = \frac{1}{4}mb^2 + \frac{1}{3}ml^2$
 <p>Conical shell</p>	$\bar{z} = \frac{2h}{3}$	$I_{yy} = \frac{1}{4}mr^2 + \frac{1}{2}mh^2$ $I_{y_1y_1} = \frac{1}{4}mr^2 + \frac{1}{6}mh^2$ $I_{zz} = \frac{1}{2}mr^2$ $\bar{I}_{yy} = \frac{1}{4}mr^2 + \frac{1}{18}mh^2$



**Table 3.2 (continued)**

Body	Mass Center	Mass Moments of Inertia
 <p>Half conical shell</p>	$\bar{x} = \frac{4r}{3\pi}$ $\bar{z} = \frac{2h}{3}$	$I_{xx} = I_{yy} = \frac{1}{4}mr^2 + \frac{1}{2}mh^2$ $I_{x_1x_1} = I_{y_1y_1} = \frac{1}{4}mr^2 + \frac{1}{6}mh^2$ $I_{zz} = \frac{1}{2}mr^2$ $\bar{I}_{zz} = \left(\frac{1}{2} - \frac{16}{9\pi^2}\right)mr^2$
 <p>Right-circular cone</p>	$\bar{z} = \frac{3h}{4}$	$I_{yy} = \frac{3}{20}mr^2 + \frac{3}{5}mh^2$ $I_{y_1y_1} = \frac{3}{20}mr^2 + \frac{1}{10}mh^2$ $I_{zz} = \frac{3}{10}mr^2$ $\bar{I}_{yy} = \frac{3}{20}mr^2 + \frac{3}{80}mh^2$
 <p>Half cone</p>	$\bar{x} = \frac{r}{\pi}$ $\bar{z} = \frac{3h}{4}$	$I_{xx} = I_{yy} = \frac{3}{20}mr^2 + \frac{3}{5}mh^2$ $I_{x_1x_1} = I_{y_1y_1} = \frac{3}{20}mr^2 + \frac{1}{10}mh^2$ $I_{zz} = \frac{3}{10}mr^2$ $\bar{I}_{zz} = \left(\frac{3}{10} - \frac{1}{\pi^2}\right)mr^2$
 <p>Semiellipsoid</p>	$\bar{z} = \frac{3c}{8}$	$I_{xx} = \frac{1}{5}m(b^2 + c^2)$ $I_{yy} = \frac{1}{5}m(a^2 + c^2)$ $I_{zz} = \frac{1}{5}m(a^2 + b^2)$ $\bar{I}_{xx} = \frac{1}{5}m\left(b^2 + \frac{19}{64}c^2\right)$ $\bar{I}_{yy} = \frac{1}{5}m\left(a^2 + \frac{19}{64}c^2\right)$
 <p>Elliptic paraboloid</p>	$\bar{z} = \frac{2c}{3}$	$I_{xx} = \frac{1}{6}mb^2 + \frac{1}{2}mc^2$ $I_{yy} = \frac{1}{6}ma^2 + \frac{1}{2}mc^2$ $I_{zz} = \frac{1}{6}m(a^2 + b^2)$ $\bar{I}_{xx} = \frac{1}{6}m\left(b^2 + \frac{1}{3}c^2\right)$ $\bar{I}_{yy} = \frac{1}{6}m\left(a^2 + \frac{1}{3}c^2\right)$

**Table 3.2** (continued)

Body	Mass Center	Mass Moments of Inertia
 <p>Rectangular tetrahedron</p>	$\bar{x} = \frac{a}{4}$ $\bar{y} = \frac{b}{4}$ $\bar{z} = \frac{c}{4}$	$I_{xx} = \frac{1}{10} m(b^2 + c^2)$ $I_{yy} = \frac{1}{10} m(a^2 + c^2)$ $I_{zz} = \frac{1}{10} m(a^2 + b^2)$ $\bar{I}_{xx} = \frac{3}{80} m(b^2 + c^2)$ $\bar{I}_{yy} = \frac{3}{80} m(a^2 + c^2)$ $\bar{I}_{zz} = \frac{3}{80} m(a^2 + b^2)$
 <p>Half torus</p>	$\bar{x} = \frac{a^2 + 4R^2}{2\pi R}$	$I_{xx} = I_{yy} = \frac{1}{2} mR^2 + \frac{5}{8} ma^2$ $I_{zz} = mR^2 + \frac{3}{4} ma^2$

## Defining Terms

**Area moment of inertia:** Defined as  $\int_R (\text{distance})^2 d(\text{area})$ .

**Mass moment of inertia:** A measure of the inertial resistance to angular acceleration.

**Product of inertia:** Defined as  $\int xy dA$  for areas ( $xy$  plane) and  $\int xy dm$ ,  $\int xz dm$ , and  $\int yz dm$  for masses.

**Radius of gyration:** Defined as  $\sqrt{I/A}$  for areas and  $\sqrt{I/m}$  for masses.

## Reference

Meriam, J. L. and Kraige, L. G. 1992. *Engineering Mechanics*, 3rd ed. John Wiley & Sons, New York.

## Further Information

Consult mechanics textbooks found in any engineering library.

Bauld, Jr.N. R. "Axial Loads and Torsion"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Axial Loads and Torsion

---

## 10.1 Axially Loaded Bars

Axial Strain • Axial Stress • Axial Stress-Strain Relation • Relative Displacement of Cross Sections • Uniform Bar • Nonuniform Bars • Statically Indeterminate Bars

## 10.2 Torsion

Power Transmission • Kinematics of Circular Shafts • Equilibrium • Elastic Twisting of Circular Shafts • Uniform Shaft • Nonuniform Shaft • Statically Indeterminate Circular Shafts

**Nelson R. Bauld, Jr.**

*Clemson University*

---

## 10.1 Axially Loaded Bars

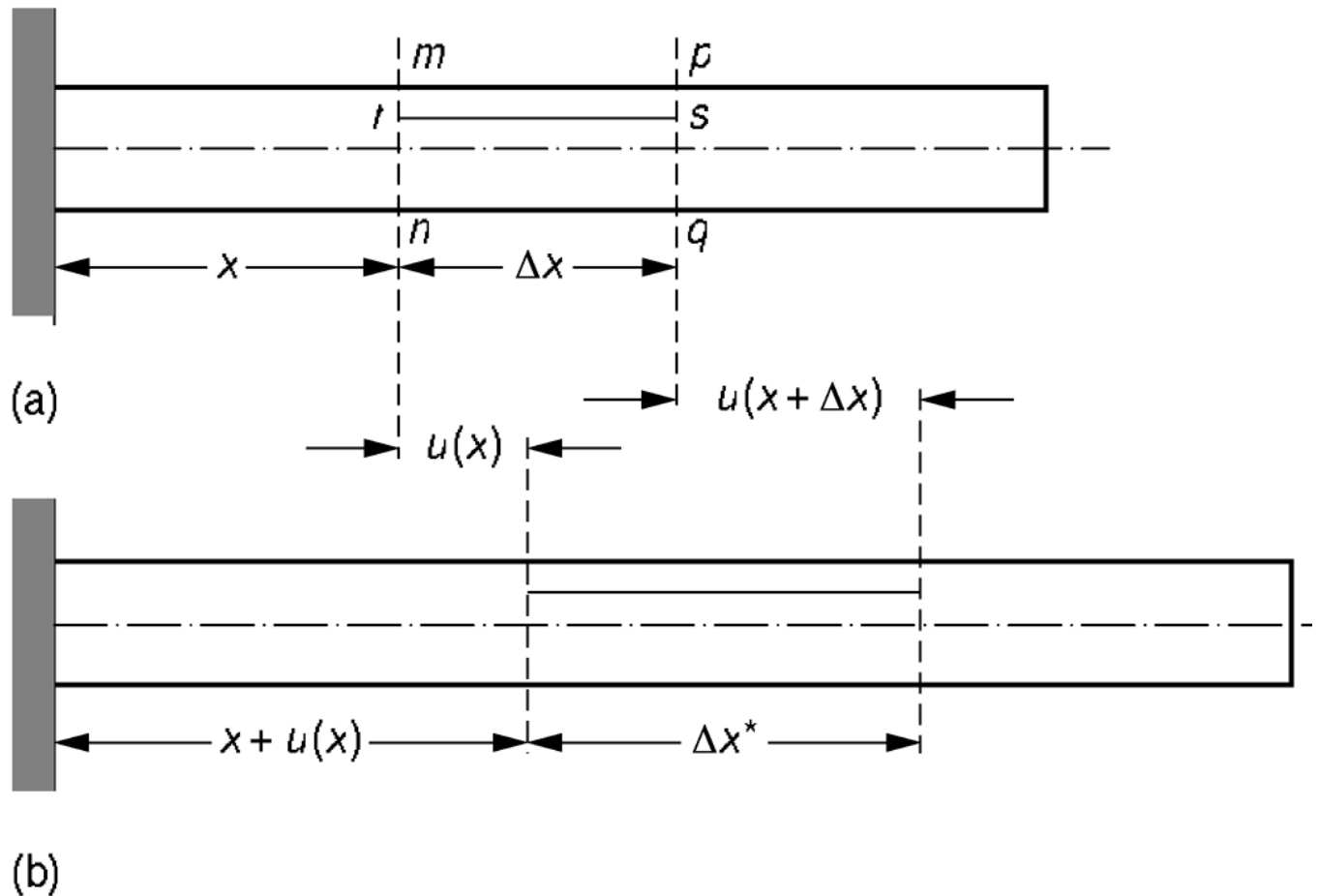
---

A bar is said to be axially loaded if the action lines of all the applied forces coincide with the axis of the bar. The **bar axis** is defined as the locus of the centroids of the cross-sectional areas along the length of the bar. This locus of centroids must form a straight line, and the action lines of the applied forces must coincide with it in order for the theory of this section to apply.

### Axial Strain

The axial strain in an axially loaded bar is based on the geometric assumptions that plane cross sections in the unloaded bar, such as sections  $m\bar{n}$  and  $p\bar{q}$  in [Fig. 10.1\(a\)](#), remain plane in the loaded bar as shown in [Fig. 10.1\(b\)](#), and that they displace only axially.

**Figure 10.1** Axial displacements of an axially loaded bar.



The axial strain of a **line element** such as  $rs$  in Fig. 10.1(a) is defined as the limit of the ratio of its change in length to its original length as its original length approaches zero. Thus, the axial strain  $\epsilon$  at an arbitrary cross section  $x$  is

$$\epsilon(x) = \lim_{\Delta x \rightarrow 0} \frac{(\Delta x + \Delta u) - \Delta x}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{[u(x + \Delta x) - u(x)]}{\Delta x} = \frac{du}{dx} \quad (10:1)$$

where  $u(x)$  and  $u(x + \Delta x)$  are axial displacements of the cross sections at  $x$  and  $x + \Delta x$ . Common units for axial strain are in./in. or mm/mm. Because axial strain is the ratio of two lengths, units for axial strain are frequently not recorded.

## Axial Stress

The axial stress  $\sigma$  at cross section  $x$  of an axially loaded bar is

$$\sigma(x) = \frac{N(x)}{A(x)} \quad (10:2)$$

where  $N(x)$  is the internal force and  $A(x)$  is the cross-sectional area, each at section  $x$ .

Common units for axial stress are pounds per square inch (psi) or megapascals (MPa). Equation (10.2) is valid at cross sections that satisfy the geometric assumptions stated previously. It ceases to be valid at abrupt changes in cross section and at points of load application. Cross sections at such locations distort and therefore violate the plane cross-section assumption. Also, Eq. (10.2) requires that the material at cross section  $x$  be homogeneous; that is, the cross section cannot be made of two or more different materials.

## Axial Stress-Strain Relation

The allowable stress for axially loaded bars used in most engineering structures falls within the proportional limit of the material from which they are made. Consequently, material behavior considered in this section is confined to the linearly elastic range and is given by

$$\epsilon(x) = E(x)^{-1} \sigma(x) \quad (10:3)$$

where  $E(x)$  is the modulus of elasticity for the material at section  $x$ . Common units for the modulus of elasticity are pounds per square inch (psi) or gigapascals (GPa).

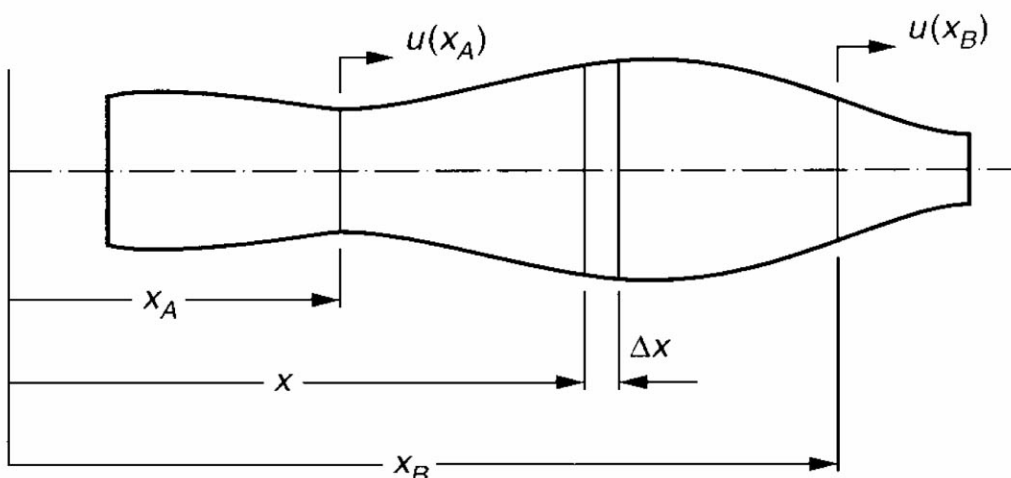
## Relative Displacement of Cross Sections

The relative displacement  $e_{B=A}$  of a cross section at  $x_B$  with respect to a cross section at  $x_A$  is obtained by combining Eqs. (10.1–10.3) and integrating from section  $x_A$  to  $x_B$ . Using [Fig. 10.2](#),

$$e_{B=A} = u(x_B) - u(x_A) = \int_{x_A}^{x_B} N(x) / [A(x)E(x)] dx \quad (10:4)$$

where  $e_{B=A}$  denotes the change in length between the cross sections at  $x_A$  and  $x_B$ .

**Figure 10.2**



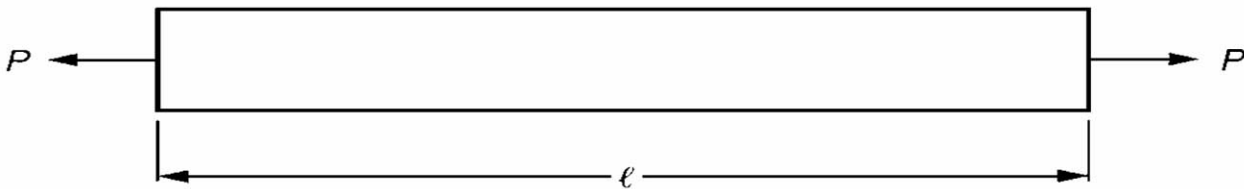
Equation (10.4) must be interpreted as the sum of several integrals for a bar for which the integrand exhibits discontinuities. Discontinuities occur for cross sections where either  $N$ ,  $A$ ,  $E$ , or combinations thereof change abruptly and can usually be detected by inspection.

## Uniform Bar

A bar for which the internal force  $N(x)$ , the cross-sectional area  $A(x)$ , and the modulus of elasticity  $E(x)$  do not change over its length is referred to as a **uniform bar**. If  $P$  denotes equilibrating forces applied to the ends of the bar and  $L$  its length, as shown in Fig. 10.3, then Eq. (10.4) gives the change in length of the bar as

$$e = PL/AE \quad (10:5)$$

**Figure 10.3** Uniform bar.



## Nonuniform Bars

A **nonuniform bar** is one for which either  $A$ ,  $E$ ,  $N$ , or combinations thereof change abruptly along the length of the bar. Three important methods are available to analyze axially loaded bars for which the integrand in Eq. (10.4) contains discontinuities. They are as follows.

### Direct Integration

Equation (10.4) is integrated directly. The internal force  $N(x)$  is obtained in terms of the applied forces via the axial equilibrium equation,  $A(x)$  from geometric considerations, and  $E(x)$  by observing the type of material at a given section.

### Discrete Elements

The bar is divided into a finite number of segments, for each of which  $N/AE$  is constant. Each segment is a uniform bar for which its change in length is given by Eq. (10.5). The change in length of the nonuniform bar is the sum of the changes in length of the various segments. Accordingly, if  $e_i$  denotes the change in length of the  $i$ th segment, then the change in length  $e$  of the nonuniform bar is

$$e = \sum e_i \quad (10:6)$$

## Superposition

The superposition principle applied to axially loaded bars asserts that the change in length between two cross sections caused by several applied forces acting simultaneously is equal to the algebraic sum of the changes in length between the same two cross sections caused by each applied force acting separately. Thus, letting  $e_{B=A}$  represent the change in length caused by several applied forces acting simultaneously, and  $e_{B=A}^0, e_{B=A}^{00}, \dots$  represent the changes in length caused by each applied force acting separately,

$$e_{B=A} = e_{B=A}^0 + e_{B=A}^{00} + \dots \quad (10:7)$$

Superposition of displacements requires that the axial forces be linearly related to the displacements they cause, and this implies that the stress at every cross section cannot exceed the proportional limit stress of the material of the bar. This requirement must be satisfied for each separate loading as well as for the combined loading.

## Statically Indeterminate Bars

The internal force  $N(x)$  in statically determinate axially loaded bars is determined via axial equilibrium alone. Subsequently, axial stress, axial strain, and axial displacements can be determined via the foregoing equations.

The internal force  $N(x)$  in statically indeterminate axially loaded bars cannot be determined via axial equilibrium alone. Thus, it is necessary to augment the axial equilibrium equation with an equation (geometric compatibility equation) that accounts for any geometric constraints imposed on the bar—that is, that takes into account how the supports affect the deformation of the bar.

Three basic mechanics concepts are required to analyze statically indeterminate axially loaded bars: axial equilibrium, geometric compatibility of axial deformations, and material behavior (stress-strain relation).

**Example 10.1.** Determine the stresses in the aluminum and steel segments of the composite bar of Fig. 10.4(a) when  $P = 7000$  lb. The cross-sectional areas of the steel and aluminum segments are  $2 \text{ in}^2$  and  $4 \text{ in}^2$ , respectively, and the moduli of elasticity are  $30 \times 10^6$  psi and  $10 \times 10^6$  psi, respectively.

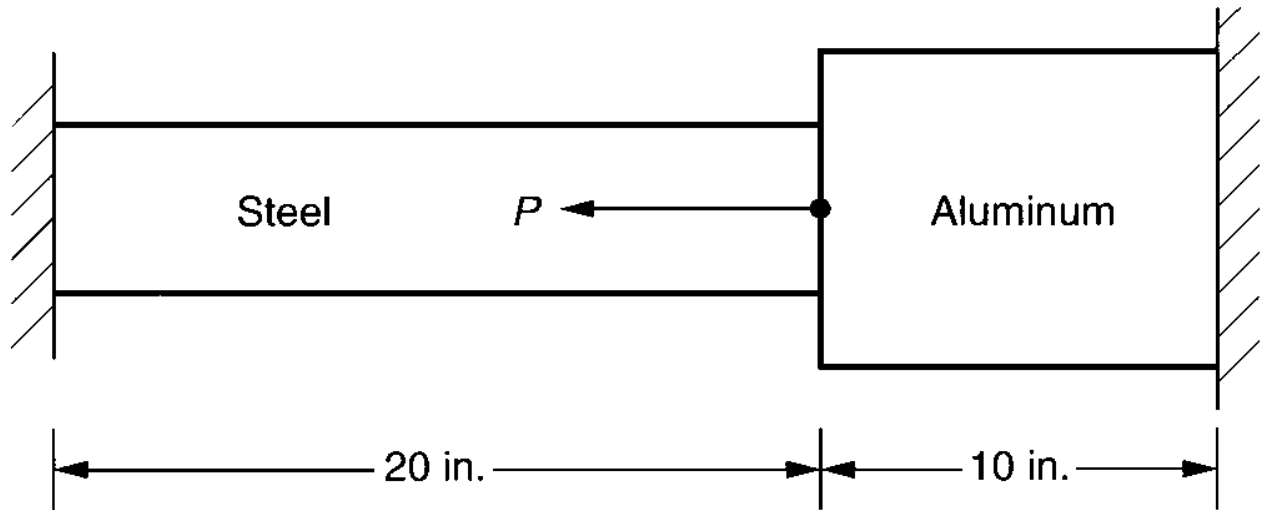
**Solution.** The bar is statically indeterminate; therefore, the solution requires the use of the three mechanics concepts discussed in the previous paragraph.

*Equilibrium.* The axial equilibrium equation is obtained from the free-body diagram of Fig. 10.4(b) as

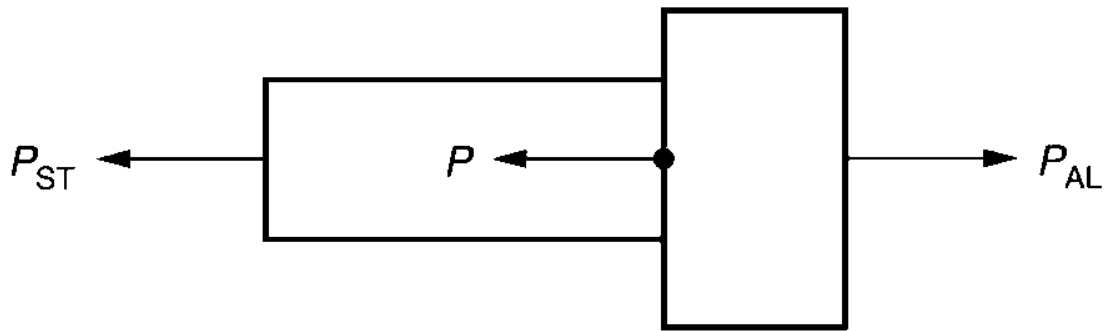
$$-P_{ST} + P_{AL} - 7000 = 0 \quad (10:8)$$



**Figure 10.4** Statically indeterminate composite step-bar.



(a)



(b)

*Geometric compatibility.* The compatibility equation is obtained by noting that the total elongation of the bar is zero. Accordingly,

$$e = e_{ST} + e_{AL} = 0 \quad (10:9)$$

*Material behavior.* The steel and aluminum segments are assumed to behave in a linearly elastic manner, so their elongations are given by

$$e_{ST} = P_{ST} L_{ST} / (A_{ST} E_{ST}) \quad \text{and} \quad e_{AL} = P_{AL} L_{AL} / (A_{AL} E_{AL}) \quad (10:10)$$

Combining Eqs. (10.9) and (10.10) yields

$$\begin{aligned}
 P_{ST} &= \frac{1}{2} (L_{AL}=L_{ST})(E_{ST}=E_{AL})(A_{ST}=A_{AL})P_{AL} \\
 &= \frac{1}{2} (10=20)(30=10)(2=4)P_{AL} = \frac{1}{2} 3=4P_{AL} \quad (10:11)
 \end{aligned}$$

Solving Eqs. (10.8) and (10.11) simultaneously yields

$$P_{ST} = \frac{1}{2} 3000 \text{ lb} \quad \text{and} \quad P_{AL} = 4000 \text{ lb} \quad (10:12)$$

from which the stresses in the steel and aluminum are found as follows:

$$\begin{aligned}
 \sigma_{ST} &= \frac{1}{2} 3000=2 = \frac{1}{2} 1500 \text{ psi} = 1500 \text{ psi (compression)} \\
 \sigma_{AL} &= 4000=4 = 1000 \text{ psi (tension)}
 \end{aligned}$$

**Example 10.2.** Assuming that  $P = 0$  in Fig. 10.4(a), determine the stress in the steel and aluminum segments of the bar due to a temperature increase of  $10^\circ\text{F}$ . The *thermal expansion coefficients* for steel and aluminum are  $\alpha_{ST} = 6.5 \times 10^{-6}$  inches per inch per degree Fahrenheit (in./in./ $^\circ\text{F}$ ) and  $\alpha_{AL} = 13 \times 10^{-6}$  in./in./ $^\circ\text{F}$ .

**Solution.** Because free thermal expansion of the bar is prevented by the supports, internal stresses are induced in the two segments.

*Equilibrium.* The axial equilibrium equation is obtained from the free-body diagram of Fig. 10.4(b). Thus,

$$\frac{1}{2} P_{ST} + P_{AL} = 0 \quad (10:13)$$

*Compatibility.* The compatibility equation is obtained by noting that if the bar could expand freely, its total elongation  $\delta$  would be

$$\delta = \delta_{ST} + \delta_{AL} \quad (10:14)$$

where  $\delta_{ST}$  and  $\delta_{AL}$  denote the free thermal expansions of the separate segments. Because the net change in length of the bar is zero, internal strains are induced in the steel and aluminum such that the sum of the changes in lengths of the steel and aluminum segments must be equal to  $\delta$ . Therefore, the compatibility equation becomes

$$e_{ST} + e_{AL} \frac{1}{2} \delta = 0 \quad (10:15)$$

*Material behavior.* Assuming linear elastic behavior for both materials

$$e_{ST} = \frac{P_{ST} L_{ST}}{A_{ST} E_{ST}} \quad \text{and} \quad e_{AL} = \frac{P_{AL} L_{AL}}{A_{AL} E_{AL}} \quad (10:16)$$

Also, because

$$\phi_{ST} = \frac{P_{ST}}{E_{ST} L_{ST}} \quad \text{and} \quad \phi_{AL} = \frac{P_{AL}}{E_{AL} L_{AL}} \quad (10:17)$$

it follows that

$$\phi = (6.5 \times 10^6)(20)(10) + (13 \times 10^6)(10)(10) = 0.0026 \text{ in.} \quad (10:18)$$

Equations (10.13), (10.15), (10.16), and (10.18) yield

$$P_{ST} f_1 + (E_{ST} = E_{AL})(A_{ST} = A_{AL})(L_{AL} = L_{ST})g = (E_{ST} A_{ST} = L_{ST})\phi$$

or

$$P_{ST} f_1 + (30 \times 10^6)(2 \times 4)(10 \times 20)g = f[30 \times 10^6(2)] = 20g(0.0026)$$

Thus

$$P_{ST} = P_{AL} = 4457 \text{ lb} \quad (10:19)$$

The corresponding stresses in the steel and aluminum are compression and equal to

$$\frac{3}{4}_{ST} = 4457/2 = 2228 \text{ psi} \quad \text{and} \quad \frac{3}{4}_{AL} = 4457/4 = 1114 \text{ psi}$$

## 10.2 Torsion

---

Torsionally loaded bars occur frequently in industrial applications such as shafts connecting motor-pump and motor-generator sets; propeller shafts in airplanes, helicopters, and ships; and torsion bars in automobile suspension systems. Many tools or tool components possess a dominant torsional component such as screwdrivers and drill and router bits. (These tools also rely on an axial force component for their effectiveness.)

### Power Transmission

The specifications for a motor customarily list the power it transmits in horsepower (hp), and its angular speed in either revolutions per minute (rpm) or in cycles per second (Hz). To design or analyze a shaft, the **torque** that it is to transmit is required. Therefore, a relationship between horsepower, angular speed, and torque is required. In U.S. customary units and in the International System of Units (SI units) these relationships are

$$\begin{aligned} \text{hp} &= \frac{2\pi nT}{33,000} = \frac{2\pi fT}{5252} \quad (\text{U.S. customary units}) \\ &= \frac{2\pi fT}{9.549} = fT \quad (\text{SI units}) \end{aligned} \quad (10:20)$$

where  $f$  and  $n$  denote the angular speed in cycles per second and revolutions per minute, respectively, and  $T$  denotes the torque transmitted in Newton-meters (N·m) or inch-pounds (in.-lb), depending on the system of units used.

## Kinematics of Circular Shafts

The theory of circular shafts is based on the geometric assumption that a plane cross section simply rotates about the axis of the shaft and can be visualized as being composed of a series of **thin rigid disks** that rotate about the axis of the shaft.

To obtain a formula that expresses the rotation of one cross section relative to another infinitesimally close to it, consider a shaft of radius  $c$  and examine the angular deformations of an interior segment of radius  $r$  and length  $\Delta x$ . This portion of the bar is indicated in Fig. 10.5(a). Before twisting, line element AB is parallel to the shaft axis, and line element AC lies along a cross-sectional circle of radius  $r$ . The angle between these elements is 90 degrees. Due to twisting, AC merely moves to a new location on the circumference, but AB becomes A'B', which is no longer parallel to the shaft axis, as is indicated in Fig. 10.5(b). The shearing deformation  $e_r$  at radius  $r$  is

$$e_r = r\Delta\theta = \gamma_r \Delta x \quad (10:21)$$

where  $\gamma_r$  denotes the shearing strain between line elements A'B' and A'C', and  $\Delta\theta$  represents the angular rotation of the cross section at B relative to the cross section at A. In the limit, as  $\Delta x$  becomes infinitesimal, Eq. (10.21) becomes

$$\gamma_r = r d\theta = dx \quad (10:22)$$

Because a cross section is considered rigid, Eq. (10.22) indicates that the shearing strain varies linearly with distance from the center of the shaft. Consequently, because  $c$  denotes the outside radius of the shaft, the shearing strain at radius  $r$  is

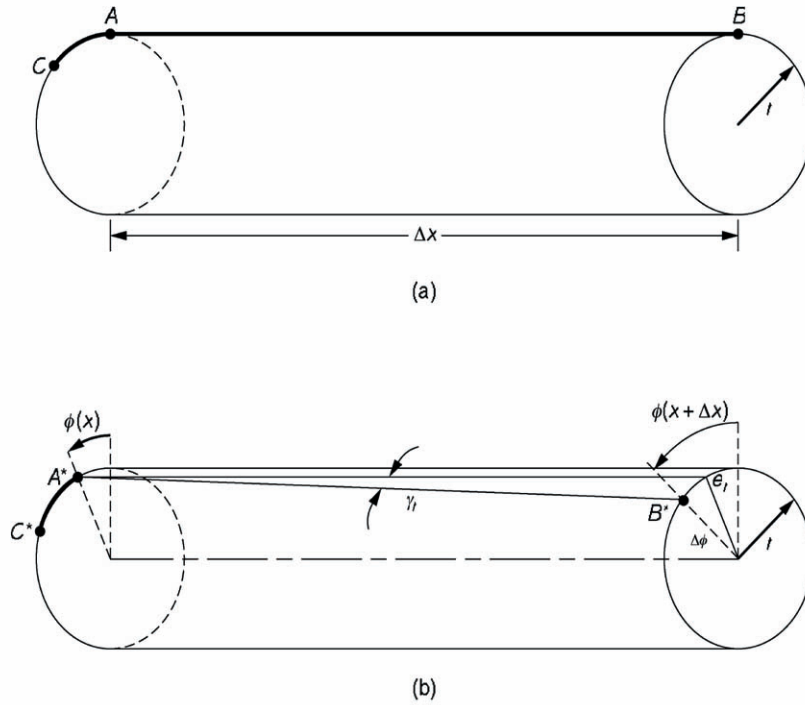
$$\gamma_r = (r/c)\gamma_c \quad (10:23)$$

## Equilibrium

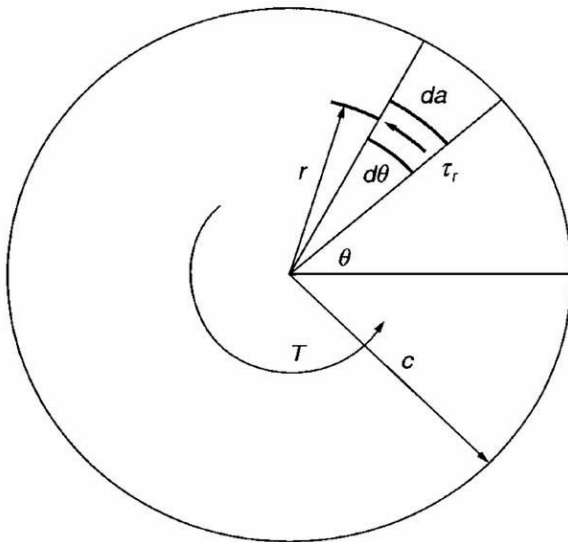
The shearing stress  $\tau_r$  that acts on a differential element of cross-sectional area  $da$  is shown in Fig. 10.6. A concentrated torque  $T$  that is equivalent to the torque produced by the distributed shearing stress  $\tau_r$  is

$$T = \int_{\text{area}} (\tau_r da)r \quad (10:24)$$

**Figure 10.5**



**Figure 10.6**



## Elastic Twisting of Circular Shafts

Explicit formulas for the angle of twist per unit length and for the shearing stress at any point  $r$  in a cross section of a circular shaft made from a linearly elastic material are obtained from Eqs. (10.22) and (10.24) and the stress-strain relation

$$\gamma_r = G \phi_r \quad (10:25)$$

in which  $G$  is the shearing modulus of elasticity. Common units for  $G$  are pounds per square inch (psi) or gigapascals (GPa). Accordingly,

$$T = \int_{\text{area}} (G \tau_r = r) r^2 da = G \int_{\text{area}} d\bar{A} = dx \int_{\text{area}} r^2 da$$

or

$$d\bar{A} = dx = T/JG \quad (10:26)$$

in which  $J$  is the polar moment of inertia of the cross-sectional area of the bar. Common units for  $J$  are inches to the fourth power (in<sup>4</sup>) or meters to the fourth power (m<sup>4</sup>), depending on the system of units used.

The shearing stress at radius  $r$  is obtained by combining Eqs. (10.22), (10.25), and (10.26). Thus,

$$\tau_r = Tr/J \quad (10:27)$$

Equations (10.26) and (10.27) provide the means needed to analyze the strength and stiffness of linearly elastic shafts with circular cross sections. These formulas remain valid for annular shafts for which the hollow and solid portions are concentric. Formulas for the polar moments of inertia  $J$  are

$$J = \begin{cases} \frac{1}{4} \pi d^4 & \text{(solid cross section)} \\ \frac{1}{4} \pi (d_o^4 - d_i^4) & \text{(annular cross section)} \end{cases} \quad (10:28)$$

where  $d_o$  and  $d_i$  denote external and internal diameters.

## Uniform Shaft

A **uniform shaft** is one for which the cross-sectional area, the shearing modulus of elasticity, and the applied torque do not change along its length. Because  $J$ ,  $G$ , and  $T$  are constants over the length  $L$ ; Eq. (10.26) integrates to give the angle of twist of one end relative to the other end as

$$\bar{A} = TL/JG \quad (10:29)$$

The shearing stress on any cross section at radial distance  $r$  is

$$\tau_r = Tr/J \quad (10:30)$$

## Nonuniform Shaft

A **nonuniform shaft** is one for which either  $J$ ;  $G$ ;  $T$ ; or a combination thereof changes abruptly along the length of the shaft. Three procedures are available to determine the angle of twist for circular shafts made from linearly elastic materials.

### Direct Integration

Equation (10.26) is integrated directly. Because the integrand  $T = JG$  can possess discontinuities at cross sections for which  $J$ ;  $G$ ; or  $T$  changes abruptly, the integration must be interpreted as a sum of several integrations. Discontinuities in  $J$ ;  $G$ ; and  $T$  can usually be detected by inspection. The polar moment of inertia  $J$  is discontinuous at abrupt changes in cross-sectional area,  $G$  is discontinuous at cross sections where the material changes abruptly, and the internal torque  $T$  is discontinuous at points where concentrated torques are applied.

### Discrete Elements

The shaft is divided into a finite number of segments for each of which  $T = JG$  is constant. Consequently, the shaft is perceived to be a series of connected uniform shafts for each of which Eq. (10.29) applies. Thus, if  $\hat{A}_i$  denotes the angle of twist of the  $i$ th segment, then the angle of twist for the shaft is

$$\hat{A} = \sum \hat{A}_i \quad (10:31)$$

### Superposition

The superposition principle applied to the twisting of circular shafts stipulates that the relative rotation of one cross section with respect to another cross section due to several torques applied simultaneously is equal to the algebraic sum of the relative rotations of the same cross sections due to each torque applied separately. If  $\hat{A}_{B=A}^0$ ;  $\hat{A}_{B=A}^{00}$ ;  $\dots$  denote relative angles of twist for each torque applied separately, then

$$\hat{A}_{B=A} = \hat{A}_{B=A}^0 + \hat{A}_{B=A}^{00} + \dots \quad (10:32)$$

Superposition of angles of twist requires that the torques be linearly related to the angles of twist that they produce, which in turn implies that the shearing stress must not exceed the proportional limit stress for the material involved. This requirement must be satisfied for each separate loading, as well as for the combined loading.

## Statically Indeterminate Circular Shafts

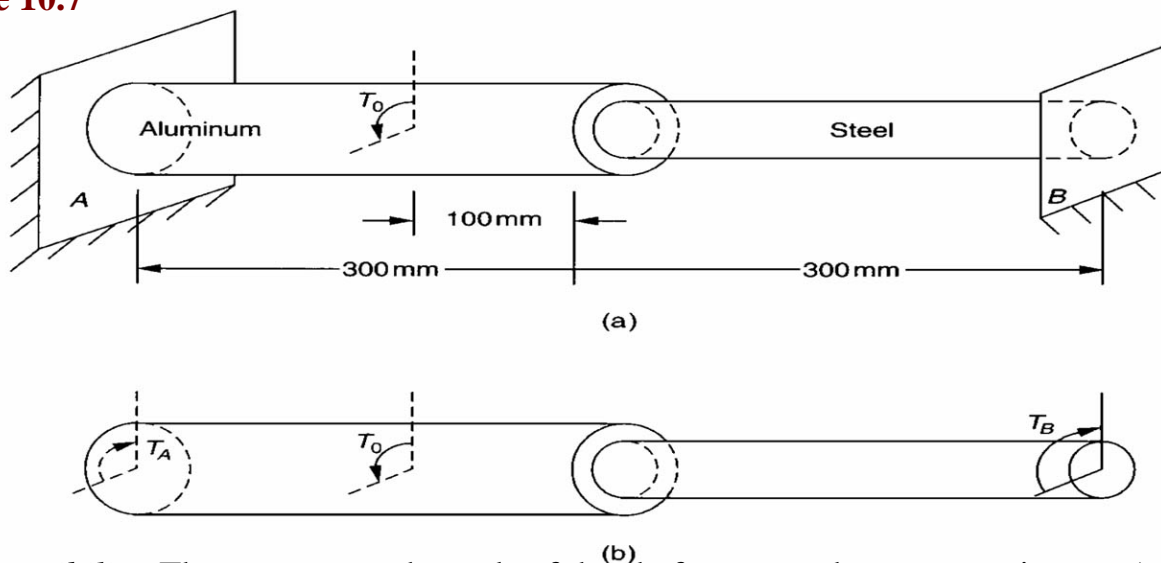
A shaft is statically indeterminate if the internal torque at a cross section cannot be determined from moment equilibrium about the axis of the shaft. In such cases an additional equation is obtained by requiring that angles of twist be compatible with the geometric constraints imposed on the shaft. As with axially loaded bars, three basic concepts of mechanics are involved in the solution of statically indeterminate shafts: equilibrium, geometric compatibility, and material behavior.

**Example 10.3.** The diameters of the aluminum and steel segments of the statically indeterminate step-shaft of Fig. 10.7(a) are 50 mm and 25 mm, respectively. Knowing that  $G_{AL} = 28 \text{ GPa}$ ,  $G_{ST} = 84 \text{ GPa}$ , and  $T_0 = 200 \text{ N} \cdot \text{m}$ , determine the maximum shearing stresses in the aluminum and in the steel.

**Solution.** *Equilibrium.* From Fig. 10.7(b), moment equilibrium about the axis of the shaft gives

$$T_A + T_B + T_0 = 0 \quad (10:33)$$

**Figure 10.7**



*Compatibility.* The supports at the ends of the shaft prevent the cross sections at A and B from rotating; hence, the required compatibility equation is

$$\hat{A}_{B=A} = 0 \quad (10:34)$$

and, with the aid of the superposition principle, it can be written as

$$\hat{A}_{B=A} = \hat{A}_{B=A}^0 + \hat{A}_{B=A}^{00} = 0 \quad (10:35)$$

Here  $\hat{A}_{B=A}^0$  and  $\hat{A}_{B=A}^{00}$  denote the relative angular rotations of the cross section at B with respect to the cross section at A due to the torques  $T_B$  and  $T_0$  acting separately.



To convert Eq. (10.35) into an algebraic equation involving the torques  $T_B$  and  $T_0$ , the discrete element procedure is used. First calculate the polar moments of inertia for the two segments:

$$\begin{aligned} J_{AL} &= \frac{1}{2} \pi (0.050)^4 = 0.613 \times 10^{-6} \text{ m}^4 \\ J_{ST} &= \frac{1}{2} \pi (0.025)^4 = 0.038 \times 10^{-6} \text{ m}^4 \end{aligned} \quad (10.36)$$

Using Eq. (10.29) for a uniform shaft, determine that

$$\begin{aligned} \phi_{B=A}^0 &= 0.3 T_B / J_{AL} (28 \times 10^9) + 0.3 T_B / J_{ST} (84 \times 10^9) = 111.47 \times 10^6 T_B / \text{m}^4 \\ \phi_{B=A}^{00} &= 0.2 T_0 / J_{AL} (28 \times 10^9) = 11.65 \times 10^6 T_0 / \text{m}^4 \end{aligned} \quad (10.37)$$

Consequently,

$$\phi_{B=A} = 111.47 T_B + 11.65 T_0 \times 10^6 = 0 \quad (10.38)$$

Equation (10.38) gives  $T_B$  and Eq. (10.33) gives  $T_A$ : Thus,

$$T_A = 179 \text{ N} \cdot \text{m} \quad \text{and} \quad T_B = 21 \text{ N} \cdot \text{m} \quad (10.39)$$

The maximum shearing stress in each material occurs at the most remote point on a cross section. Thus,

$$\begin{aligned} (\tau_{AL})_{\max} &= T_{AL} C / J_{AL} = 179 (0.025) / 0.613 \times 10^{-6} = 22.9 \text{ MPa} \\ (\tau_{ST})_{\max} &= T_{ST} C / J_{ST} = 21 (0.0125) / 0.038 \times 10^{-6} = 21.7 \text{ MPa} \end{aligned} \quad (10.40)$$

## Defining Terms

**Bar axis:** Straight line locus of centroids of cross sections along the length of a bar.

**Line element:** Imaginary fiber of material along a specific direction.

**Nonuniform bar:** A bar for which the cross-sectional area or the material composition changes abruptly along its length, or external forces are applied intermediate to its ends.

**Nonuniform shaft:** A bar of circular cross section for which the diameter or material composition changes abruptly along its length, or external twisting moments are applied intermediate to its ends.

**Thin rigid disk:** Imaginary circular cross section of infinitesimal thickness that is assumed to undergo no deformations in its plane.

**Torque:** Twisting moment.

**Uniform bar:** A bar of uniform cross-sectional area that is made of one material and is subjected to axial forces only at its ends.

**Uniform shaft:** A bar of uniform, circular cross-sectional area that is made of one material and is subjected to twisting moments only at its ends.

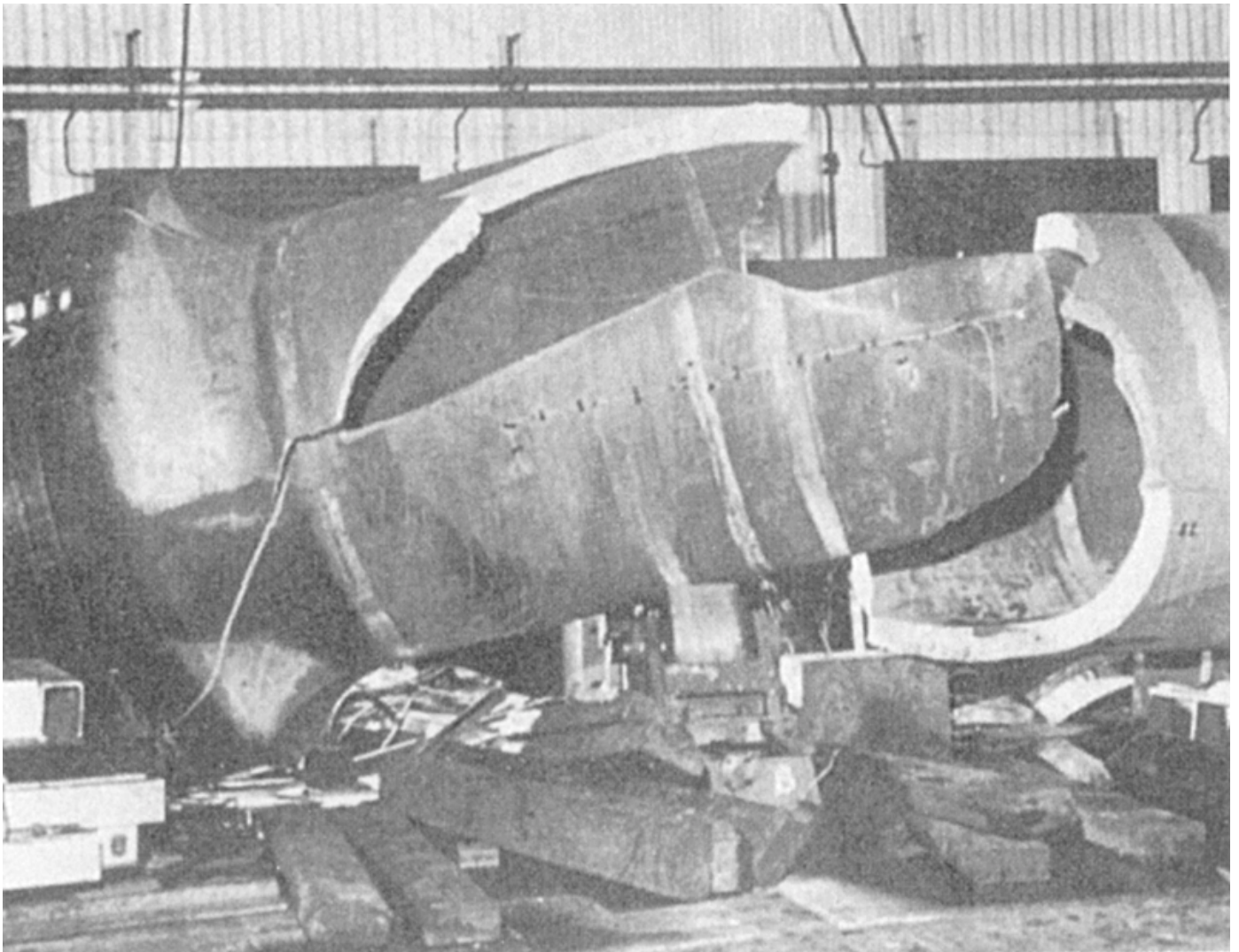
## References

- Bauld, N. R., Jr. 1986. Axially loaded members and torsion. In *Mechanics of Materials*, 2nd ed.
- Beer, F. P. and Johnston, E. R., Jr. 1981. Stress and strain—axial loading and torsion. In *Mechanics of Materials*.
- Gere, J. M. and Timoshenko, S. P. 1990. Axially loaded members and torsion. In *Mechanics of Materials*, 2nd ed.

## Further Information

Formulas for the twisting of shafts with the following cross-sectional shapes can be found in Bauld [1986]: thin-wall, open sections of various shapes; solid elliptical, rectangular, and equilateral triangular sections; open sections composed of thin rectangles; and circular sections composed of two different concentric materials. Also available in the same reference are formulas for the twisting of circular shafts in the inelastic range.

Anderson, T. L. "Mechanics of Materials"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



During hydrostatic testing, *brittle failure* of a high pressure, thick-walled chemical reactor vessel occurred in the manufacturer's test shop. Brittle failure can occur without any prior noticeable deformation; it is characterized by a very rapid crack propagation of up to 6 thousand feet per second. Brittle fracture is the most dangerous type of failure. (Source: Harvey, J. 1974. *Theory and Design of Modern Pressure Vessels*, 2nd ed. Van Nostrand Reinhold. With permission.)

# II

## Mechanics of Materials

---

**Ted L. Anderson**

*Structural Reliability Technology*

- 4    **Reactions** *T. Anagnos*  
Types of Supports • Actual versus Idealized Support Conditions • Static Determinacy and Indeterminacy • Computation of Reactions
- 5    **Bending Stresses in Beams** *J. M. Gere*  
Longitudinal Strains in Beams • Normal Stresses in Beams (Linearly Elastic Materials)
- 6    **Shear Stresses in Beams** *J. M. Gere*  
Shear Stresses in Rectangular Beams • Shear Stresses in Circular Beams • Shear Stresses in the Webs of Beams with Flanges
- 7    **Shear and Moment Diagrams** *G. R. Buchanan*  
Sign Convention • Shear and Moment Diagrams • Shear and Moment Equations
- 8    **Columns** *L. W. Zachary and J. B. Ligon*  
Fundamentals • Examples • Other Forms of Instability
- 9    **Pressure Vessels** *E. Livingston and R. J. Scavuzzo*  
Design Criteria • Design Formulas • Opening Reinforcement
- 10   **Axial Loads and Torsion** *N. R. Bauld, Jr.*  
Axially Loaded Bars • Torsion
- 11   **Fracture Mechanics** *T. L. Anderson*  
Fundamental Concepts • The Energy Criterion • The Stress Intensity Approach • Time-Dependent Crack Growth and Damage Tolerance • Effect of Material Properties on Fracture

THE RESPONSE OF MATERIALS TO STRESS is an important topic in engineering. Excessive stress can lead to failure by plastic deformation, buckling, or brittle fracture. Thus it is essential for the design engineer to estimate stresses correctly and to determine the limit state for the material and structure of interest.

Stress is defined as force per unit cross-sectional area. In general, the stress state in a body can be three-dimensional and can vary from point to point. This section, however, focuses primarily on simple loading cases such as bending and axial loading of beams, columns, and shafts. Readers who are concerned with more complicated situations are encouraged to consult a textbook on solid mechanics or theory of elasticity. Forces and moments that induce stresses in a structure can be considered as either loads or reactions. Wind loading, gravitational forces, and hydrostatic pressure are examples of loads, while reactions are forces that arise from supports that resist movement of a structure that is subject to loads. **Chapter 4** discusses reactions in more detail. Beams and columns are the primary load-bearing members in a range of structures, including bridges and buildings. **Chapters 5 and 6** address bending stresses and shear stresses in beams, respectively, while **Chapter 7** introduces shear and moment diagrams for beams. **Chapter 8** considers buckling

instability in columns subject to compressive axial loads. Vessels that are subject to hydrostatic pressure, such as pressure vessels, pipes, and storage tanks, are another important class of structure. **Chapter 9** lists the basic equations that relate stress to hydrostatic pressure in these configurations. **Chapter 10** covers the analysis of axial and torsion loading in bars and shafts. This chapter addresses both statically determinate and statically indeterminate loading cases. Structures can fail by brittle fracture when the material contains cracks or other flaws. Traditional strength-of-materials approaches do not address brittle fracture or cracks in materials. Fracture mechanics is a relatively new engineering discipline that relates critical combinations of stress, crack size, and a property called *toughness* that quantifies material resistance to fracture. **Chapter 11** gives a brief introduction to this field.

Anagnos, T. "Reactions"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

This chapter is not available because of  
copyright issues



Gere, J. M. "Bending Stresses in Beams"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Bending Stresses in Beams

This chapter contains selected material (text and figures) from Gere, J. M. and Timoshenko, S. P. 1990. *Mechanics of Materials*, 3rd ed. PWS, Boston. With permission.

### 5.1 Longitudinal Strains in Beams

### 5.2 Normal Stresses in Beams (Linearly Elastic Materials)

**James M. Gere**

*Stanford University*

A *beam* is a slender structural member subjected to lateral loads. In this chapter we consider the bending stresses (i.e., normal stresses) in beams having initially straight longitudinal axes, such as the cantilever beam of Fig. 5.1(a). For reference, we direct the positive  $x$  axis to the right along the longitudinal axis of the beam and the positive  $y$  axis downward (because the deflections of most beams are downward). The  $z$  axis, which is not shown in the figure, is directed away from the viewer, so that the three axes form a right-handed coordinate system. All cross sections of the beam are assumed to be symmetric about the  $xy$  plane, and all loads are assumed to act in this plane. Consequently, the beam will deflect in this same plane [Fig. 5.1(b)], which is called the **plane of bending**.

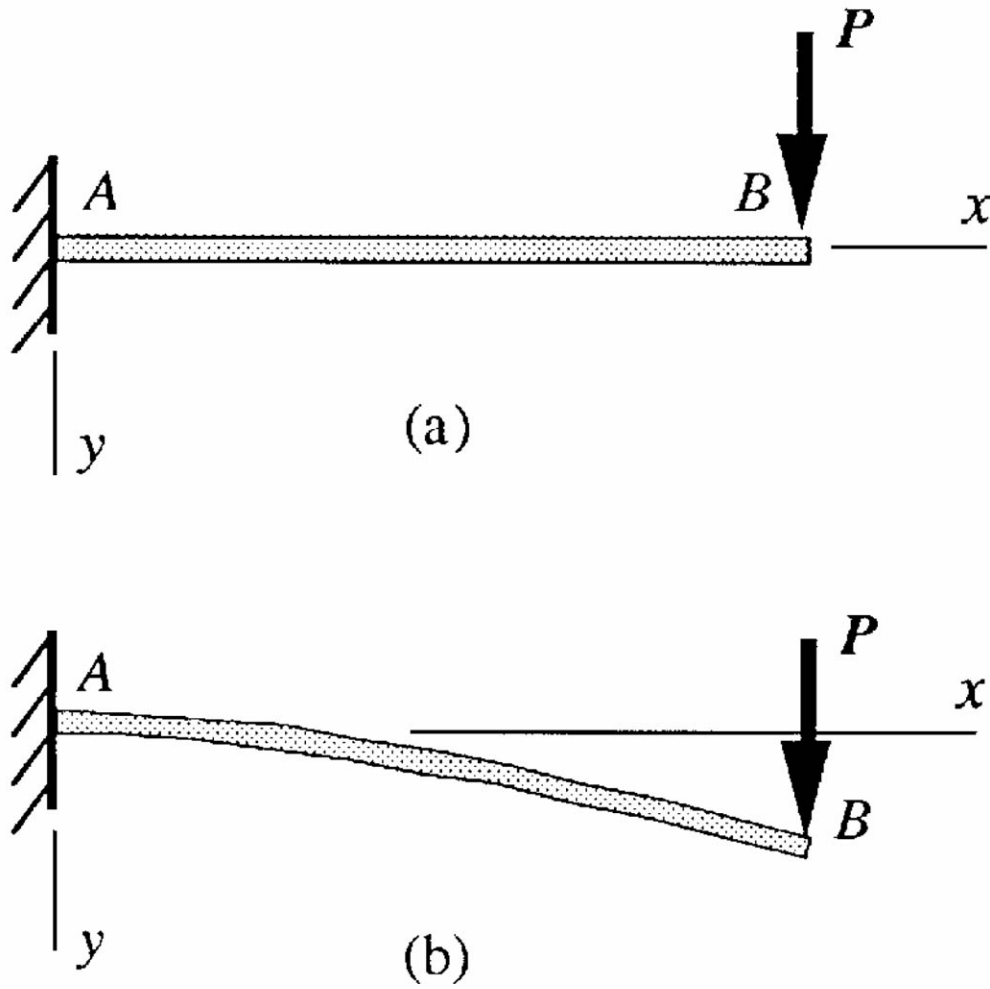
**Pure bending** refers to bending of a beam under a constant bending moment  $M$ , which means that the shear force  $V$  is zero (because  $V = dM/dx$ ). **Nonuniform bending** refers to bending in the presence of shear forces, in which case the bending moment varies along the axis of the beam. The sign convention for bending moments is shown in Fig. 5.2; note that positive bending moment produces tension in the lower part of the beam and compression in the upper part.

The stresses and strains in a beam are directly related to the *curvature*  $\kappa$  of the deflection curve. Because the  $x$  axis is positive to the right and the  $y$  axis is positive downward, the curvature is positive when the beam is bent concave downward and negative when the beam is bent concave upward (Fig. 5.2).

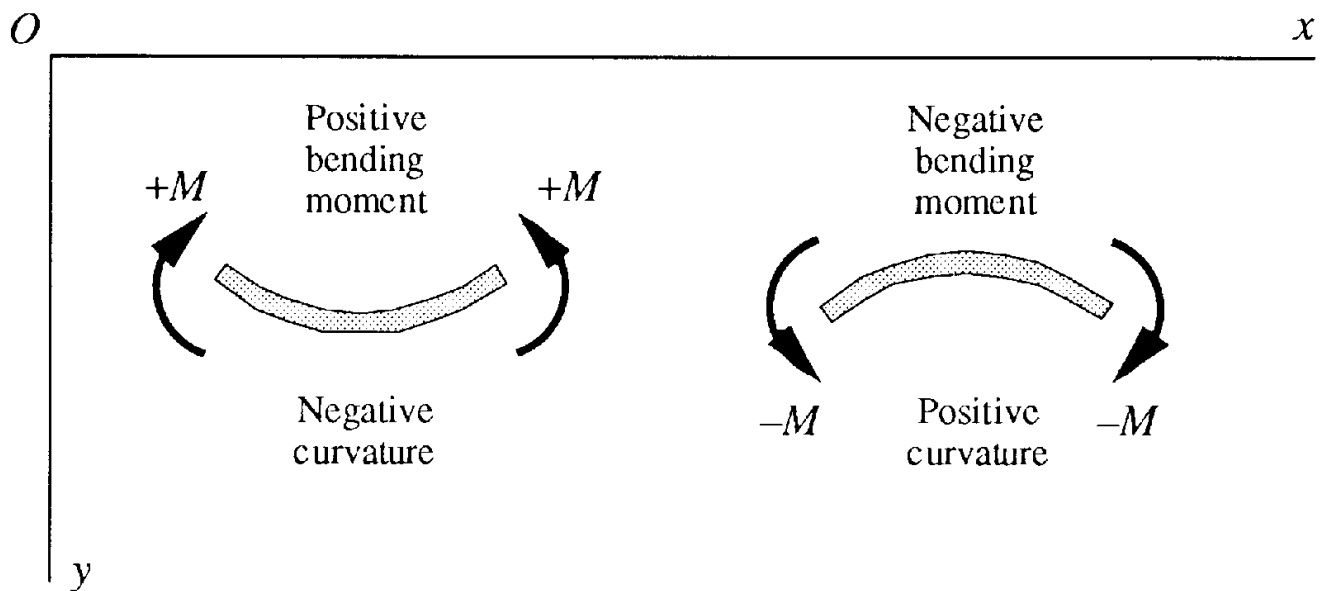
## 5.1 Longitudinal Strains in Beams

Consider a segment  $DE$  of a beam subjected to pure bending by positive bending moments  $M$  [Fig. 5.3(a)]. The cross section of the beam at section  $mn$  is of arbitrary shape except that it must be symmetrical about the  $y$  axis [Fig. 5.3(b)]. All cross sections of the beam (such as  $mn$ ) that were plane before bending remain plane after bending, a fact that can be proven theoretically using arguments based on symmetry. Therefore, *plane sections remain plane regardless of the material properties, whether elastic or inelastic, linear or nonlinear*. (Of course, the material properties, like the dimensions, must be symmetric about the plane of bending.)

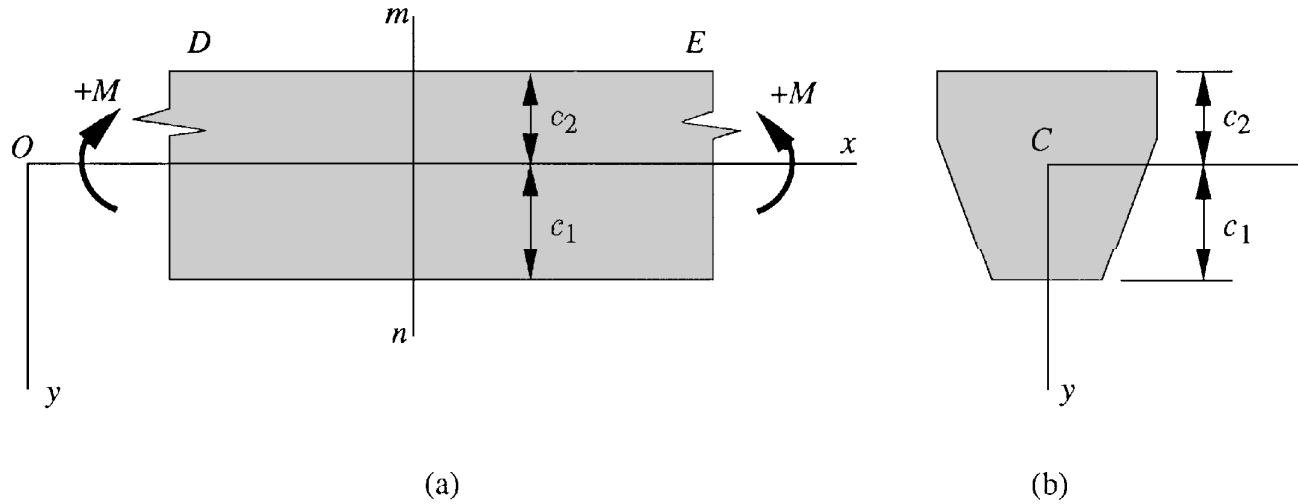
**Figure 5.1** Bending of a cantilever beam.



**Figure 5.2** Sign conventions for bending moment and curvature.



**Figure 5.3** Beam in pure bending. (a) Side view of segment of beam showing bending moments  $M$  and typical section  $mn$ . (b) Cross section of beam at section  $mn$ .



With positive bending moments, the lower part of the beam is in tension and the upper part is in compression. Therefore, longitudinal lines (i.e., line segments parallel to the  $x$  axis) in the lower part of the beam are elongated and those in the upper part are shortened. The intermediate surface in which longitudinal lines do not change in length is called the **neutral surface** of the beam. We place the origin  $O$  of coordinates in this plane, so that the  $xz$  plane becomes the neutral surface. The intersection of this surface with any cross-sectional plane is called the **neutral axis of the cross section**, for instance, the  $z$  axis in Fig. 5.3(b).

Because plane sections remain plane, the longitudinal strains  $\varepsilon_x$  in the beam vary linearly with the distance  $y$  from the neutral surface, regardless of the material properties. It can also be shown that the strains are proportional to the curvature  $\kappa$ . Thus, the strains are given by the equation

$$\varepsilon_x = -\kappa y \quad (5.1)$$

The sign convention for  $\varepsilon_x$  is positive for elongation and negative for shortening. Note that when the curvature is positive (Fig. 5.2) and  $y$  is positive (Fig. 5.3), the strain is negative.

## 5.2 Normal Stresses in Beams (Linearly Elastic Materials)

Since longitudinal line elements in the beam are subjected only to tension or compression (elongation or shortening), they are in a state of uniaxial stress. Therefore, we can use the stress-strain diagram of the material to obtain the normal stresses  $\sigma_x$  from the normal strains  $\varepsilon_x$ . If the shape of the stress-strain curve can be expressed analytically, a formula can be derived for the stresses in the beam; otherwise, they must be calculated numerically.

The simplest and most common stress-strain relationship is for a linearly elastic material, in which case we can combine Hooke's law for uniaxial stress ( $\sigma = E\varepsilon$ ) with Eq. (5.1) and obtain

$$\sigma_x = E\varepsilon_x = -E\kappa y \quad (5.2)$$

in which  $E$  is the modulus of elasticity of the material. Equation (5.2) shows that the normal stresses acting on a cross section vary linearly with the distance  $y$  from the neutral surface when the material follows Hooke's law.

Since the beam is in pure bending (Fig. 5.3), the resultant of the stresses  $\sigma_x$  acting over the cross section must equal the bending moment  $M$ . This observation provides two equations of statics—the first expressing that the resultant force in the  $x$  direction is equal to zero and the second expressing that the resultant moment is equal to  $M$ . The first equation of statics leads to the equation

$$\int y \, dA = 0 \quad (5.3)$$

which shows that the first moment of the cross-sectional area with respect to the  $z$  axis is zero. Therefore, the  $z$  axis must pass through the *centroid* of the cross section. Since the  $z$  axis is also the neutral axis, we arrive at the following conclusion: The neutral axis passes through the centroid  $C$  of the cross section provided the material follows Hooke's law and there is no axial force acting on the cross section.

Since the  $y$  axis is an axis of symmetry, the  $y$  axis also passes through the centroid. Therefore, the origin of coordinates  $O$  is located at the centroid  $C$  of the cross section. Furthermore, the symmetry of the cross section about the  $y$  axis means that the  $y$  axis is a *principal axis*. The  $z$  axis is also a principal axis since it is perpendicular to the  $y$  axis. Therefore, when a beam of linearly elastic material is subjected to pure bending, the  $y$  and  $z$  axes are principal centroidal axes.

The second equation of statics leads to the *moment-curvature equation*

$$M = -\kappa EI \quad (5.4)$$

in which

$$I = \int y^2 \, dA \quad (5.5)$$

is the moment of inertia of the cross-sectional area with respect to the  $z$  axis (that is, with respect to the neutral axis). Moments of inertia have dimensions of length to the fourth power, and typical units are in.<sup>4</sup>, mm<sup>4</sup>, and m<sup>4</sup> for beam calculations. The quantity  $EI$  is a measure of the resistance of the beam to bending and is called the *flexural rigidity* of the beam.

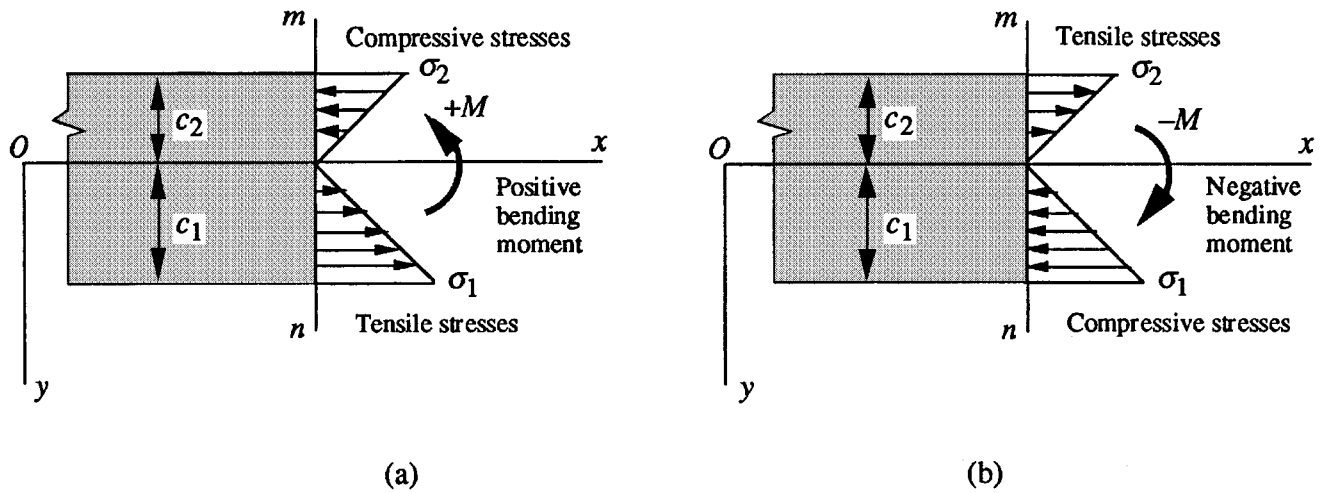
The minus sign in the moment-curvature equation is a consequence of the sign conventions we have adopted for bending moments and coordinate axes (Fig. 5.2). We see that a positive bending moment produces negative curvature and a negative bending moment produces positive curvature. If the opposite sign convention for bending moments is used, or if the  $y$  axis is positive upward, then the minus sign is omitted in Eq. (5.4) but a minus sign must be inserted in the flexure formula [Eq. (5.6)] that follows.

The normal stresses in the beam can be related to the bending moment  $M$  by eliminating the curvature  $\kappa$  between Eqs. (5.2) and (5.4), yielding

$$\sigma_x = \frac{My}{I} \quad (5.6)$$

This equation, called the **flexure formula**, shows that the stresses are directly proportional to the bending moment  $M$  and inversely proportional to the moment of inertia  $I$  of the cross section. Furthermore, the stresses vary linearly with the distance  $y$  from the neutral axis, as shown in Fig. 5.4. Stresses calculated from the flexure formula are called **bending stresses**.

**Figure 5.4** Bending stresses obtained from the flexure formula.



The maximum tensile and compressive bending stresses occur at points located farthest from the neutral axis. Let us denote by  $c_1$  and  $c_2$  the distances from the neutral axis to the extreme elements in the positive and negative  $y$  directions, respectively (see Figs. 5.3 and 5.4). Then the corresponding maximum normal stresses  $\sigma_1$  and  $\sigma_2$  are

$$\sigma_1 = \frac{Mc_1}{I} = \frac{M}{S_1} \quad \sigma_2 = -\frac{Mc_2}{I} = -\frac{M}{S_2} \quad (5.7)$$

in which

$$S_1 = \frac{I}{c_1} \quad S_2 = \frac{I}{c_2} \quad (5.8)$$

The quantities  $S_1$  and  $S_2$  are known as the **section moduli** of the cross-sectional area. From Eq. (5.8) we see that a section modulus has dimensions of length to the third power (for example, in.<sup>3</sup>, mm<sup>3</sup>, or m<sup>3</sup>).

If the cross section is symmetric with respect to the  $z$  axis, which means that it is a *doubly*

*symmetric cross section*, then  $c_1 = c_2 = c$ , and the maximum tensile and compressive stresses are equal numerically:

$$\sigma_1 = -\sigma_2 = \frac{Mc}{I} = \frac{M}{S} \quad (5.9)$$

in which

$$S = \frac{I}{c} \quad (5.10)$$

is the section modulus. For a beam of *rectangular cross section* with width  $b$  and height  $h$  [Fig. 5.5(a)], the moment of inertia and section modulus are

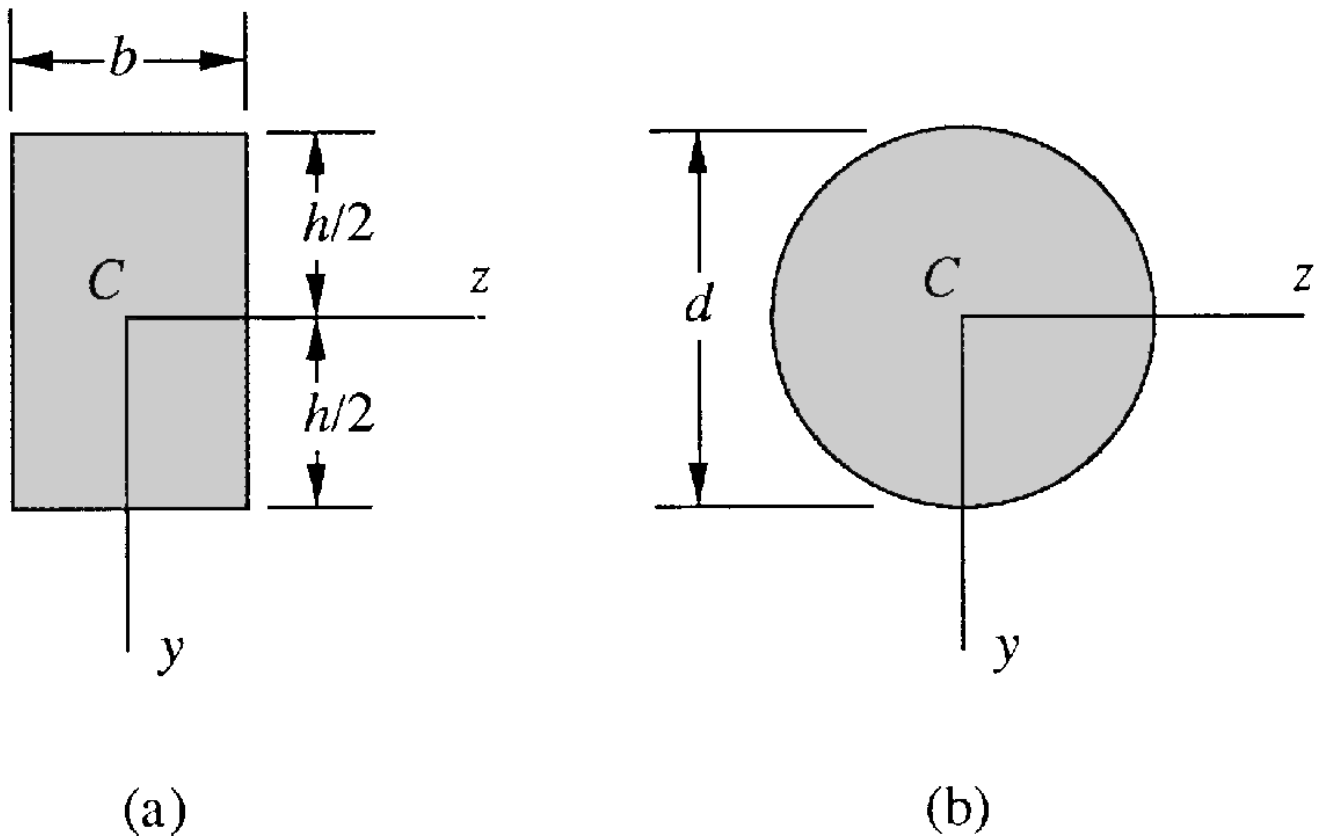
$$I = \frac{bh^3}{12} \quad S = \frac{bh^2}{6} \quad (5.11)$$

For a *circular cross section* of diameter  $d$  [Fig. 5.5(b)], these properties are

$$I = \frac{\pi d^4}{64} \quad S = \frac{\pi d^3}{32} \quad (5.12)$$

The properties of many other plane figures are listed in textbooks and handbooks.

**Figure 5.5** Doubly symmetric cross-sectional shapes.

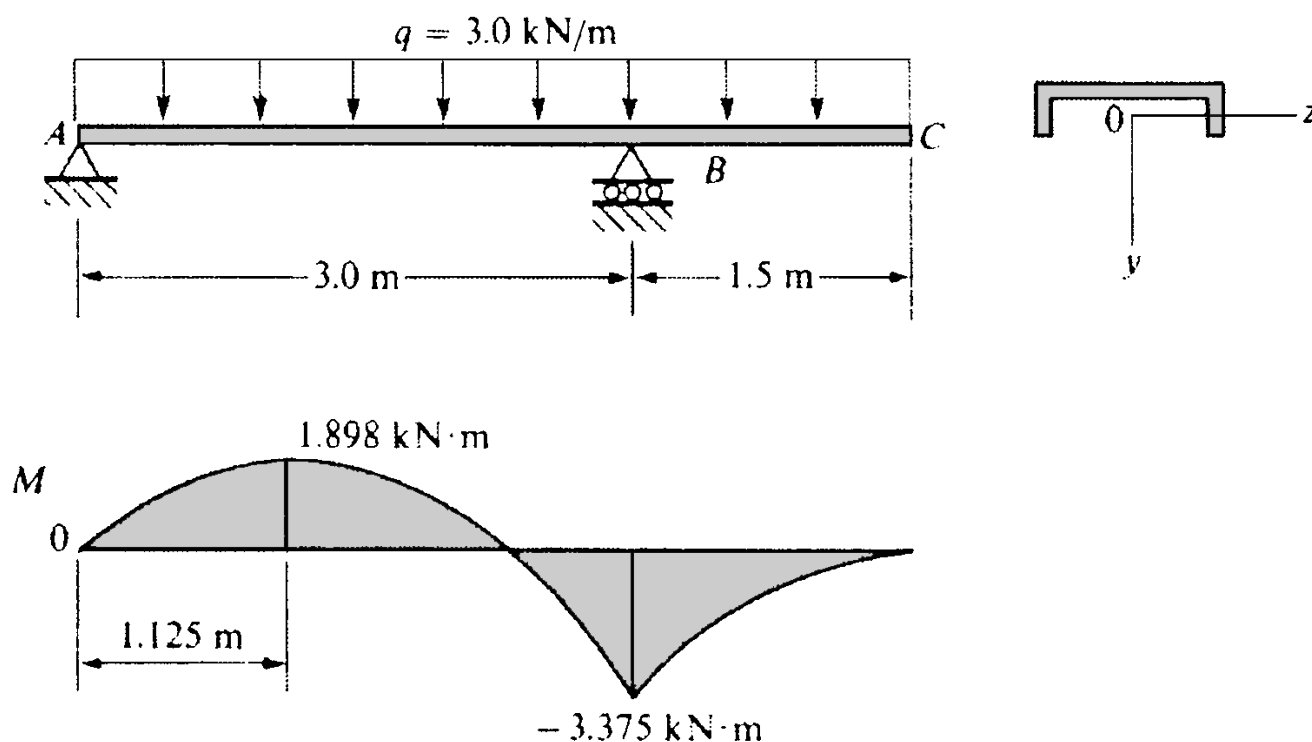


The preceding equations for the normal stresses apply rigorously only for pure bending, which means that no shear forces act on the cross sections. The presence of shear forces produces warping, or out-of-plane distortion, of the cross sections, and a cross section that is plane before bending is no longer plane after bending. Warping due to shear greatly complicates the behavior of the beam, but detailed investigations show that the normal stresses calculated from the flexure formula are not significantly altered by the presence of the shear stresses and the associated warping. Thus, under ordinary conditions we may use the flexure formula for calculating normal stresses even when we have nonuniform bending.

The flexure formula gives results that are accurate only in regions of the beam where the stress distribution is not disrupted by abrupt changes in the shape of the beam or by discontinuities in loading. For instance, the flexure formula is not applicable at or very near the supports of a beam, where the stress distribution is irregular. Such irregularities produce localized stresses, or *stress concentrations*, that are much greater than the stresses obtained from the flexure formula. With ductile materials and static loads, we may usually disregard the effects of stress concentrations. However, they cannot be ignored when the materials are brittle or when the loads are dynamic in character.

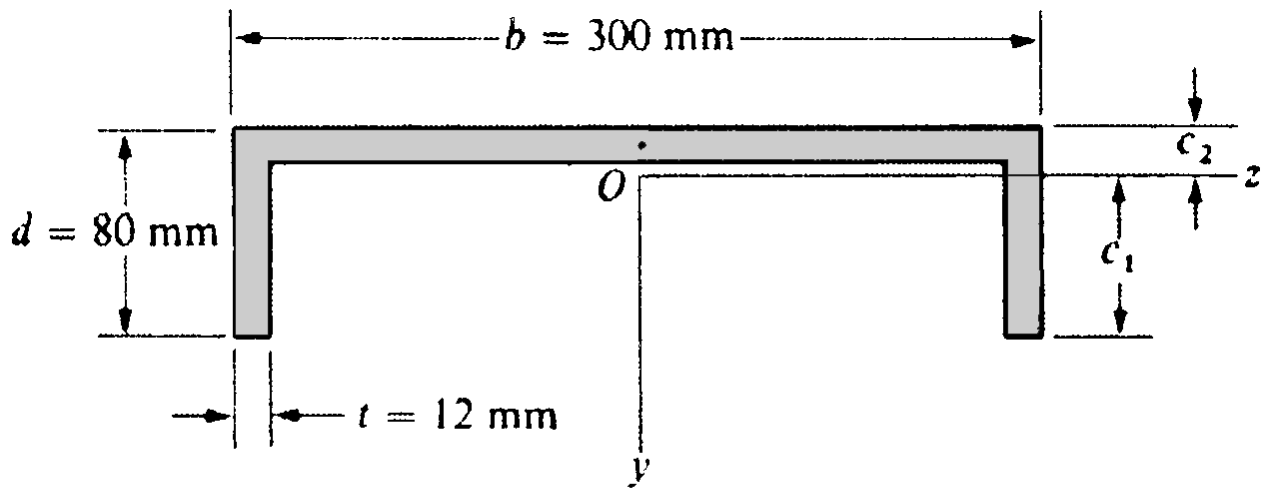
**Example.** The beam  $ABC$  shown in Fig. 5.6 has simple supports at  $A$  and  $B$  and an overhang from  $B$  to  $C$ . A uniform load of intensity  $q = 3.0 \text{ kN/m}$  acts throughout the length of the beam. The beam is constructed of steel plates (12 mm thick) welded to form a channel section, the dimensions of which are shown in Fig. 5.7(a). Calculate the maximum tensile and compressive stresses in the beam due to the uniform load.

**Figure 5.6** Beam dimensions.

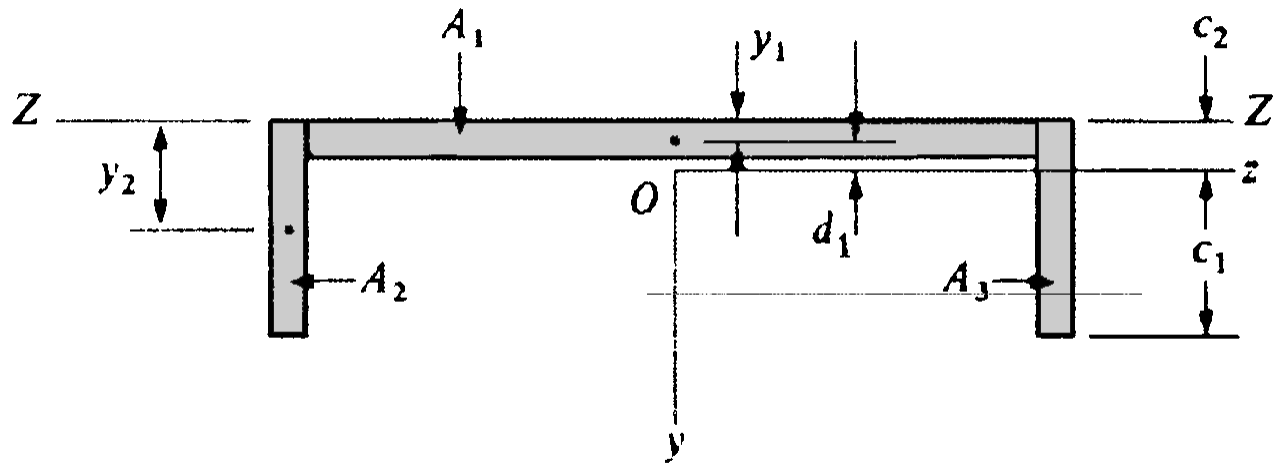




**Figure 5.7** Cross section of beam.



(a)



(b)

**Solution.** The maximum tensile and compressive stresses occur at the cross sections where the bending moments have their maximum numerical values. Therefore, we construct the bending-moment diagram for the beam (Fig. 5.6) and note that the maximum positive and negative moments equal  $1.898 \text{ kN} \cdot \text{m}$  and  $-3.375 \text{ kN} \cdot \text{m}$ , respectively.

Next, we determine the position of the neutral axis by locating the centroid of the cross-sectional area shown in Fig. 5.7(a). The results are as follows:

$$c_1 = 61.52 \text{ mm} \quad c_2 = 18.48 \text{ mm}$$

The moment of inertia of the cross-sectional area about the neutral axis (the  $z$  axis) is calculated

with the aid of the parallel-axis theorem for moments of inertia; the result is

$$I = 2.469 \cdot 10^6 \text{ mm}^4$$

Also, the section moduli for the bottom and top of the beam, respectively, are

$$S_1 = \frac{I}{c_1} = 40\,100 \text{ mm}^3 \quad S_2 = \frac{I}{c_2} = 133\,600 \text{ mm}^3$$

At the cross section of maximum positive bending moment, the largest tensile stress occurs at the bottom of the beam ( $\sigma_1$ ) and the largest compressive stress occurs at the top ( $\sigma_2$ ) :

$$\sigma_t = \sigma_1 = \frac{M}{S_1} = \frac{1.898 \text{ kN} \cdot \text{m}}{40\,100 \text{ mm}^3} = 47.3 \text{ MPa}$$

$$\sigma_c = \sigma_2 = -\frac{M}{S_2} = -\frac{1.898 \text{ kN} \cdot \text{m}}{133\,600 \text{ mm}^3} = -14.2 \text{ MPa}$$

Similarly, the largest stresses at the section of maximum negative moment are

$$\sigma_t = \sigma_2 = -\frac{M}{S_2} = -\frac{-3.375 \text{ kN} \cdot \text{m}}{133\,600 \text{ mm}^3} = 25.3 \text{ MPa}$$

$$\sigma_c = \sigma_1 = \frac{M}{S_1} = \frac{-3.375 \text{ kN} \cdot \text{m}}{40\,100 \text{ mm}^3} = -84.2 \text{ MPa}$$

A comparison of these four stresses shows that the maximum tensile stress due to the uniform load  $q$  is 47.3 MPa and occurs at the bottom of the beam at the section of maximum positive bending moment. The maximum compressive stress is  $-84.2 \text{ MPa}$  and occurs at the bottom of the beam at the section of maximum negative moment.

## Defining Terms

**Bending stresses:** Longitudinal normal stresses  $\sigma_x$  in a beam due to bending moments.

**Flexure formula:** The formula  $\sigma_x = My/I$  for the bending stresses in a beam (linearly elastic materials only).

**Neutral axis of the cross section:** The intersection of the neutral surface with a cross-sectional plane; that is, the line in the cross section about which the beam bends and where the bending stresses are zero.

**Neutral surface:** The surface perpendicular to the plane of bending in which longitudinal lines in the beam do not change in length (no longitudinal strains).

**Nonuniform bending:** Bending in the presence of shear forces (which means that the bending moment varies along the axis of the beam).

**Plane of bending:** The plane of symmetry in which a beam bends and deflects.

**Pure bending:** Bending of a beam under a constant bending moment (no shear forces).

**Section modulus:** A property of the cross section of a beam, equal to  $I/c$  see Eq. (5.8).

## References

- Beer, F. P., Johnston, E. R., and DeWolf, J. T. 1992. *Mechanics of Materials*, 2nd ed. McGraw-Hill, New York.
- Gere, J. M. and Timoshenko, S. P. 1990. *Mechanics of Materials*, 3rd ed. PWS, Boston.
- Hibbeler, R. C. 1991. *Mechanics of Materials*. Macmillan Publishing Company, New York.
- Popov, E. P. 1990. *Engineering Mechanics of Solids*. Prentice Hall, Englewood Cliffs, NJ.
- Riley, W. F. and Zachary, L. 1989. *Introduction to Mechanics of Materials*. John Wiley & Sons, New York.

## Further Information

Extensive discussions of bending, with derivations, examples, and problems, can be found in textbooks on mechanics of materials, such as those listed in the References. These books also cover many additional topics pertaining to bending stresses in beams. For instance, nonprismatic beams, fully stressed beams, beams with axial loads, stress concentrations in bending, composite beams, beams with skew loads, and stresses in inelastic beams are discussed in Gere and Timoshenko [1990].

Gere, J. M. "Shear Stresses in Beams"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Shear Stresses in Beams

Selected material (text and figures) from Chapter 5 of Gere, J. M. and Timoshenko, S. P. 1990. *Mechanics of Materials*, 3rd ed. PWS, Boston. With permission.

## 6.1 Shear Stresses in Rectangular Beams

## 6.2 Shear Stresses in Circular Beams

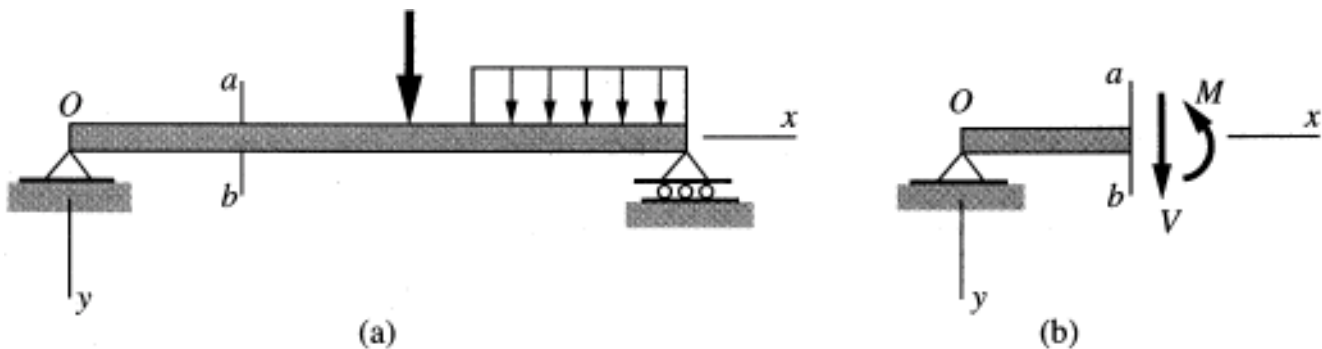
## 6.3 Shear Stresses in the Webs of Beams with Flanges

### James M. Gere

Stanford University

The loads acting on a beam [Fig. 6.1(a)] usually produce both bending moments  $M$  and shear forces  $V$  at cross sections such as  $ab$  [Fig. 6.1(b)]. The longitudinal normal stresses  $\sigma_x$  associated with the bending moments can be calculated from the flexure formula (see **Chapter 5**). The transverse shear stresses  $\tau_{xy}$  associated with the shear forces are described in this chapter.

**Figure 6.1** Beam with bending moment  $M$  and shear force  $V$  acting at cross section  $ab$ .

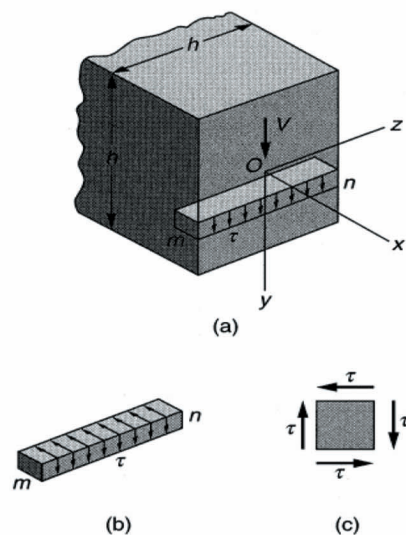


Since the formulas for shear stresses are derived from the flexure formula, they are subject to the same limitations: (1) the beam is symmetric about the  $xy$  plane and all loads act in this plane (the *plane of bending*); (2) the beam is constructed of a linearly elastic material; and (3) the stress distribution is not disrupted by abrupt changes in the shape of the beam or by discontinuities in loading (*stress concentrations*).

## 6.1 Shear Stresses in Rectangular Beams

A segment of a beam of rectangular cross section (width  $b$  and height  $h$ ) subjected to a vertical shear force  $V$  is shown in Fig. 6.2(a). We assume that the shear stresses  $\tau$  acting on the cross section are parallel to the sides of the beam and uniformly distributed across the width (although they vary as we move up or down on the cross section). A small element of the beam cut out between two adjacent cross sections and between two planes that are parallel to the neutral surface is shown in Fig. 6.2(a) as element  $m$  $n$ . Shear stresses acting on one face of an element are always accompanied by complementary shear stresses of equal magnitude acting on perpendicular faces of the element, as shown in Figs. 6.2(b) and 6.2(c). Thus, there are horizontal shear stresses acting between horizontal layers of the beam as well as transverse shear stresses acting on the vertical cross sections.

**Figure 6.2** Shear stresses in a beam of rectangular cross section.



The equality of the horizontal and vertical shear stresses acting on element  $m$  $n$  leads to an interesting conclusion regarding the shear stresses at the top and bottom of the beam. If we imagine that the element  $m$  $n$  is located at either the top or the bottom, we see that the horizontal shear stresses vanish because there are no stresses on the outer surfaces of the beam. It follows that the vertical shear stresses also vanish at those locations; thus,  $\tau = 0$  where  $y = \pm h/2$ . (Note that the origin of coordinates is at the centroid of the cross section and the  $z$  axis is the neutral axis.)

The magnitude of the shear stresses can be determined by a lengthy derivation that involves only the flexure formula and static equilibrium (see References). The result is the following formula for the shear stress:

$$\tau = \frac{V}{Ib} \int y \, dA \quad (6:1)$$

in which  $V$  is the shear force acting on the cross section,  $I$  is the moment of inertia of the cross-sectional area about the neutral axis, and  $b$  is the width of the beam. The integral in Eq. (6.1) is the first moment of the part of the cross-sectional area below (or above) the level at which the stress is being evaluated. Denoting this first moment by  $Q$ , that is,

$$Q = \int y \, dA \quad (6:2)$$

we can write Eq. (6.1) in the simpler form

$$\tau = \frac{VQ}{Ib} \quad (6:3)$$

This equation, known as the **shear formula**, can be used to determine the shear stress  $\tau$  at any point in the cross section of a rectangular beam. Note that for a specific cross section, the shear force  $V$ , moment of inertia  $I$ , and width  $b$  are constants. However, the first moment  $Q$  (and hence the shear stress  $\tau$ ) varies depending upon where the stress is to be found.

To evaluate the shear stress at distance  $y_1$  below the neutral axis (Fig. 6.3), we must determine the first moment  $Q$  of the area in the cross section below the level  $y = y_1$ . We can obtain this first moment by multiplying the partial area  $A_1$  by the distance  $\bar{y}_1$  from its centroid to the neutral axis:

$$Q = A_1 \bar{y}_1 = b \left[ \frac{h}{2} - y_1 \right] y_1 + \frac{b(h - 2y_1)^2}{2} = \frac{b}{2} \left[ \frac{h^2}{4} - y_1^2 \right] \quad (6:4)$$

Of course, this same result can be obtained by integration using Eq. (6.2):

$$Q = \int y \, dA = \int_{y_1}^{h/2} yb \, dy = \frac{b}{2} \left[ \frac{h^2}{4} - y_1^2 \right] \quad (6:5)$$

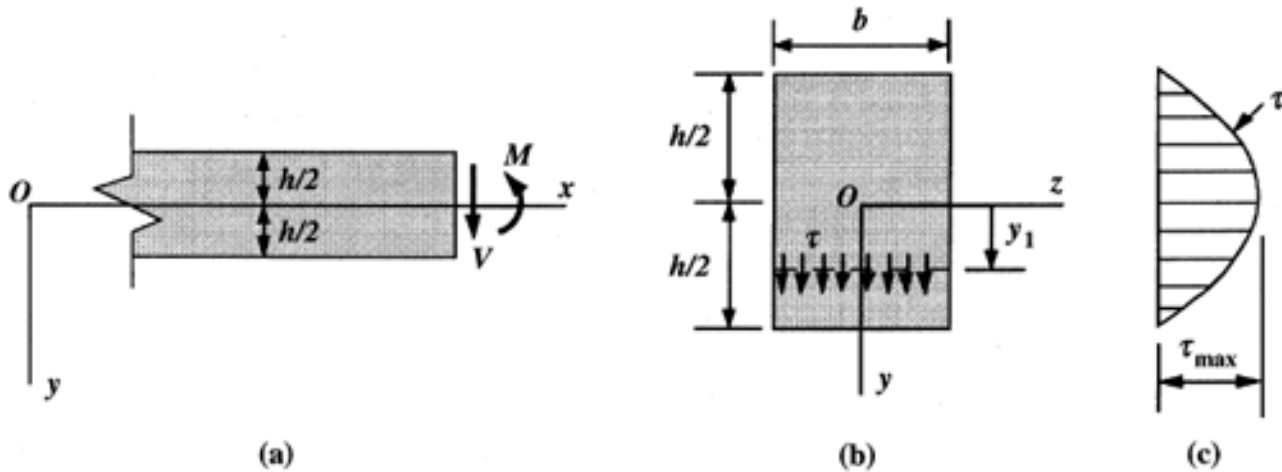
Substituting this expression for  $Q$  into the shear formula [Eq. (6.3)], we get

$$\tau = \frac{V}{2I} \left[ \frac{h^2}{4} - y_1^2 \right] \quad (6:6)$$

This equation shows that the shear stresses in a rectangular beam vary quadratically with the distance  $y_1$  from the neutral axis. Thus, when plotted over the height of the beam,  $\tau$  varies in

the manner shown by the parabolic diagram of Fig. 6.3(c). Note that the shear stresses are zero when  $y_1 = \pm h/2$ :

**Figure 6.3** Distribution of shear stresses in a beam of rectangular cross section. (a) Side view of beam showing the shear force  $V$  and bending moment  $M$  acting at a cross section. (b) Cross section of beam showing shear stresses  $\tau$  acting at distance  $y_1$  from the neutral axis. (c) Diagram showing the parabolic distribution of shear stresses.



The maximum value of the shear stress occurs at the neutral axis, where the first moment  $Q$  has its maximum value. Substituting  $y_1 = 0$  into Eq. (6.6), we get

$$\tau_{\max} = \frac{V h^2}{8I} = \frac{3V}{2A} \quad (6.7)$$

in which  $A = bh$  is the cross-sectional area. Thus, the maximum shear stress is 50% larger than the average shear stress (equal to  $V/A$ ): Note that the preceding equations for the shear stresses can be used to calculate either vertical shear stresses acting on a cross section or horizontal shear stresses acting between horizontal layers of the beam.

The shear formula is valid for rectangular beams of ordinary proportions—it is exact for very narrow beams (width  $b$  much less than height  $h$ ) but less accurate as  $b$  increases relative to  $h$ . For instance, when  $b = h$ ; the true maximum shear stress is about 13% larger than the value given by Eq. (6.7).

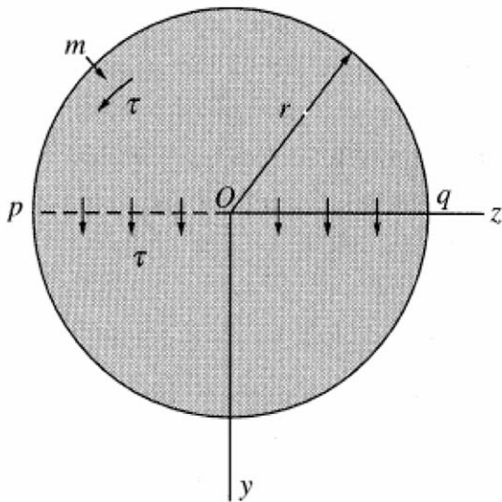
A common error is to apply the shear formula to cross-sectional shapes, such as a triangle, for which it is not applicable. The reasons it does not apply to a triangle are (1) we assumed the cross section had sides parallel to the  $y$  axis (so that the shear stresses acted parallel to the  $y$  axis), and (2) we assumed that the shear stresses were uniform across the width of the cross section. These assumptions hold only in particular cases, including beams of narrow rectangular cross section.



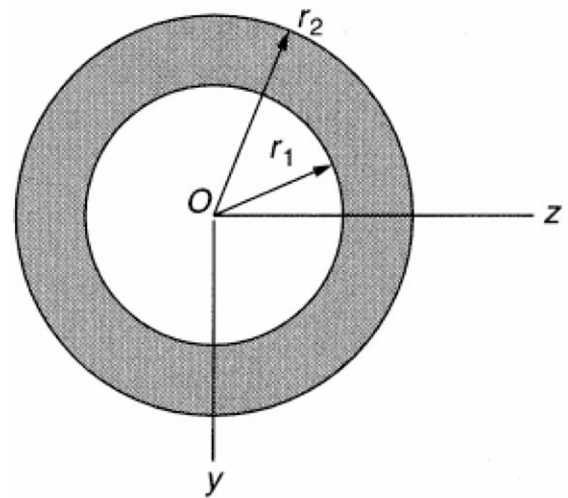
## 6.2 Shear Stresses in Circular Beams

When a beam has a circular cross section (Fig. 6.4), we can no longer assume that all of the shear stresses act parallel to the  $y$  axis. For instance, we can easily demonstrate that at a point on the boundary of the cross section, such as point  $m$ , the shear stress  $\tau$  acts tangent to the boundary. This conclusion follows from the fact that the outer surface of the beam is free of stress, and therefore the shear stress acting on the cross section can have no component in the radial direction (because shear stresses acting on perpendicular planes must be equal in magnitude).

**Figure 6.4** Shear stresses in a beam of circular cross section.



**Figure 6.5** Shear stresses in a beam of hollow circular cross section.



Although there is no simple way to find the shear stresses throughout the entire cross section, we can readily determine the stresses at the neutral axis (where the stresses are the largest) by making some reasonable assumptions about the stress distribution. We assume that the stresses act parallel to the  $y$  axis and have constant intensity across the width of the beam (from point  $p$  to point  $q$  in Fig. 6.4). Inasmuch as these assumptions are the same as those used in deriving the shear formula [Eq. (6.3)], we can use that formula to calculate the shear stresses at the neutral axis. For a cross section of radius  $r$ , we obtain

$$I = \frac{\frac{1}{4}r^4}{4} \quad b = 2r \quad (6:8)$$

$$Q = A_1 \bar{y}_1 = \frac{\frac{1}{4}r^2}{2} \frac{4r}{\frac{3}{4}} = \frac{2r^3}{3}$$

in which  $Q$  is the first moment of a semicircle. Substituting these expressions for  $I$ ,  $b$ , and  $Q$  into the shear formula, we obtain

$$\tau_{\max} = \frac{VQ}{Ib} = \frac{V(2r^3/3)}{(\pi r^4/4)(2r)} = \frac{4V}{3\pi r^2} = \frac{4V}{3A} \quad (6:9)$$

in which  $A$  is the area of the cross section. This equation shows that the maximum shear stress in a circular beam is equal to 4/3 times the average shear stress  $V/A$ :

Although the preceding theory for the maximum shear stress in a circular beam is approximate, it gives results that differ by only a few percent from those obtained by more exact theories.

If a beam has a *hollow circular cross section* (Fig. 6.5), we may again assume with good accuracy that the shear stresses along the neutral axis are parallel to the  $y$  axis and uniformly distributed. Then, as before, we may use the shear formula to find the maximum shear stress. The properties of the hollow section are

$$I = \frac{\pi}{4} (r_2^4 - r_1^4) \quad b = 2(r_2 - r_1) \quad Q = \frac{2\pi}{3} (r_2^3 - r_1^3) \quad (6:10)$$

and the maximum stress is

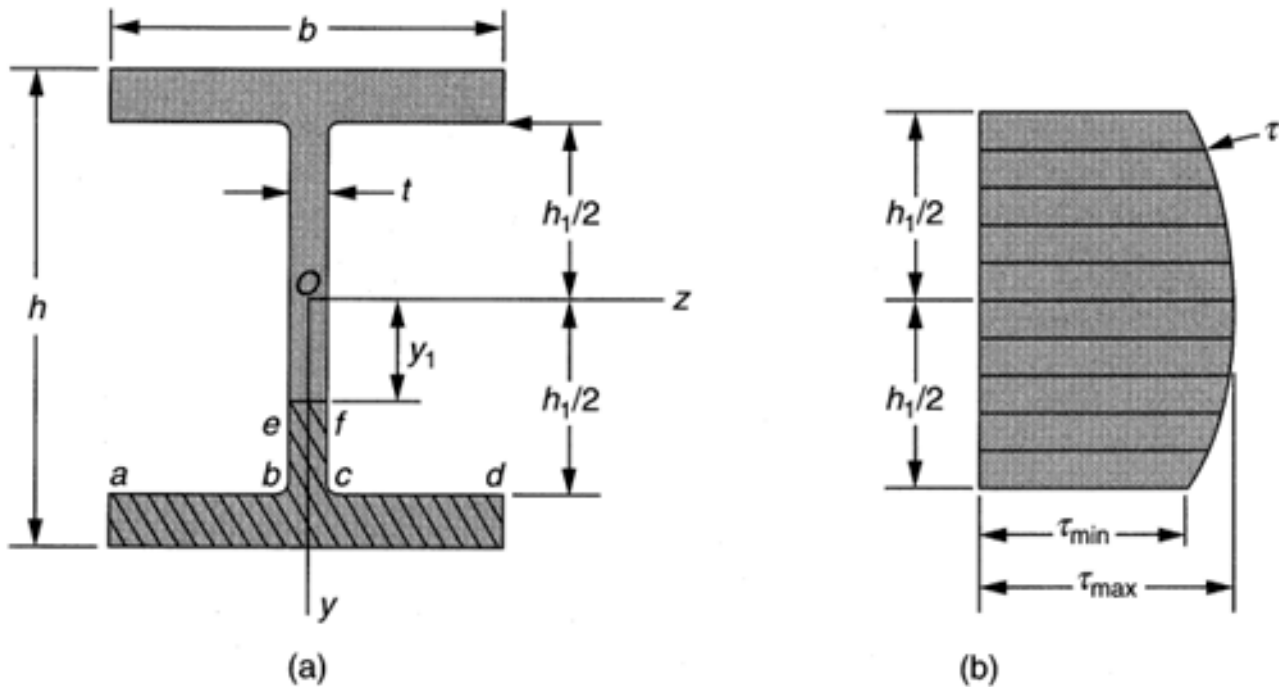
$$\tau_{\max} = \frac{VQ}{Ib} = \frac{4V}{3A} \frac{r_2^2 + r_2 r_1 + r_1^2}{r_2^2 - r_1^2} \quad (6:11)$$

in which  $A = \pi(r_2^2 - r_1^2)$  is the area of the cross section. Note that if  $r_1 = 0$ ; this equation reduces to Eq. (6.9) for a solid circular beam.

### 6.3 Shear Stresses in the Webs of Beams with Flanges

When a beam of wide-flange shape [Fig. 6.6(a)] is subjected to a vertical shear force, the distribution of shear stresses is more complicated than in the case of a rectangular beam. For instance, in the flanges of the beam, shear stresses act in both the vertical and horizontal directions (the  $y$  and  $z$  directions). Fortunately, the largest shear stresses occur in the web, and we can determine those stresses using the same techniques we used for rectangular beams.

**Figure 6.6** Shear stresses in the web of a wide-flange beam. (a) Cross section of beam. (b) Graph showing distribution of vertical shear stresses in the web.



Consider the shear stresses at level  $ef$  in the web of the beam [Fig. 6.6(a)]. We assume that the shear stresses act parallel to the  $y$  axis and are uniformly distributed across the thickness of the web. Then the shear formula will still apply. However, the width  $b$  is now the thickness  $t$  of the web, and the area used in calculating the first moment  $Q$  is the area between  $ef$  and the bottom edge of the cross section [that is, the shaded area of Fig. 6.6(a)]. This area consists of two rectangles—the area of the flange (that is, the area below the line  $abcd$ ) and the area  $efcb$  (note that we disregard the effects of the small fillets at the juncture of the web and flange). After evaluating the first moments of these areas and substituting into the shear formula, we get the following formula for the shear stress in the web of the beam at distance  $y_1$  from the neutral axis:

$$\tau = \frac{VQ}{It} = \frac{V}{8It} [b(h^2 - h_1^2) + t(h_1^2 - 4y_1^2)] \quad (6:12)$$

in which  $I$  is the moment of inertia of the entire cross section,  $t$  is the thickness of the web,  $b$  is the flange width,  $h$  is the height, and  $h_1$  is the distance between the insides of the flanges. The expression for the moment of inertia is

$$I = \frac{bh^3}{12} + \frac{(b - t)h_1^3}{12} = \frac{1}{12}(bh^3 + bh_1^3 + th_1^3) \quad (6:13)$$

Equation (6.12) is plotted in Fig. 6.6(b), and we see that  $\tau$  varies quadratically throughout the height of the web (from  $y_1 = 0$  to  $y_1 = \pm h_1/2$ ):

The maximum shear stress in the beam occurs in the web at the neutral axis ( $y_1 = 0$ ); and the minimum shear stress in the web occurs where the web meets the flanges ( $y_1 = \pm h_1/2$ ): Thus, we find

$$\tau_{\max} = \frac{V}{8It} (bh^2 \mp bh_1^2 + th_1^2) \quad \tau_{\min} = \frac{Vb}{8It} (h^2 \mp h_1^2) \quad (6:14)$$

For wide-flange beams having typical cross-sectional dimensions, the maximum stress is 10 to 60% greater than the minimum stress. Also, the shear stresses in the web typically account for 90 to 98% of the total shear force; the remainder is carried by shear in the flanges.

When designing wide-flange beams, it is common practice to calculate an approximation of the maximum shear stress by dividing the total shear force by the area of the web. The result is an average shear stress in the web:

$$\tau_{\text{ave}} = \frac{V}{th_1} \quad (6:15)$$

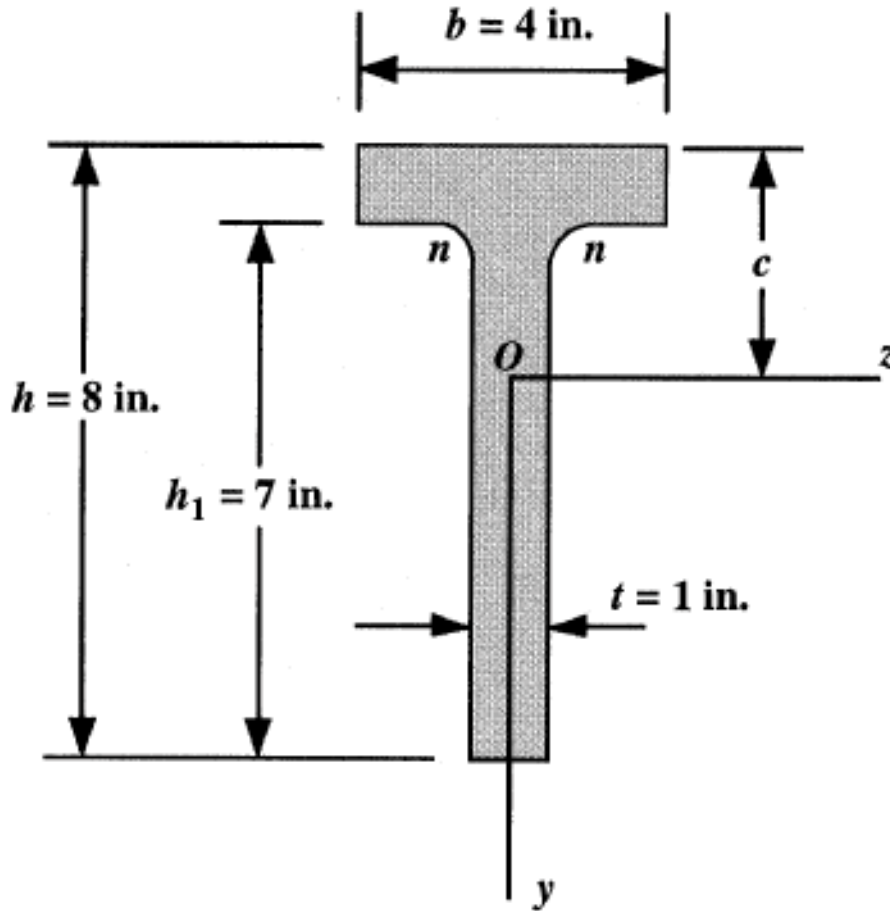
For typical beams, the average stress is within 10% (plus or minus) of the actual maximum shear stress.

The elementary theory presented in the preceding paragraphs is quite satisfactory for determining shear stresses in the web. However, when investigating shear stresses in the flanges, we can no longer assume that the shear stresses are constant across the width of the section, that is, across the width  $b$  of the flanges [Fig. 6.6(a)]. For instance, at the junction of the web and lower flange ( $y_1 = h_1/2$ ); the width of the section changes abruptly from  $t$  to  $b$ . The shear stress at the free surfaces  $ab$  and  $cd$  [Fig. 6.6(a)] must be zero, whereas across the web at  $bc$  the stress is  $\tau_{\min}$ : These observations indicate that at the junction of the web and either flange the distribution of shear stresses is more complex and cannot be investigated by an elementary analysis. The stress analysis is further complicated by the use of fillets at the reentrant corners, such as corners  $b$  and  $c$ . Without fillets, the stresses would become dangerously large. Thus, we conclude that the shear formula cannot be used to determine the vertical shear stresses in the flanges. (Further discussion of shear stresses in thin-walled beams can be found in the references.)

The method used above to find the shear stresses in the webs of wide-flange beams can also be used for certain other sections having thin webs, such as T-beams.

**Example.** A beam having a T-shaped cross section (Fig. 6.7) is subjected to a vertical shear force  $V = 10\,000$  lb. The cross-sectional dimensions are  $b = 4$  in.,  $t = 1$  in.,  $h = 8$  in., and  $h_1 = 7$  in. Determine the shear stress  $\tau_1$  at the top of the web (level  $nn$ ) and the maximum shear stress  $\tau_{\max}$ : (Disregard the areas of the fillets.)

**Figure 6.7** Example.



**Solution.** The neutral axis is located by calculating the distance  $c$  from the top of the beam to the centroid of the cross section. The result is

$$c = 3.045 \text{ in.}$$

The moment of inertia  $I$  of the cross-sectional area about the neutral axis (calculated with the aid of the parallel-axis theorem) is

$$I = 69.66 \text{ in.}^4$$

To find the shear stress at the top of the web we need the first moment  $Q_1$  of the area above level  $nn$ . Thus,  $Q_1$  is equal to the area of the flange times the distance from the neutral axis to the centroid of the flange:

$$Q_1 = A_1 \bar{y}_1 = (4 \text{ in.})(1 \text{ in.})(c + 0.5 \text{ in.}) = 10.18 \text{ in.}^3$$

Substituting into the shear formula, we find

$$\tau_1 = \frac{V Q_1}{I t} = \frac{(10\,000 \text{ lb})(10.18 \text{ in.}^3)}{(69.66 \text{ in.}^4)(1 \text{ in.})} = 1460 \text{ psi}$$

Like all shear stresses in beams, this stress exists both as a vertical shear stress and as a horizontal shear stress. The vertical stress acts on the cross section at level  $nn$  and the horizontal stress acts on the horizontal plane between the flange and the web. The maximum shear stress occurs in the web at the neutral axis. The first moment  $Q_2$  of the area below the neutral axis is

$$Q_2 = A_2 \bar{y}_2 = (1 \text{ in.})(8 \text{ in.} - c) \left( \frac{8 \text{ in.} - c}{2} \right) = 12.28 \text{ in.}^3$$

Substituting into the shear formula, we obtain

$$\tau_{\max} = \frac{V Q_2}{I t} = \frac{(10\,000 \text{ lb})(12.28 \text{ in.}^3)}{(69.66 \text{ in.}^4)(1 \text{ in.})} = 1760 \text{ psi}$$

which is the maximum shear stress in the T-beam.

## Defining Terms

**Shear formula:** The formula  $\tau = VQ/Ib$  giving the shear stresses in a rectangular beam of linearly elastic material [Eq. (6.3)].  
(See also Defining Terms for Chapter 5.)

## References

- Beer, F. P., Johnston, E. R., and DeWolf, J. T. 1992. *Mechanics of Materials*, 2nd ed. McGraw-Hill, New York.
- Gere, J. M. and Timoshenko, S. P. 1990. *Mechanics of Materials*, 3rd ed. PWS, Boston.
- Hibbeler, R. C. 1991. *Mechanics of Materials*. Macmillan, New York.
- Popov, E. P. 1990. *Engineering Mechanics of Solids*. Prentice-Hall, Englewood Cliffs, NJ.
- Riley, W. F. and Zachary, L. 1989. *Introduction to Mechanics of Materials*. John Wiley & Sons, New York.

## Further Information

Extensive discussions of bending—with derivations, examples, and problems—can be found in textbooks on mechanics of materials, such as those listed in the References. These books also cover many additional topics pertaining to shear stresses in beams. For instance, built-up

beams, nonprismatic beams, shear centers, and beams of thin-walled open cross section are discussed in Gere and Timoshenko [1990].

Buchanan, G. R. "Shear and Moment Diagrams"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



# Shear and Moment Diagrams

---

- 7.1 Sign Convention
- 7.2 Shear and Moment Diagrams
- 7.3 Shear and Moment Equations

**George R. Buchanan**

*Tennessee Technological University*

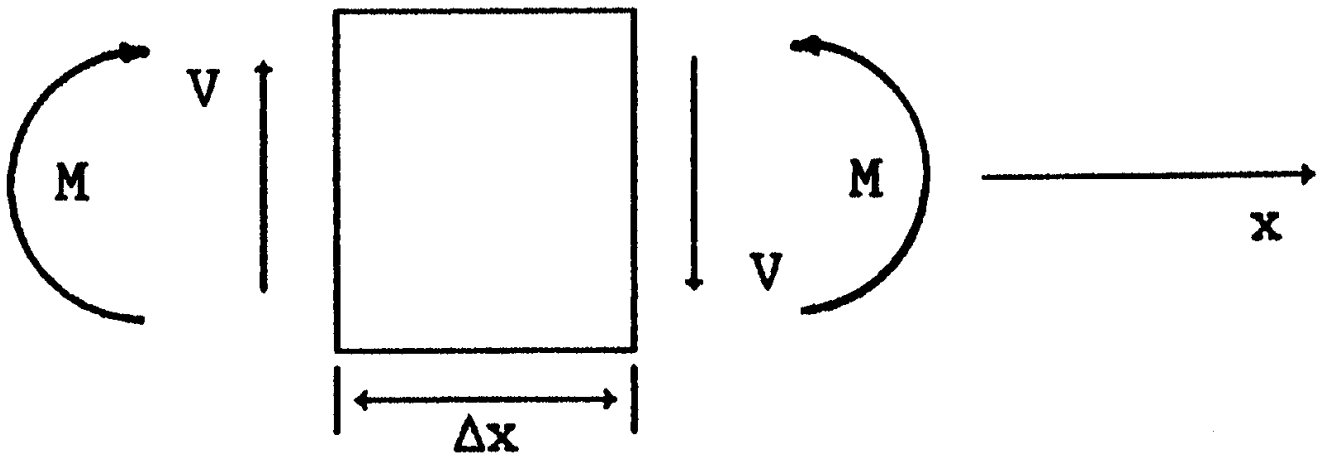
Computations for shear force and bending moment are absolutely necessary for the successful application of the theory and concepts presented in the previous chapters. The discussion presented here is an extension of **Chapter 4** and the reader should already be familiar with computations for reactions. This chapter will concentrate on **statically determinate** beams. Statically indeterminate beams and frames constitute an advanced topic and the reader is referred to the end of the chapter for further information on the topic. Even though the problems that illustrate shear and moment concepts appear as structural beams, the reader should be aware that the same concepts apply to structural machine parts. A distinction should not be drawn between civil engineering and mechanical engineering problems because the methods of analysis are the same.

## 7.1 Sign Convention

---

The sign convention for moment in a beam is based on the behavior of the loaded beam. The sign convention for shear in a beam is dictated by the convenience of constructing a shear diagram using a load diagram. The sign convention is illustrated in [Fig. 7.1](#). The  $x$  axis corresponds to the longitudinal axis of the beam and must be directed from left to right. This convention dictates that shear and moment diagrams should be drawn from left to right. The direction of the positive  $y$  axis will be assumed upward, and loads that act downward on the beam will be negative. Note that it is not mandatory in shear and moment computations for positive  $y$  to be directed upward or even defined since the sign convention is independent of the vertical axis. However, for the more advanced topic of **beam deflections** positive  $y$  must be defined [[Buchanan, 1988](#)]. Positive bending moment causes compression at the top of the beam and negative bending moment causes compression at the bottom of the beam. Positive shear forces act downward on the positive face of the **free body** as shown in [Fig. 7.1](#).

**Figure 7.1** Beam element showing positive shear force  $V$  and positive bending moment  $M$ .



## 7.2 Shear and Moment Diagrams

Two elementary differential equations govern the construction of shear and moment diagrams and can be derived using a free-body diagram similar to [Fig. 7.1](#), with  $w$  corresponding to a continuous load acting along the length of the beam.

$$dV = w \, dx \quad \text{or} \quad \int_{V_1}^{V_2} dV = \int_{x_1}^{x_2} w \, dx \quad (7.1)$$

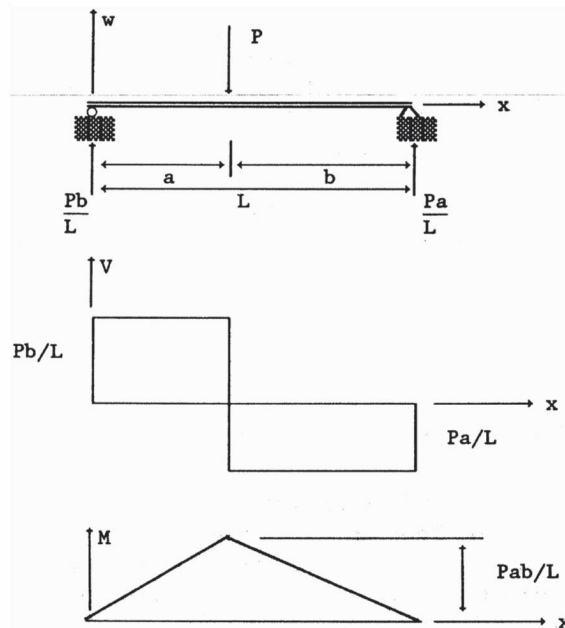
$$dM = V \, dx \quad \text{or} \quad \int_{M_1}^{M_2} dM = \int_{x_1}^{x_2} V \, dx \quad (7.2)$$

The differential equations show that the change in shear  $V$  between any two points  $x_1$  and  $x_2$  on a beam is equal to the area of the load diagram between those same two points and, similarly, that the change in bending moment  $M$  is equal to the area of the shear diagram. It follows that the slope of a tangent drawn at any point on the moment diagram is given by  $dM/dx$  and corresponds to the magnitude of  $V$  at that point. When the tangent has zero slope,  $dM/dx = 0$ , that corresponds to a **maximum or minimum moment** and can be located by examining the shear diagram for a point ( $x$  location) where  $V = 0$ . Locating the largest positive or negative bending moment is important for properly designing beam structures when using the equations of the previous chapters.

Shear and moment diagrams (as opposed to shear and moment equations) offer the most efficient method for analyzing beam structures for shear and moment when the beam loading can be represented as **concentrated loads** or **uniform continuous loads**. An elementary example will serve to illustrate the concept. Consider the **simply supported beam** of [Fig. 7.2](#). There is a single concentrated load with reactions as shown. The shear is obtained by directly plotting the load; the sign convention of [Fig 7.1](#) specifically allows for this. The reaction on the left is plotted upward as the change in the shear at a point,  $x = 0$ . The area of the load diagram between  $x = 0$  and  $x = a$  is zero since the load is zero; it follows that the change in shear is zero and the shear diagram is a

straight horizontal line extending from the left end of the beam to the concentrated load  $P$ . The load changes abruptly by an amount  $P$  downward and a corresponding change is noted on the shear diagram. Positive shear is above the axis of the shear diagram. There is no change in load between  $x = a$  and  $x = L$  and the shear remains constant. The reaction at the right end of the beam is upward and the shear is plotted upward to return to zero. The beam is simply supported, indicating that the moment must be zero at the supports. The change in moment between the left support and the point where the load is applied,  $x = a$ , is equal to the area of the shear diagram or a positive  $Pab/L$ . The variation in moment appears as a straight line (a line with constant slope) connecting  $M = 0$  at  $x = 0$  with  $M = Pab/L$  at  $x = a$ . The area of the remaining portion of the shear diagram is  $-Pab/L$  and, when plotted on the moment diagram, returns the moment to zero and satisfies the simply supported boundary condition.

**Figure 7.2** Shear and moment diagrams for a simple beam with a concentrated load.

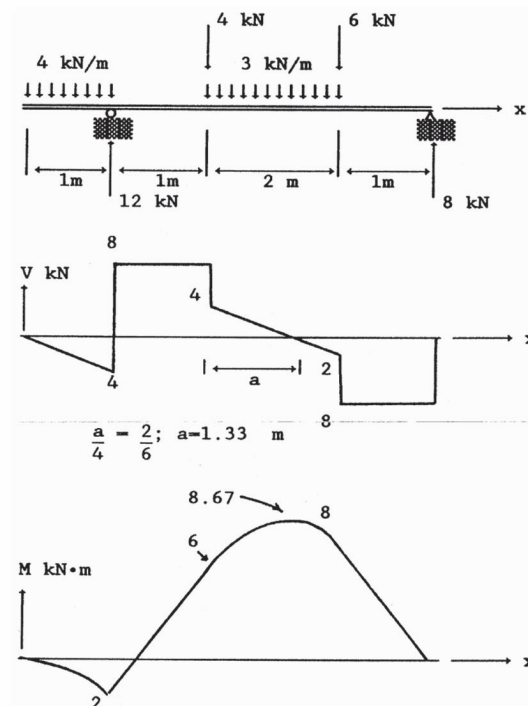


Equations (7.1) and (7.2) indicate that the load function is integrated to give the shear function, and similarly the shear function is integrated to give the moment function. The diagrams are a graphical illustration of the integration. An important point is that the order (power) of each

function increases by one as the analyst moves from load to shear to moment. In Fig. 7.2, note that when the load function is zero (1) the corresponding shear function is a constant and (2) the corresponding moment function is linear in  $x$  and is plus or minus corresponding to the sign of the area of the shear diagram.

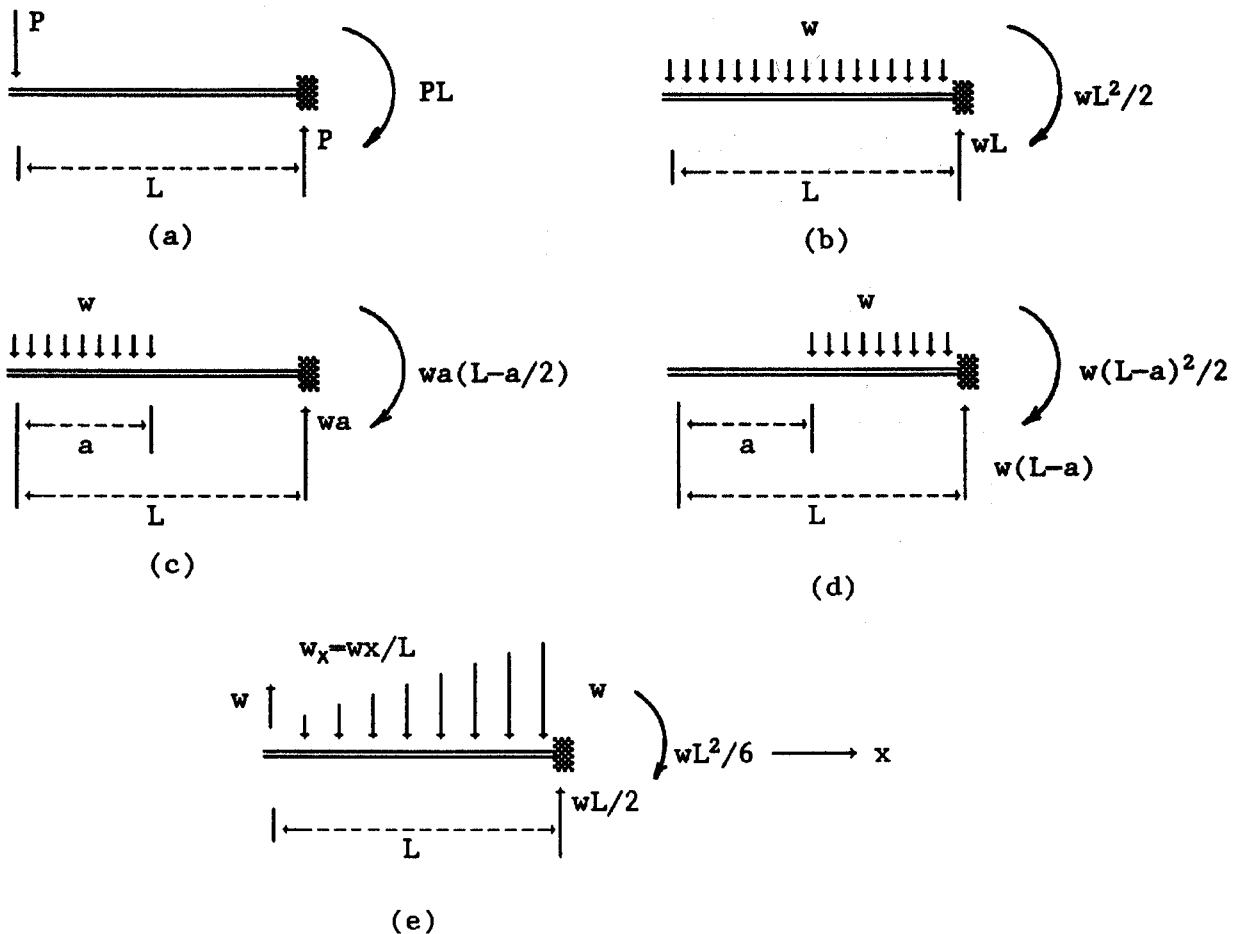
Consider, as a second example, the beam of Fig. 7.3, where a series of uniform loads and concentrated loads is applied to a beam with an overhang. The reactions are computed and shown in the figure. The shear diagram is plotted starting at the left end of the beam. Shear and moment diagrams are always constructed from left to right because Eqs. (7.1) and (7.2) were derived in a coordinate system that is positive from left to right. The area of the load between  $x = 0$  and  $x = 1$  m is  $-4$  kN; this value is the change in shear, which is plotted as a sloping line. The corresponding change in bending moment equals the area of the shear diagram,  $(-4 \text{ kN})(1 \text{ m})/2 = 2 \text{ kN} \cdot \text{m}$ . A curve with continually changing negative slopes between  $x = 0$  and  $x = 1$  m is shown in Fig. 7.3. The point of zero shear occurs in the beam section,  $2 \text{ m} \leq x \leq 4 \text{ m}$ , and is located using similar triangles as shown in the space between the shear and moment diagrams of Fig. 7.3. A textbook on mechanics of materials or structural analysis should have a complete discussion of the topic. Again, refer to "Further Information."

**Figure 7.3** Shear and moment diagrams for a beam with an overhang. The concentrated loads and uniform loads illustrate the concept of maximum moment and corresponding zero shear.



Concentrated loads and uniform loads lead to shear diagrams with areas that will always be rectangles or triangles; the change in moment is easily computed. The uniformly varying load shown in Fig. 7.4(e) sometimes occurs in practice; locating the point of maximum moment (point of zero shear) for some boundary conditions is not so elementary when using geometrical relationships. In such cases the use of shear and moment equations becomes a valuable analysis tool.

**Figure 7.4** Shear and moment reactions for free-fixed beams.



## 7.3 Shear and Moment Equations

Shear and moment equations are equations that represent the functions shown in Figs. 7.2 and 7.3. As with any mathematical function, they must be referenced to a coordinate origin, usually the left end of the beam. Shear and moment equations are piecewise continuous functions. Note that two separate equations are required to describe the shear diagram of Fig. 7.2 and, similarly, four equations are required to describe the shear diagram of Fig. 7.3. The same is true for the moment

diagram.

The following procedure can be used to write shear and moment equations for almost any beam loading: (1) Choose a coordinate origin for the equation, usually, but not limited to, the left end of the beam. (2) Pass a free-body cut through the beam section where the shear and moment equations are to be written. (3) Choose the free body that contains the coordinate origin. (4) Assume positive unknown shear and moment at the free-body cut using the sign convention defined by Fig. 7.1. (5) View the free body as a **free-fixed beam** with the fixed end being at the free-body cut and the beam extending toward the coordinate origin. (6) Statically solve for the unknown shear and moment as if they were the reactions at the fixed end of any beam, that is,  $\Sigma F = 0$  and  $\Sigma M = 0$ . (7) Always sum moments at the free-body cut such that the unknown shear passes through that point. An example should illustrate the concept.

A complete description of the beam of Fig. 7.3 would require four shear and moment equations. Consider a free body of the first section—the uniformly loaded overhang. The free body is shown in Fig. 7.5(a). Compare Fig. 7.4(b) with Fig. 7.5(a);  $L$  is merely replaced with  $x$ , the length of the free-body section. The seven steps outlined above have been followed to give

$$V_x = -wx = -4x \text{ kN}, \quad M_x = -wx^2/2 = -4x^2/2 \text{ kN} \cdot \text{m}, \quad 0 \leq x \leq 1 \text{ m} \quad (7.3)$$

The second segment ( $1 \text{ m} \leq x \leq 2 \text{ m}$ ) is shown in Fig. 7.5(b) and can be compared with Figs. 7.4(c) and 7.4(a): merely replace  $L$  with an  $x$ ,  $a$  with 1 m,  $P$  with 12 kN, and  $w$  with 4 kN.

$$V_x = -wa + P = -(4 \text{ kN/m})(1 \text{ m}) + 12 \text{ kN}, \quad (7.4)$$

$$\begin{aligned} M_x &= -wa(x - a/2) + P(x - a) \\ &= -(4)(1)(x - 1/2) \text{ kN} \cdot \text{m} + 12(x - 1) \text{ kN} \cdot \text{m} \end{aligned} \quad (7.5)$$

The general idea is that any shear or moment equation can be broken down into a series of individual problems that always correspond to computing the reactions at the fixed end of a free-fixed beam. Continuing with the shear and moment equations for the beam of Fig. 7.3, the third section ( $2 \text{ m} \leq x \leq 4 \text{ m}$ ) is shown in Fig. 7.5(c) and should be compared with Figs. 7.4(a), (c), and (d). There are four terms in each equation since there are four separate loadings on the free body.

$$\begin{aligned} V_x &= -(4 \text{ kN/m})(1 \text{ m}) + 12 \text{ kN} \\ &\quad - 4 \text{ kN} - (3 \text{ kN/m})(x - 2 \text{ m}), \end{aligned} \quad (7.6)$$

$$\begin{aligned} M_x &= -(4 \text{ kN/m})(1 \text{ m})(x - 1 \text{ m}/2) + (12 \text{ kN})(x - 1 \text{ m}) - (4 \text{ kN})(x - 2 \text{ m}) \\ &\quad - (3 \text{ kN/m})(x - 2 \text{ m})^2/2 \end{aligned} \quad (7.7)$$

The last beam section ( $4 \text{ m} \leq x \leq 5 \text{ m}$ ) is shown in Fig. 7.5(d). There are five separate loadings

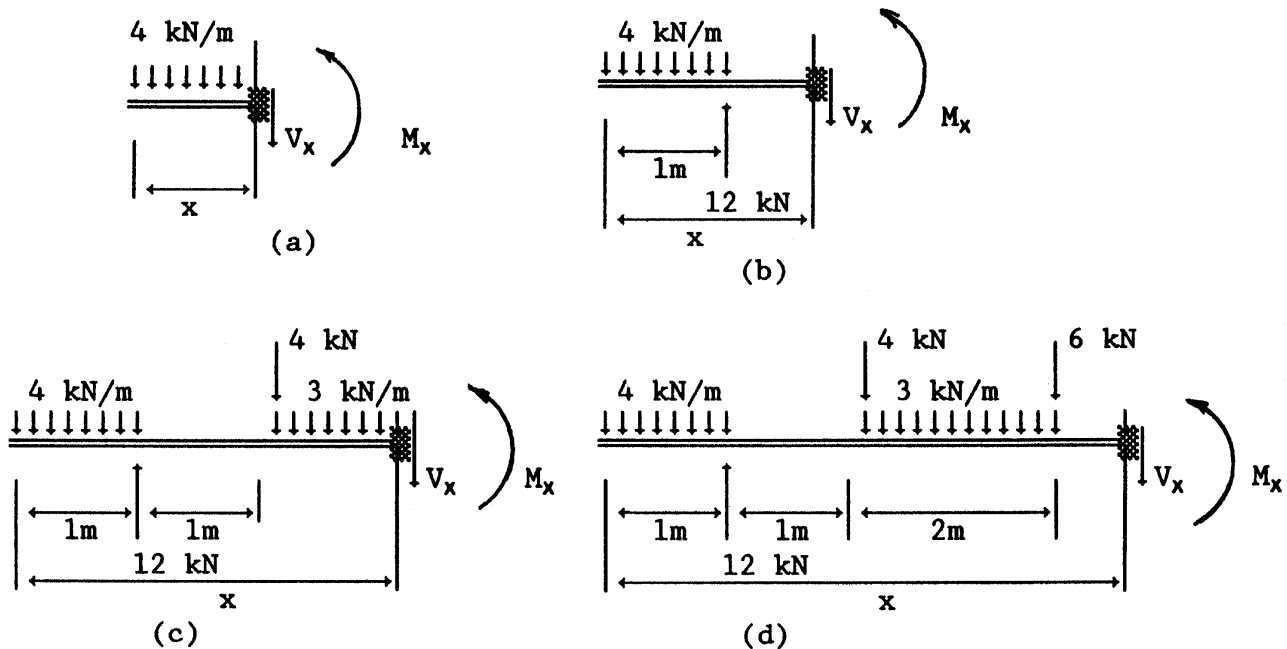
on the beam. The shear and moment equations are written again by merely summing forces and moments on the free-body section.

$$V_x = -(4 \text{ kN/m})(1 \text{ m}) + 12 \text{ kN} - 4 \text{ kN} - (3 \text{ kN/m})(2 \text{ m}) - 6 \text{ kN}, \quad (7.8)$$

$$M_x = -(4 \text{ kN/m})(1 \text{ m})(x - 1 \text{ m}/2) + (12 \text{ kN})(x - 1 \text{ m}) - (4 \text{ kN})(x - 2 \text{ m}) - (3 \text{ kN/m})(2 \text{ m})(x - 3 \text{ m}) - (6 \text{ kN})(x - 4 \text{ m}) \quad (7.9)$$

The point of maximum moment corresponds to the point of zero shear in the third beam section. The shear equation, Eq. (7.6), becomes  $V_x = 10 - 3x$ . Setting  $V_x$  to zero and solving for  $x$  gives  $x = 3.33 \text{ m}$ . Substituting into the corresponding moment equation, Eq. (7.7), gives the maximum moment as  $8.67 \text{ kN} \cdot \text{m}$ .

**Figure 7.5** Free-body diagrams for the beam shown in Fig. 7.3.



## Defining Terms

**Beam deflections:** A theory that is primarily based upon the moment behavior for a beam and leads to a second-order differential equation that can be solved to give a mathematical equation describing the deflection of the beam.

**Concentrated load:** A single load, with units of force, that can be assumed to act at a point on a beam.

**Free body:** A section that is removed from a primary structural system and is assumed to be in equilibrium mathematically.

**Free-fixed beam:** A beam that is free to rotate and deflect at one end but is completely clamped or rigid at the other end (also known as a *cantilever beam*).

**Maximum or minimum moment:** The bending moment that usually governs the design and analysis of beam structures.

**Simply supported beam:** A beam that is supported using a pin (hinge) at one end and a surface at the other end, with freedom to move along the surface.

**Statically determinate beam:** A beam that can be analyzed for external reactions using only the equations of engineering mechanics and statics.

**Uniform continuous load:** A distributed beam loading of constant magnitude, with units of force per length, that acts continuously along a beam segment.

## References

Buchanan, G. R. 1988. Shear and moment in beams, Chap. 5, and Deflection of beams, Chap. 10, in *Mechanics of Materials*. Holt, Rinehart and Winston, New York.

## Further Information

Hibbeler, R. C. 1985. *Structural Analysis*. Macmillan, New York. Chapter 3 contains a discussion of shear and moment concepts. Chapters 8 and 9 cover fundamental concepts for analysis of indeterminate beams.

McCormac, J. and Elling, R. E. 1988. *Structural Analysis*. Harper & Row, New York. Chapter 3 contains a discussion of shear and moment concepts. Chapters 10, 11, and 13 cover fundamental concepts for analysis of indeterminate structures.

Gere, J. M. and Timoshenko, S. P. 1990. *Mechanics of Materials*, 3rd ed. PWS, Boston. Chapter 4 contains a discussion of shear and moment concepts. Beam deflections are covered in Chaps. 7, 8, and 10.

Nash, W. A. 1994. *Theory and Problems of Strength of Materials*, 3rd ed., McGraw-Hill, New York. Numerous solved problems for shear and moment are given in Chap. 6



Zachary, L. W., Ligon, J. B. "Columns"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 8

## Columns

---

### 8.1 Fundamentals

Buckling of Long Straight Columns • Effective Lengths • Compression Blocks and Intermediate Lengths

### 8.2 Examples

### 8.3 Other Forms of Instability

**Loren W. Zachary**

*Iowa State University*

**John B. Ligon**

*Michigan Technological University*

A column is an initially straight load-carrying member that is subjected to a compressive axial load. The failure of a column in compression is different from one loaded in tension. Under compression, a column can deform laterally, or buckle, and this deflection can become excessive. The buckling of columns is a major cause of failure. To illustrate the fundamental aspects of the buckling of long, straight, prismatic bars, consider a thin meter stick. If a tensile axial load is applied to the meter stick, the stable equilibrium position is that of a straight line. If the stick is given a momentary side load to cause a lateral deflection, upon its release the stick immediately returns to the straight line configuration. If a compressive axial load is applied, a different result may occur. At small axial loads the meter stick will again return to a straight line configuration after being displaced laterally. At larger loads the meter stick will remain in the displaced position. With an attempt to increase the axial load acting on the buckled column, the lateral deformations become excessive and failure occurs.

In theory, the column that is long and perfectly straight is in stable equilibrium for small loads up to a specific **critical buckling load**. At this critical buckling load the beam will remain straight unless it is perturbed and displays large lateral deformations. This is a bifurcation point since the column can be in equilibrium with two different shapes—laterally displaced or perfectly straight. The load in this neutral equilibrium state is the critical buckling load and, for long slender columns, is referred to as the **Euler buckling load**. At loads higher than the critical load the beam is in unstable equilibrium.

## 8.1 Fundamentals

---

### Buckling of Long Straight Columns

In 1757 Leonhard Euler published the solution to the problem of long slender columns buckling under compressive loads. [Figure 8.1\(a\)](#) shows a column that is deflected in the lateral direction. The load  $P_{cr}$  is the smallest load that will just hold the column in the laterally deflected shape. The ends of the beam are free to rotate and are commonly referred to as being *pinned*, *hinged*, or simply *supported*. The following assumptions are used in determining  $P_{cr}$ :

1. The beam is initially straight with a constant cross section along its length.
2. The material is linearly elastic, isotropic, and homogeneous.
3. The load is applied axially through the centroidal axis.
4. The ends are pinned–pinned (pinned end condition at both ends).
5. No residual stresses exist in the column prior to loading.
6. No distortion or twisting of the cross section occurs during loading.
7. The classical differential equation for the elastic curve can be used since the deflections are small.

Standard mechanics of materials textbooks [[Riley and Zachary, 1989](#)] and structural stability textbooks [[Chajes, 1974](#)] contain the derivation of the following formula:

$$\sqrt{\frac{P_{cr}}{EI}} L = n\pi; \quad n = 1; 2; 3; \dots \quad (8:1)$$

The smallest value for  $P_{cr}$  occurs when  $n = 1$ . Larger values of  $n$  give magnitudes of  $P_{cr}$  that will never be reached in practice.

$$P_{cr} = \frac{\pi^2 EI}{L^2} \quad (8:2)$$

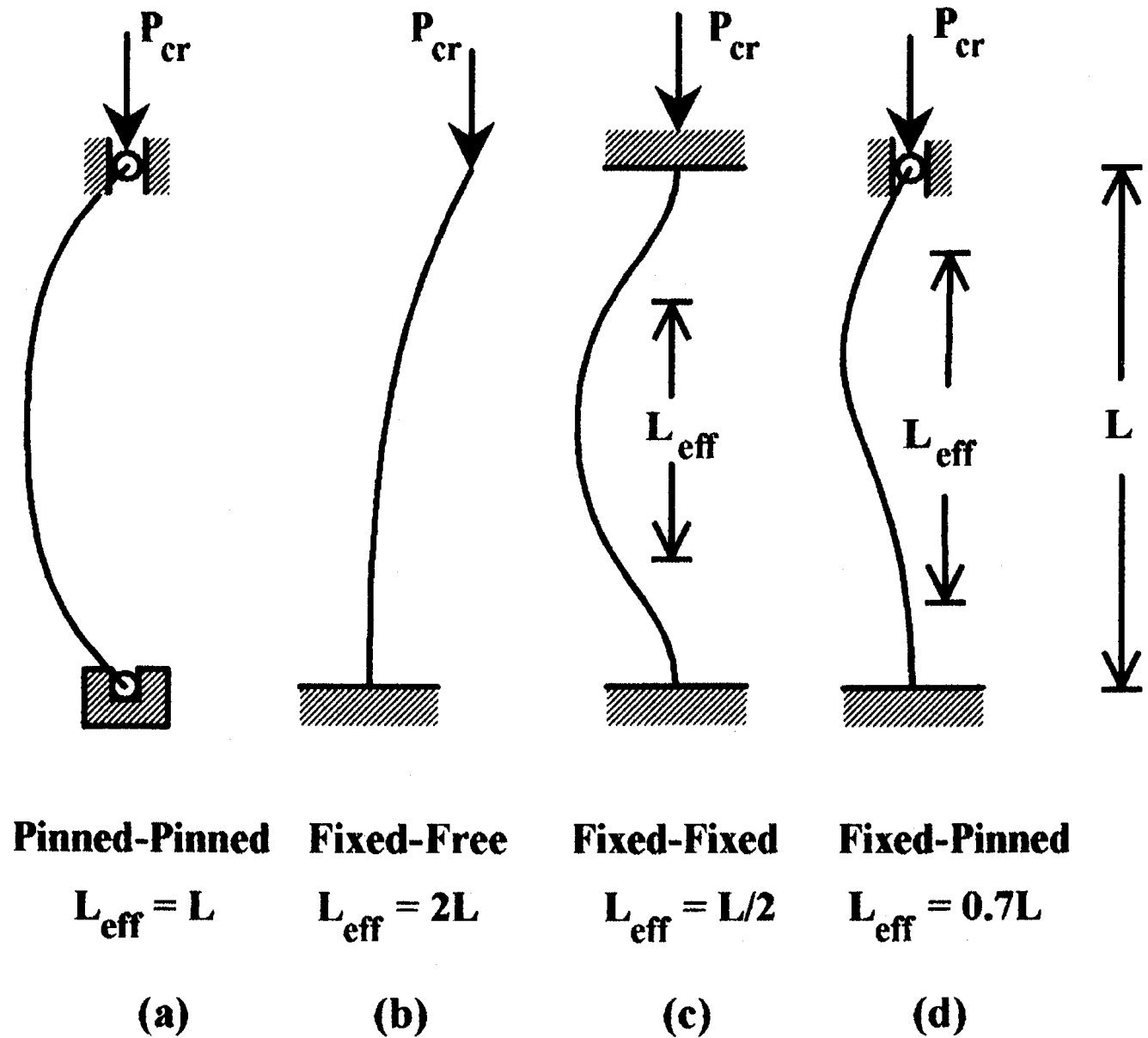
The Euler buckling load,  $P_{cr}$ , is calculated using the moment of inertia  $I$  of the column cross section about which axis buckling (bending) occurs. The moment of inertia can also be written in terms of the radius of gyration about the same axis:

$$I = Ar^2 \quad (8:3)$$

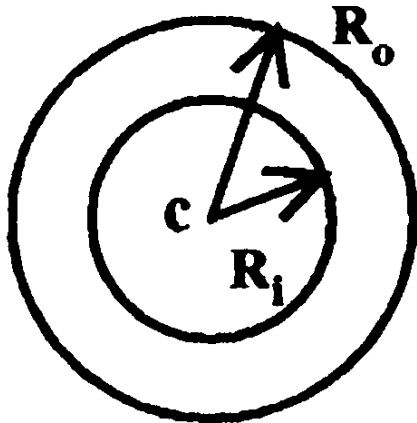
[Table 8.1](#) gives some formulas that are helpful in determining the radius of gyration and moment of inertia. Using Eqs. (8.2) and (8.3), the Euler buckling stress,  $\sigma_{cr}$ , in terms of the **slenderness ratio**  $L/r$ , is

$$\frac{3}{4}P_{cr} = \frac{P_{cr}}{A} = \frac{\frac{1}{4}E}{(L=r)^2} \quad (8:4)$$

**Fig. 8.1** Effective column lengths.



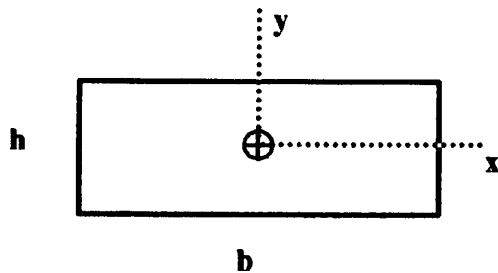
**Table 8.1** Properties of Selected Areas



$$A = \frac{1}{4} \pi (R_o^2 - R_i^2)$$

$$I_c = \frac{1}{4} \pi (R_o^4 - R_i^4)$$

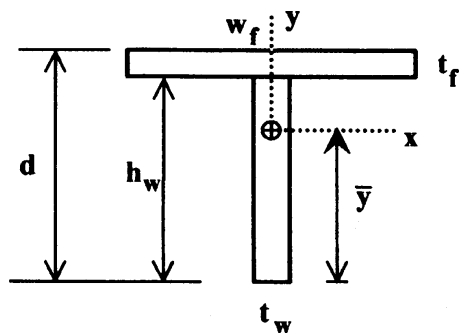
$$r_c = \frac{1}{2} \sqrt{\frac{R_o^4 - R_i^4}{R_o^2 - R_i^2}}$$



$$A = bh$$

$$I_x = \frac{1}{12}bh^3 \quad I_y = \frac{1}{12}hb^3$$

$$r_x = \frac{h}{\sqrt{12}} \quad r_y = \frac{b}{\sqrt{12}}$$



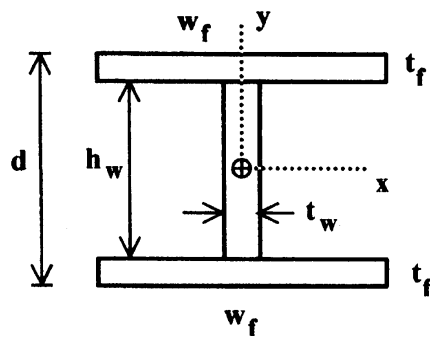
$$A = w_f t_f + h_w t_w$$

$$\bar{y} = \frac{w_f t_f \left( \frac{h_w}{2} + \frac{t_f}{2} \right) + \frac{h_w^2 t_w}{2}}{A}$$

$$I_x = \frac{1}{3} w_f (d - \bar{y})^3 + \frac{1}{3} t_w \bar{y}^3 + \frac{1}{3} (w_f - t_w) (h_w - \bar{y})^3$$

$$I_y = \frac{1}{12} t_f w_f^3 + \frac{1}{12} h_w t_w^3$$

$$r_x = \frac{\rho}{I_x=A} \quad r_y = \frac{\rho}{I_y=A}$$



$$A = 2w_f t_f + h_w t_w$$

$$\bar{y} = \frac{d}{2}$$

$$I_x = \frac{1}{12} w_f d^3 + \frac{1}{12} (w_f - t_w) h_w^3$$

$$I_y = \frac{1}{6} t_f w_f^3 + \frac{1}{12} h_w t_w^3$$

$$r_x = \frac{\rho}{I_x=A} \quad r_y = \frac{\rho}{I_y=A}$$

## Effective Lengths

The development given above is for a beam with simple supports at both ends. Other boundary conditions give equations similar to Eq. (8.4) if the physical length of the beam,  $L$ , is replaced by the effective length  $L_{e^{\circ}}$ .

$$\frac{3}{4}P_{cr} = \frac{\frac{1}{4}E}{(L_{e^{\circ}}/r)^2} \quad (8:5)$$

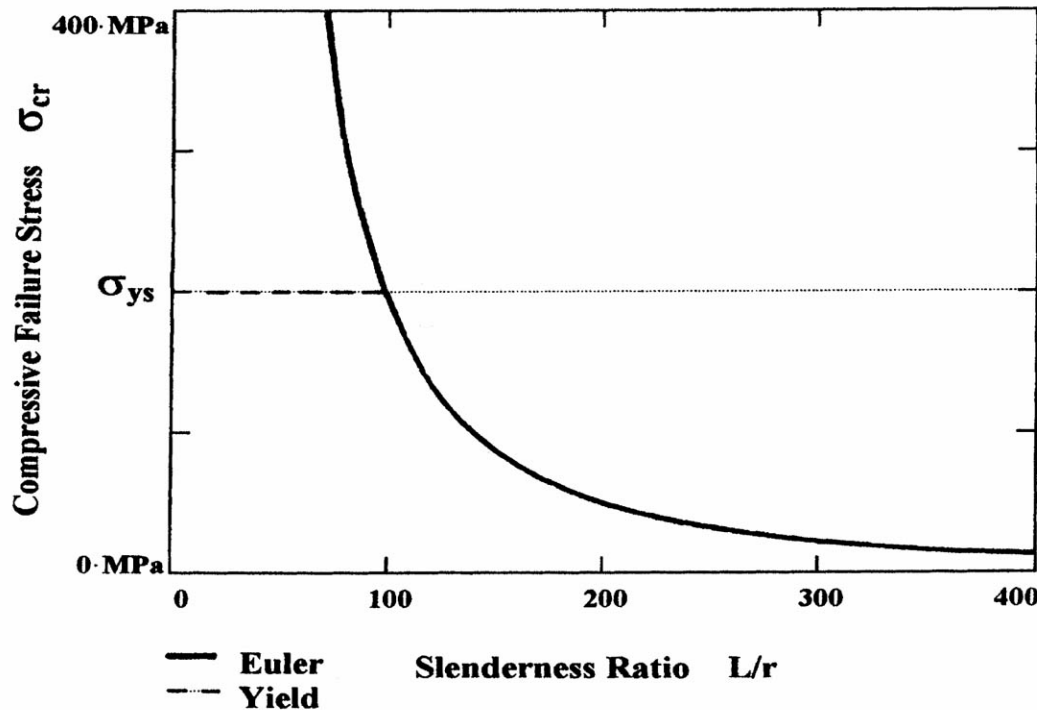
Figure 8.1 gives the effective lengths for four classic end condition cases. The effective lengths are all measured between the inflection points of the elastic curves. The moments at the inflection points,  $y'' = 0$ , are zero. This observation can be used to estimate the effective length of other boundary condition cases. If one can estimate the elastic curve and visualize the location of the inflection points, a rough estimate of the effective length can be obtained.

## Compression Blocks and Intermediate Lengths

A plot of Eq. (8.5), using a generic steel with a Young's modulus,  $E$ , of 200 GPa for illustration purposes, is shown as the solid curved line in Fig. 8.2. For columns that are very short and stocky where  $L/r$  approaches zero, Eq. (8.5) predicts that the column will support a very large load or normal stress. However, the mechanism of failure changes when the column becomes a short compression block. The compressive yield strength limits the compressive normal load that can be carried by the column. The horizontal dotted line in Fig. 8.2 represents the yield stress limit that the column can sustain due to compressive block failure.

Critical buckling loads for large values of  $L/r$  are predicted with a high degree of confidence using Euler's column equation. Failure loads for small values of  $L/r$  are reliably predicted from compressive yield strength criteria for the compression block. Columns that have effective lengths in the region near the yield strength magnitude on the Euler curve may behave neither as an Euler column or a compressive block. Experiments using the particular material in question establish the shape of the curve between the compression block values and the Euler column values. Standard mechanics of materials textbooks [Riley and Zachary, 1989] give empirical formulas for this range. Using a horizontal line to cover the complete compression block and intermediate ranges usually predicts a higher critical buckling load than is found experimentally.

**Fig. 8.2** Effect of slenderness ratio on compressive failure stress.



## 8.2 Examples

The following examples illustrate the procedure for determining the critical buckling load and stress for columns of several different cross sections and effective lengths. Initially, the slenderness ratio for the two principal directions of possible buckling must be calculated to determine which direction controls. The direction with the largest slenderness ratio will give the smallest buckling load.

**Example 8.1: Checking Both Directions for Buckling.** Consider the rectangular cross section shown in Fig. 8.3. Both ends of the column are pinned for the pinned–pinned condition in Fig. 8.1(a). The physical length of the column is 1.5 m. Determine the Euler buckling load if the column is made of steel ( $E = 200$  GPa). Using Table 8.1,

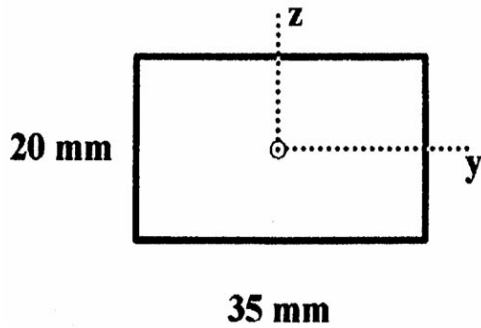
$$A = (20 \text{ mm})(35 \text{ mm}) = 700 \text{ mm}^2$$

$$r_y = \sqrt{\frac{I_y}{A}} = \sqrt{\frac{\frac{bh^3}{12}}{A}} = 5.77 \text{ mm}$$

$$r_z = \sqrt{\frac{I_z}{A}} = \sqrt{\frac{\frac{hb^3}{12}}{A}} = 10.10 \text{ mm}$$

The column tends to bend (buckle) about the  $y$  axis since  $r_y$  is smaller than  $r_z$ , producing the maximum  $L/r$  and the smallest  $P_{cr}$  and  $\sigma_{cr}$  for the 1.5-m length.

**Fig. 8.3** Cross section used in Example 8.1.



If the ends of the column are later restrained or fixed with respect to buckling about one of the axes—say the  $y$  axis—this changes the boundary conditions for buckling to the fixed–fixed condition [Fig. 8.1(c)] for that direction. The effective length of the beam is then half the physical length or 750 mm for buckling about the  $y$  axis.

$$\frac{L_y}{r_y} = \frac{L=2}{r_y} = \frac{750 \text{ mm}}{5.77 \text{ mm}} = 130$$

$$\frac{L_z}{r_z} = \frac{L}{r_z} = \frac{1500 \text{ mm}}{10.10 \text{ mm}} = 149$$

The column will now buckle about the  $z$  axis before the load can become large enough to cause buckling about the  $y$  axis. Although the moment of inertia and radius of gyration about the  $y$  axis are smaller than about the  $z$  axis, the end conditions significantly influence the slenderness ratio and the axis about which buckling occurs.

$$P_{cr} = \frac{\frac{1}{4} E A}{(L_z=r_z)^2} = \frac{\frac{1}{4} 200(10)^9 \text{ N} \cdot \text{m}^2 \cdot 700(10)^{-6} \text{ m}^2}{(149)^2} = 62.2 \text{ kN} \quad \text{Ans:}$$

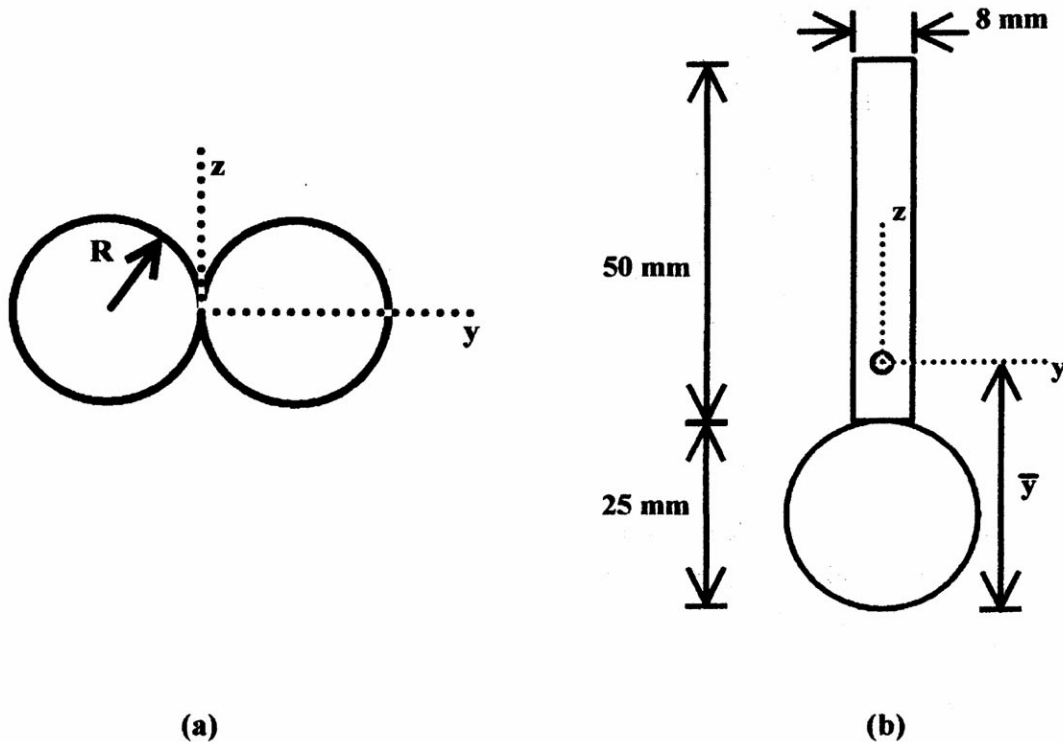
$$\frac{3}{4} \sigma_{cr} = \frac{P_{cr}}{A} = \frac{62.2(10)^3 \text{ N}}{700(10)^{-6} \text{ m}^2} = 88.9 \text{ MPa} \quad \text{Ans:}$$

According to the Euler buckling formula, any load below 62.2 kN will not cause the column to buckle. No factor of safety is included in the calculations. In practice, however, such a factor should be used since real columns are never ideally straight nor is column loading purely axial. Some inadvertent bending moment or load eccentricity is always possible. The  $\frac{3}{4} \sigma_{cr}$  calculated above is the axial compressive stress in the beam just before the beam deforms laterally. This stress is relatively small when compared to a yield strength for structural steel of approximately 250 MPa. The compressive stress in long slender beams at or below the critical buckling stress  $\frac{3}{4} \sigma_{cr}$  can be much less than the yield strength of the material. It is imperative that failure due to buckling be checked for compressive loads.



**Example 8.2a: Built-up Section.** A single 20-mm solid steel rod is being used in compression. It has been determined that the rod will buckle. It has been decided that two rods will be welded together—Fig. 8.4(a)—in order to increase the buckling load. Does this increase the Euler buckling load?

**Fig. 8.4** Cross sections used in Examples 8.2a and 8.2b.



The parallel axis theorem can be used to determine the combined moments of inertia and radii of gyration. The term  $d$  in the following equation is the transfer distance from the centroid of the component area to the centroid of the entire cross section.

$$I = \sum (I_c + Ad^2)$$

$$r_z = \sqrt{\frac{I_z}{A}} = \sqrt{\frac{2Ar_{cz}^2 + 2Ad_z^2}{2A}} = \sqrt{r_{cz}^2 + d_z^2} = \sqrt{(R=2)^2 + R^2} = \sqrt{5} \frac{R}{2}$$

$$r_y = \sqrt{\frac{I_y}{A}} = \sqrt{\frac{2Ar_{cy}^2 + 2Ad_y^2}{2A}} = \sqrt{r_{cy}^2 + d_y^2} = \sqrt{(R=2)^2 + 0} = \frac{R}{2}$$

The radius of gyration is not increased for bending about the  $y$  axis compared to the single rod value of  $R=2$ . The buckling load will remain the same even though the column has been stiffened in one direction.

**Example 8.2b: Built-up Section.** Consider the composite aluminum ( $E = 70 \text{ GPa}$ ) section in Fig. 8.4(b). Determine the maximum compressive stress that can be applied. The end conditions are as follows: about the  $y$  axis, the ends are fixed–pinned (see Fig. 8.1(d)), and about the  $z$  axis, the ends are fixed–fixed (see Fig. 8.1(d)). The column length is 2 m.

Compared to Example 2a, the radius of gyration about the  $y$  axis has the same basic definition, but the details are slightly different.

$$A_{\text{rect}} = (8 \text{ mm})(50 \text{ mm}) = 400 \text{ mm}^2$$

$$A_{\text{rod}} = \frac{1}{4}(12.5 \text{ mm})^2 = 490.9 \text{ mm}^2$$

$$\bar{y} = \frac{\sum P_i \bar{y}_i}{\sum P_i} = \frac{(400 \text{ mm}^2)50 \text{ mm} + (490.9 \text{ mm}^2)12.5 \text{ mm}}{400 \text{ mm}^2 + 490.9 \text{ mm}^2} = 29.34 \text{ mm}$$

$$r_y = \sqrt{\frac{\sum P_i I_{y_i}}{\sum P_i}} = \sqrt{\frac{A_{\text{rect}}(r_{y_{\text{rect}}}^2 + d_y^2) + A_{\text{rod}}(r_{y_{\text{rod}}}^2 + d_y^2)}{\sum P_i}}$$

$$= \sqrt{\frac{400 \left( \frac{50}{12} \right)^2 + (50 + 29.34)^2 + 490.9 \left( \frac{12.5}{2} \right)^2 + (29.34 + 12.5)^2}{890.9}}$$

$$= 21.52 \text{ mm}$$

$$r_z = \sqrt{\frac{\sum P_i I_{z_i}}{\sum P_i}} = \sqrt{\frac{A_{\text{rect}}(r_{z_{\text{rect}}}^2 + d_z^2) + A_{\text{rod}}(r_{z_{\text{rod}}}^2 + d_z^2)}{\sum P_i}}$$

$$= \sqrt{\frac{400 \left( \frac{8}{12} \right)^2 + 490.9 \left( \frac{12.5}{2} \right)^2}{890.9}} = 4.891 \text{ mm}$$

The slenderness ratio for buckling about the  $z$  axis controls since  $r_z$  is less than one-fourth of the value about the  $y$  axis:

$$\frac{L_y}{r_y} = \frac{0.7L}{r_y} = \frac{0.7(2000 \text{ mm})}{21.52 \text{ mm}} = 65.1$$

For many materials a slenderness ratio of 65.1 places the beam in the intermediate length range.

$$\frac{L_z}{r_z} = \frac{L=2}{r_z} = \frac{1000 \text{ mm}}{4.891 \text{ mm}} = 204.5$$

The beam acts as an Euler beam for buckling about the  $z$  axis:

$$\sigma_{cr} = \frac{\frac{1}{4} E}{(L_z=r_z)^2} = \frac{\frac{1}{4} 70(10^9) \text{ N/m}^2}{(204.5)^2} = 16.52 \text{ MPa} \quad \text{Ans:}$$

## 8.3 Other Forms of Instability

---

Beams can also fail due to local instabilities. There can be a crushing type of failure that is familiar in the crushing of thin-walled soda pop cans. The above formulas do not apply in such instances [Young, 1989]. Hollow rods subjected to torsion can buckle locally due to the compressive principal stress acting at a  $45^\circ$  angle to the longitudinal axis of the rod. Beams can fail in a combined bending and torsion fashion. An I-beam, with a pure moment applied, can have the flanges on the compression side buckle. When lateral loads are present in conjunction with an axial compressive load, the beam acts as a beam-column [Chen and Atsuta, 1976].

### Defining Terms

**Critical buckling load:** The smallest compressive load at which a column will remain in the laterally displaced, buckled, shape.

**Critical buckling stress:** The smallest compressive stress at which a column will remain in the laterally displaced, buckled, shape.

**E:** Young's modulus of elasticity, which is the slope of the stress versus strain diagram in the initial linear region.

**Euler buckling load:** Same as the critical buckling load if the column is long and slender.

**Euler buckling stress:** Same as the critical buckling stress if the column is long and slender.

### References

- Chajes, A. 1974. *Principles of Structural Stability Theory*. Prentice Hall, Englewood Cliffs, NJ.
- Chen, W. F. and Atsuta, T. 1976. *Theory of Beam-Columns, vol. 1*, In-Plane Behavior and

Design. McGraw-Hill, New York.

Riley, W. F. and Zachary, L. W. 1989. *Introduction to Mechanics of Materials*. John Wiley & Sons, New York.

Young, W. C. 1989. *Roark's Formulas for Stress and Strain*. McGraw-Hill, New York.

## **Further Information**

*Journal of Structural Engineering of the American Society of Civil Engineers.*

*Engineering Journal of American Institute of Steel Construction.*

Structural Stability Research Council. 1976. *Guide to Stability Design Criteria for Metal Structures*, 3rd ed. John Wiley & Sons, New York.

Salmon, C. G. and Johnson, J. E. 1990. *Steel Structures*, 3rd ed. HarperCollins, New York.

Livingston , E., Scavuzzo, R. J. “Pressure Vessels”  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Pressure Vessels

---

## 9.1 Design Criteria

Design Loads • Materials • Allowable Stress

## 9.2 Design Formulas

## 9.3 Opening Reinforcement

### Earl Livingston

*Babcock and Wilcox Company, Retired*

### Rudolph J. Scavuzzo

*University of Akron*

**Pressure vessels** used in industry are leak-tight pressure containers, usually cylindrical or spherical in shape, with different head configurations. They are usually made from carbon or stainless steel and assembled by welding. Early operation of pressure vessels and boilers resulted in numerous explosions, causing loss of life and considerable property damage. Some 80 years ago, the American Society of Mechanical Engineers formed a committee for the purpose of establishing minimum safety rules of construction for boilers. In 1925 the committee issued a set of rules for the design and construction of unfired pressure vessels. Most states have laws mandating that these **Code** rules be met. Enforcement of these rules is accomplished via a third party employed by the state or the insurance company. These Codes are living documents in that they are constantly being revised and updated by committees composed of individuals knowledgeable on the subject. Keeping current requires that the revised Codes be published every three years with addendas issued every year. This chapter covers a very generalized approach to pressure vessel design based on the ASME Boiler and Pressure Vessel Code, Section VIII, Division 1: Pressure Vessels.

## 9.1 Design Criteria

---

The Code design criteria consist of basic rules specifying the design method, design load, allowable stress, acceptable material, and fabrication—inspection certification requirements for vessel **construction**. The design method known as "design by rule" uses design pressure,

allowable stress, and a design formula compatible with the geometry of the part to calculate the minimum required thickness of the part. This procedure minimizes the amount of analysis required to ensure that the vessel will not rupture or undergo excessive distortion. In conjunction with specifying the vessel thickness, the Code contains many construction details that must be followed. Where vessels are subjected to complex loadings such as cyclic, thermal, or localized loads, and where significant discontinuities exist, the Code requires a more rigorous analysis to be performed. This method is known as the "design by analysis" method. A more complete background of both methods can be found in Bernstein, 1988.

The ASME Code [1994] is included as a standard by the American National Standards Institute (ANSI). The American Petroleum Institute (API) has also developed codes for low-pressure storage tanks, and these are also part of the ANSI standards. The ASME Boiler and Pressure Vessel Code has been used worldwide, but many other industrialized countries have also developed boiler and pressure vessel codes. Differences in these codes sometimes cause difficulty in international trade.

## Design Loads

The forces that influence pressure vessel design are internal/external pressure; dead loads due to the weight of the vessel and contents; external loads from piping and attachments, wind, and earthquakes; operating-type loads such as vibration and sloshing of the contents; and startup and shutdown loads. The Code considers design pressure, design temperature, and, to some extent, the influence of other loads that impact the **circumferential** (or hoop) and **longitudinal stresses** in shells. It is left to the designer to account for the effect of the remaining loads on the vessel. Various national and local building codes must be consulted for handling wind and earthquake loadings.

## Materials

The materials to be used in pressure vessels must be selected from Code-approved material specifications. This requirement is normally not a problem since a large catalogue of tables listing acceptable materials is available. Factors that need to be considered in picking a suitable table are:

- Cost

- Fabricability

- Service condition (wear, corrosion, operating temperature)

- Availability

- Strength requirements

Several typical pressure vessel materials for a noncorrosive environment and for service temperatures between  $-50^{\circ}\text{F}$  and  $1000^{\circ}\text{F}$  are shown in [Table 9.1](#).

**Table 9.1** Acceptable Pressure Vessel Materials

Temperature Use Limit ( $^{\circ}$ F)	Plate Material	Pipe Material	Forging Material
Down to -50	SA-516 <sup>a</sup> All grades	SA 333 Gr. 1	SA 350 Gr. LF1, LF2
+33 to +775	SA-285 Gr. C SA-515 Gr. 55, 60, 65 SA-516 All grades	SA-53 SA-106	SA-181 Gr. I, II
+776 to +1000	SA-204 Gr. B, C SA-387 Gr. 11, 12 Class 1	SA-335 Gr. P1, P11, P12	SA-182 Gr. F1, F11, F12

<sup>a</sup> Impact testing required.

*Note:* SA is a classification of steel used in the ASME Boiler and Pressure Vessel Code.

## Allowable Stress

The allowable stress used to determine the minimum vessel wall thickness is based on the material tensile and yield properties at room and design temperature. When the vessel operates at an elevated temperature, the creep properties of the material must also be considered. These properties are adjusted by design factors which limit the **hoop membrane stress** level to a value that precludes rupture, and excessive elastic or plastic distortion and creep rupture. [Table 9.2](#) shows typical allowable stresses for several carbon steels commonly used for unfired pressure vessels.

**Table 9.2** Typical Allowable Stresses for Use in Pressure Vessel Design

Material Specification	Temperature Use Limit ( $^{\circ}$ F)	Allowable Stress (psi)
SA-515 Gr. 60	700	14 400
	800	10 800
	900	6 500
SA-516 Gr. 70	700	16 600
	800	14 500



	900	12 000
SA-53 Gr. A	700	11 700
	800	9 300
	900	6 500
SA-106 Gr. B	700	14 400
	800	10 800
	900	6 500
SA-181 Gr. I	700	16 600
	800	12 000
	900	6 500

---

## 9.2 Design Formulas

The design formulas used in the "design by rule" method are based on the principal stress theory and calculate the average hoop stress. The principal stress theory of failure states that failure occurs when one of the three principal stresses reaches the yield strength of the material. Assuming that the radial stress is negligible, the other two principal stresses can be determined by simple formulas based on engineering mechanics.

The Code recognizes that the shell thickness may be such that the radial stress may not be negligible, and adjustments have been made in the appropriate formulas. [Table 9.3](#) shows various formulas used to calculate the wall thickness for numerous pressure vessel geometries.

**Table 9.3** Code Formulas for Calculation of Vessel Component Thickness

Cylindrical shell	$t = \frac{PR}{SE_j 0.6P}$
Hemispherical head or spherical shell	$t = \frac{PR}{2SE_j 0.2P}$
2:1 ellipsoidal head	$t = \frac{PD}{2SE_j 0.2P}$
Flanged and dished head	$t = \frac{1.77PL}{2SE_j 0.2P}$
Flat head	$t = d \sqrt{\frac{CP}{SE}}$

where

$t$  = Minimum required thickness (in.)

$P$  = Design pressure (psi)

$R$  = Inside radius (in.)

$S$  = Allowable stress (psi)

$D$  = Inside diameter (in.)

$L$  = Inside spherical crown radius (in.)

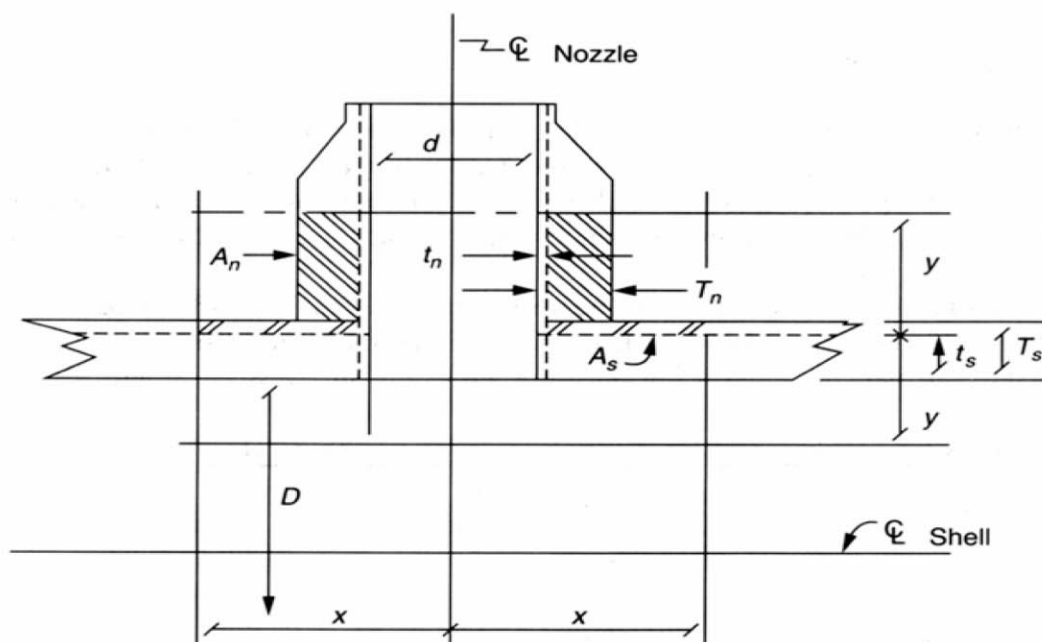
$E$  = Weld joint efficiency factor, determined by joint location and degree of examination

$C$  = Factor depending upon method of head-to-shell attachment

## 9.3 Opening Reinforcement

Vessel components are weakened when material is removed to provide openings for nozzles or access. High **stress concentrations** exist at the opening edge and decrease radially outward from the opening, becoming negligible beyond twice the diameter from the center of the opening. To avoid failure in the opening area, compensation or reinforcement is required. Some ways in which this can be accomplished are: (a) increase the vessel wall thickness, (b) increase the wall thickness of the nozzle, or (c) use a combination of extra shell and nozzle thickness. The Code procedure is to relocate the removed material to an area within an effective boundary around the opening. Figure 9.1 shows the steps necessary to reinforce an opening in a pressure vessel. Numerous assumptions have been made with the intent of simplifying the general approach.

**Figure 9.1** Opening Reinforcement Requirements.



$x$  = Larger of  $d$  or  $R_n + t_n + T_n$   
 $y$  = Smaller of  $2\frac{1}{2}T_s$  or  $2\frac{1}{2}T_n$   
 $d$  = Diameter of circular opening (in.)  
 $D$  = Inside diameter of shell (in.)  
 $t_s$  = Required thickness of shell (in.)  
 $T_s$  = Actual thickness of shell (in.)  
 $t_n$  = Required thickness of nozzle (in.)  
 $T_n$  = Actual thickness of nozzle (in.)  
 $R_n$  = Inside radius of nozzle =  $d/2$  (in.)

$A_r$  = Area of required reinforcement (in.<sup>2</sup>)  
 $A_s$  = Area available in the shell (in.<sup>2</sup>)  
 $A_n$  = Area available in the nozzle (in.<sup>2</sup>)  
 $A_r = (d)(t_s)$   
 $A_s$  = Larger of:  $d(T_s - t_s) - 2T_n(T_s - t_s)$  or  
 $2(T_s + t_n)(T_s - t_s) - 2t_n(T_s - t_s)$   
 $A_n$  = Smaller of:  $2[2\frac{1}{2}(T_s)(T_n - t_n)]$  or  
 $2[2\frac{1}{2}(T_n)(T_n - t_n)]$   
 $A_r < (A_s + A_n)$ : Acceptable configuration

The example shown in Fig. 9.2 uses the design approach indicated by the Code to perform a simple sizing calculation for a typical welded carbon steel vessel. Figure 9.3 shows a typical shell-nozzle juncture and head-shell juncture which meet the code requirements. Design specifications for the many associated vessel parts such as bolted flanges, external attachments, and saddle supports can be found in the reference materials.

**Figure 9.2** Sample vessel calculation.

#### DESIGN SPECIFICATION

Design pressure = 700 psi

Design temperature = 700° F

Material:

Shell SA-516 Gr. 70

Head SA-181 Class 70

Nozzle SA-106 Gr. B

Weld efficiency factor = 1.0 =  $E$   
(full radiographic examination)

Shell Thickness

$$\begin{aligned} t_s &= \frac{PR}{SE - 0.6P} \\ &= \frac{700(30)}{16\,600(1.0) - 0.6(700)} \\ &= 1.30 \text{ in. Use } 1\frac{1}{2}'' = T_s \end{aligned}$$

$P = 700 \text{ psi}$

$R = 30 \text{ in.}$

$E = 1.0$

$S = 16\,600 \text{ psi (SA-516 Gr. 70, Table 9.2)}$

Hemispherical Head Thickness

$$\begin{aligned} t_h &= \frac{PR}{2SE - 0.2P} \\ &= \frac{700(30)}{2(16\,600)(1.0) - 0.2(200)} \\ &= 0.64 \text{ in. Use } 1'' = T_h \end{aligned}$$

Nozzle Thickness

$$\begin{aligned} t_n &= \frac{PR}{SE - 0.6P} \\ &= \frac{700(4)}{16\,600(1.0) - 0.6(700)} \\ &= 0.17 \text{ in. Use } 1\frac{3}{4}'' = T_n \end{aligned}$$

$P = 700 \text{ psi}$

$R = 4 \text{ in.}$

$E = 1.0$

$S = 16\,600 \text{ psi (SA-106 Gr. B, Table 9.2)}$

Opening Reinforcement Calculation (see Fig. 9.1)\*

$$A_{\text{req'd}} = (d)(t_s) = (8)(1.3) = 10.4 \text{ in.}^2$$

$$A_n = \text{Smaller of: } 2[2\frac{1}{2}(T_s)(T_n - t_n)] \text{ or}$$

$$2[2\frac{1}{2}(T_n)(T_n - t_n)]$$

$$T_s < T_n \text{ Use } T_s$$

$$2[2\frac{1}{2}(1\frac{1}{2})(1\frac{3}{4} - 0.17)] = 11.7 \text{ in.}^2$$

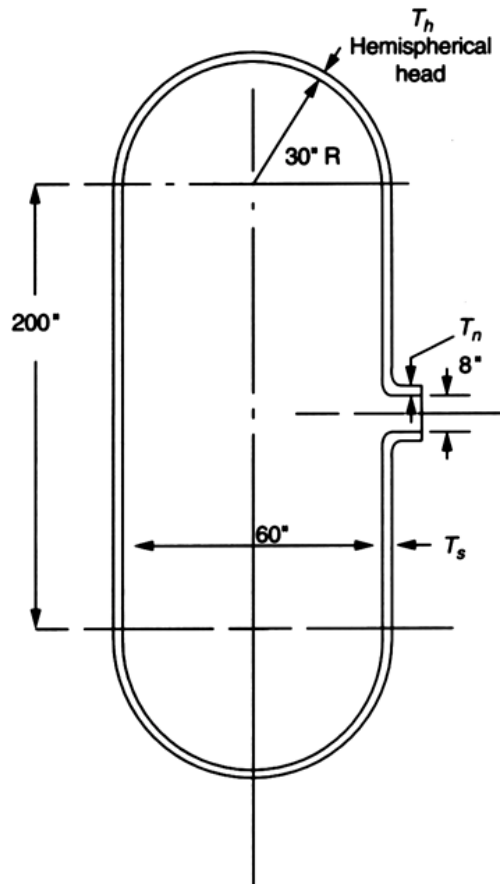
$$d = 8 \text{ in.}$$

$$t_s = 1.3 \text{ in.}$$

$$T_s = 1\frac{1}{2} \text{ in.}$$

$$T_n = 1\frac{3}{4} \text{ in.}$$

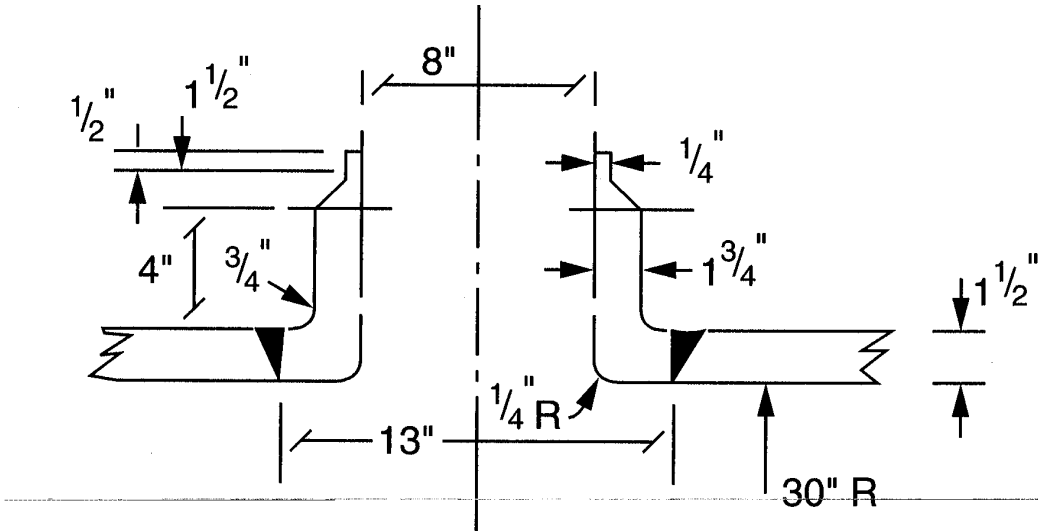
$$t_n = 0.17 \text{ in.}$$



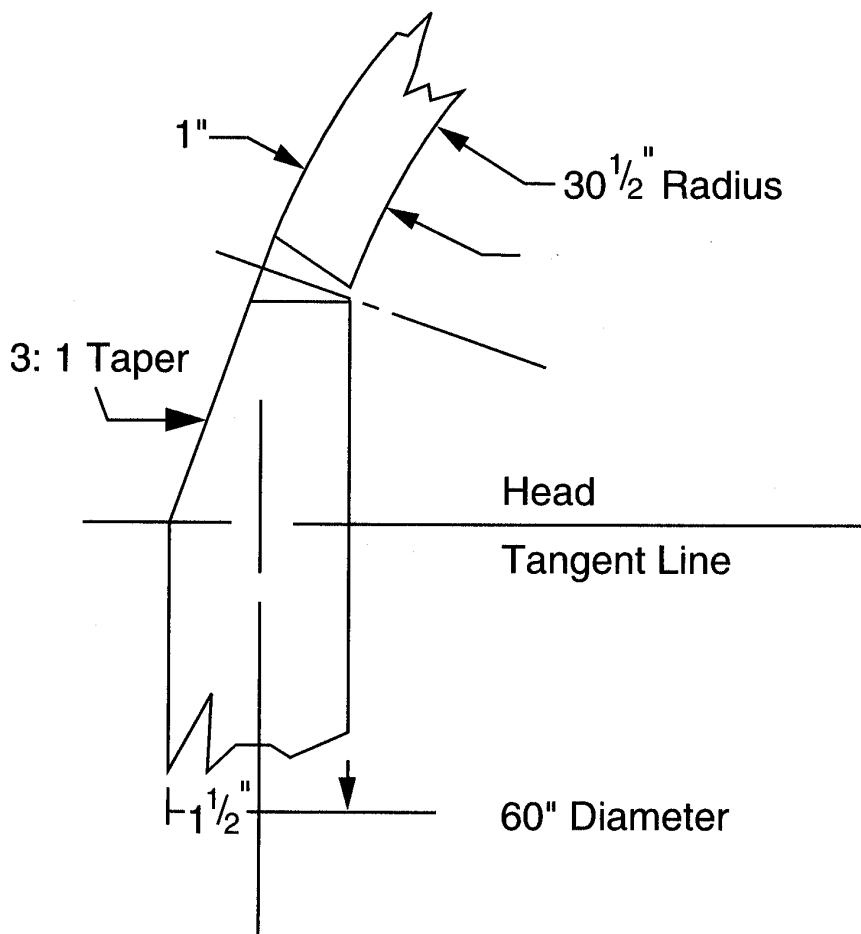
\*This example places all the required reinforcement in the nozzle wall since it is more economical than increasing the shell thickness.

$$A_n = 11.7 \text{ in.}^2 > A_{\text{req'd}} = 10.4 \text{ in.}^2 \text{ — OK}$$

**Figure 9.3** Fabrication details.



Nozzle-Shell Junction



Head-Shell Junction

## Defining Terms

- Code:** The complete rules for construction of pressure vessels as identified in ASME Boiler and Pressure Vessel Code, Section VIII, Division 1, Pressure Vessels.
- Construction:** The complete manufacturing process, including design, fabrication, inspection, examination, hydrotest, and certification. Applies to new construction only.
- Hoop membrane stress:** The average stress in a ring subjected to radial forces uniformly distributed along its circumference.
- Longitudinal stress:** The average stress acting on a cross section of the vessel.
- Pressure vessel:** A leak-tight pressure container, usually cylindrical or spherical in shape, with pressure usually varying from 15 psi to 5000 psi.
- Stress concentration:** Local high stress in the vicinity of a material discontinuity such as a change in thickness or an opening in a shell.
- Weld efficiency factor:** A factor which reduces the allowable stress. The factor depends on the degree of weld examination performed during construction of the vessel.

## References

- American Society of Mechanical Engineers. 1994. *ASME Boiler and Pressure Vessel Code, Section VIII Division 1, Pressure Vessels*. ASME, New York.
- Bednar, H. H. 1981. *Pressure Vessel Design Handbook*. Van Nostrand Reinhold, New York.
- Bernstein, M. D. 1988. Design criteria for boiler and pressure vessels in the U.S.A. *ASME J. F* –443.
- Jawad, M. H. and Farr, J. R. 1989. *Structural Analysis and Design of Process Equipment*, 2nd ed. John Wiley & Sons, New York.

## Further Information

Each summer, usually in June, the Pressure Vessels and Piping Division of the American Society of Mechanical Engineers organizes an annual meeting devoted to pressure vessel technology. Usually 300 to 500 papers are presented, many of which are published by ASME in booklets called special publications. Archival papers are also published in the *Journal of Pressure Vessel Technology*.

Research for ASME Boiler and Pressure Code work is often conducted by the Pressure Vessel Research Council (PVRC). This research is normally published in Welding Research Council bulletins. These bulletins, which number about 400, are an excellent documentation of major research contributions in the field of pressure vessel technology. The Electric Power Research Institute (EPRI) located in Palo Alto, California, also conducts extensive research for the electric power industry, and this work includes some research on pressure



Bauld, Jr.N. R. “Axial Loads and Torsion”  
*The Engineering Handbook.*  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Axial Loads and Torsion

---

## 10.1 Axially Loaded Bars

Axial Strain • Axial Stress • Axial Stress-Strain Relation • Relative Displacement of Cross Sections • Uniform Bar • Nonuniform Bars • Statically Indeterminate Bars

## 10.2 Torsion

Power Transmission • Kinematics of Circular Shafts • Equilibrium • Elastic Twisting of Circular Shafts • Uniform Shaft • Nonuniform Shaft • Statically Indeterminate Circular Shafts

**Nelson R. Bauld, Jr.**

*Clemson University*

---

## 10.1 Axially Loaded Bars

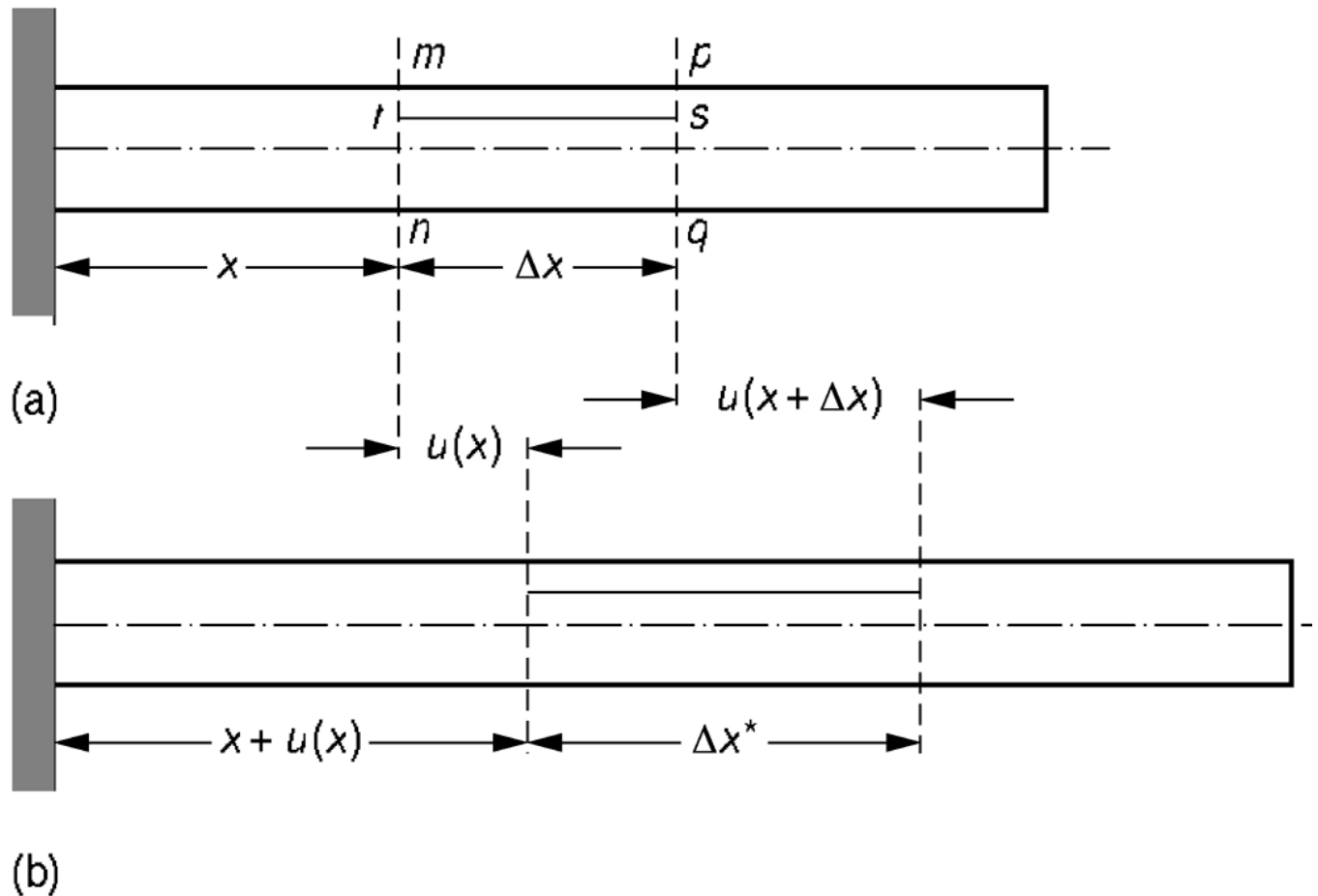
---

A bar is said to be axially loaded if the action lines of all the applied forces coincide with the axis of the bar. The **bar axis** is defined as the locus of the centroids of the cross-sectional areas along the length of the bar. This locus of centroids must form a straight line, and the action lines of the applied forces must coincide with it in order for the theory of this section to apply.

### Axial Strain

The axial strain in an axially loaded bar is based on the geometric assumptions that plane cross sections in the unloaded bar, such as sections  $m\bar{n}$  and  $p\bar{q}$  in [Fig. 10.1\(a\)](#), remain plane in the loaded bar as shown in [Fig. 10.1\(b\)](#), and that they displace only axially.

**Figure 10.1** Axial displacements of an axially loaded bar.



The axial strain of a **line element** such as  $rs$  in Fig. 10.1(a) is defined as the limit of the ratio of its change in length to its original length as its original length approaches zero. Thus, the axial strain  $\epsilon$  at an arbitrary cross section  $x$  is

$$\epsilon(x) = \lim_{\Delta x \rightarrow 0} \frac{(\Delta x + \Delta u) - \Delta x}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{[u(x + \Delta x) - u(x)]}{\Delta x} = \frac{du}{dx} \quad (10:1)$$

where  $u(x)$  and  $u(x + \Delta x)$  are axial displacements of the cross sections at  $x$  and  $x + \Delta x$ . Common units for axial strain are in./in. or mm/mm. Because axial strain is the ratio of two lengths, units for axial strain are frequently not recorded.

## Axial Stress

The axial stress  $\sigma$  at cross section  $x$  of an axially loaded bar is

$$\sigma(x) = \frac{N(x)}{A(x)} \quad (10:2)$$

where  $N(x)$  is the internal force and  $A(x)$  is the cross-sectional area, each at section  $x$ .



Common units for axial stress are pounds per square inch (psi) or megapascals (MPa). Equation (10.2) is valid at cross sections that satisfy the geometric assumptions stated previously. It ceases to be valid at abrupt changes in cross section and at points of load application. Cross sections at such locations distort and therefore violate the plane cross-section assumption. Also, Eq. (10.2) requires that the material at cross section  $x$  be homogeneous; that is, the cross section cannot be made of two or more different materials.

## Axial Stress-Strain Relation

The allowable stress for axially loaded bars used in most engineering structures falls within the proportional limit of the material from which they are made. Consequently, material behavior considered in this section is confined to the linearly elastic range and is given by

$$\epsilon(x) = E(x)^{-1} \sigma(x) \quad (10:3)$$

where  $E(x)$  is the modulus of elasticity for the material at section  $x$ . Common units for the modulus of elasticity are pounds per square inch (psi) or gigapascals (GPa).

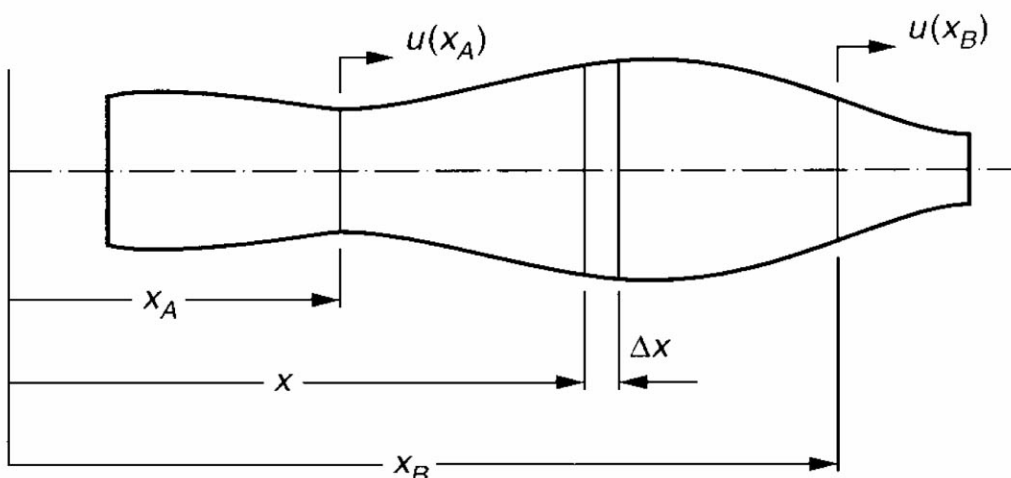
## Relative Displacement of Cross Sections

The relative displacement  $e_{B=A}$  of a cross section at  $x_B$  with respect to a cross section at  $x_A$  is obtained by combining Eqs. (10.1–10.3) and integrating from section  $x_A$  to  $x_B$ . Using [Fig. 10.2](#),

$$e_{B=A} = u(x_B) - u(x_A) = \int_{x_A}^{x_B} N(x) / [A(x)E(x)] dx \quad (10:4)$$

where  $e_{B=A}$  denotes the change in length between the cross sections at  $x_A$  and  $x_B$ .

**Figure 10.2**



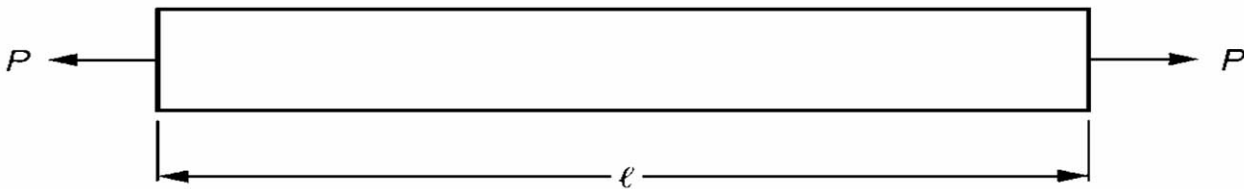
Equation (10.4) must be interpreted as the sum of several integrals for a bar for which the integrand exhibits discontinuities. Discontinuities occur for cross sections where either  $N$ ,  $A$ ,  $E$ , or combinations thereof change abruptly and can usually be detected by inspection.

## Uniform Bar

A bar for which the internal force  $N(x)$ , the cross-sectional area  $A(x)$ , and the modulus of elasticity  $E(x)$  do not change over its length is referred to as a **uniform bar**. If  $P$  denotes equilibrating forces applied to the ends of the bar and  $L$  its length, as shown in Fig. 10.3, then Eq. (10.4) gives the change in length of the bar as

$$e = PL/AE \quad (10:5)$$

**Figure 10.3** Uniform bar.



## Nonuniform Bars

A **nonuniform bar** is one for which either  $A$ ,  $E$ ,  $N$ , or combinations thereof change abruptly along the length of the bar. Three important methods are available to analyze axially loaded bars for which the integrand in Eq. (10.4) contains discontinuities. They are as follows.

### Direct Integration

Equation (10.4) is integrated directly. The internal force  $N(x)$  is obtained in terms of the applied forces via the axial equilibrium equation,  $A(x)$  from geometric considerations, and  $E(x)$  by observing the type of material at a given section.

### Discrete Elements

The bar is divided into a finite number of segments, for each of which  $N/AE$  is constant. Each segment is a uniform bar for which its change in length is given by Eq. (10.5). The change in length of the nonuniform bar is the sum of the changes in length of the various segments. Accordingly, if  $e_i$  denotes the change in length of the  $i$ th segment, then the change in length  $e$  of the nonuniform bar is

$$e = \sum e_i \quad (10:6)$$

## Superposition

The superposition principle applied to axially loaded bars asserts that the change in length between two cross sections caused by several applied forces acting simultaneously is equal to the algebraic sum of the changes in length between the same two cross sections caused by each applied force acting separately. Thus, letting  $e_{B=A}$  represent the change in length caused by several applied forces acting simultaneously, and  $e_{B=A}^0, e_{B=A}^{00}, \dots$  represent the changes in length caused by each applied force acting separately,

$$e_{B=A} = e_{B=A}^0 + e_{B=A}^{00} + \dots \quad (10:7)$$

Superposition of displacements requires that the axial forces be linearly related to the displacements they cause, and this implies that the stress at every cross section cannot exceed the proportional limit stress of the material of the bar. This requirement must be satisfied for each separate loading as well as for the combined loading.

## Statically Indeterminate Bars

The internal force  $N(x)$  in statically determinate axially loaded bars is determined via axial equilibrium alone. Subsequently, axial stress, axial strain, and axial displacements can be determined via the foregoing equations.

The internal force  $N(x)$  in statically indeterminate axially loaded bars cannot be determined via axial equilibrium alone. Thus, it is necessary to augment the axial equilibrium equation with an equation (geometric compatibility equation) that accounts for any geometric constraints imposed on the bar—that is, that takes into account how the supports affect the deformation of the bar.

Three basic mechanics concepts are required to analyze statically indeterminate axially loaded bars: axial equilibrium, geometric compatibility of axial deformations, and material behavior (stress-strain relation).

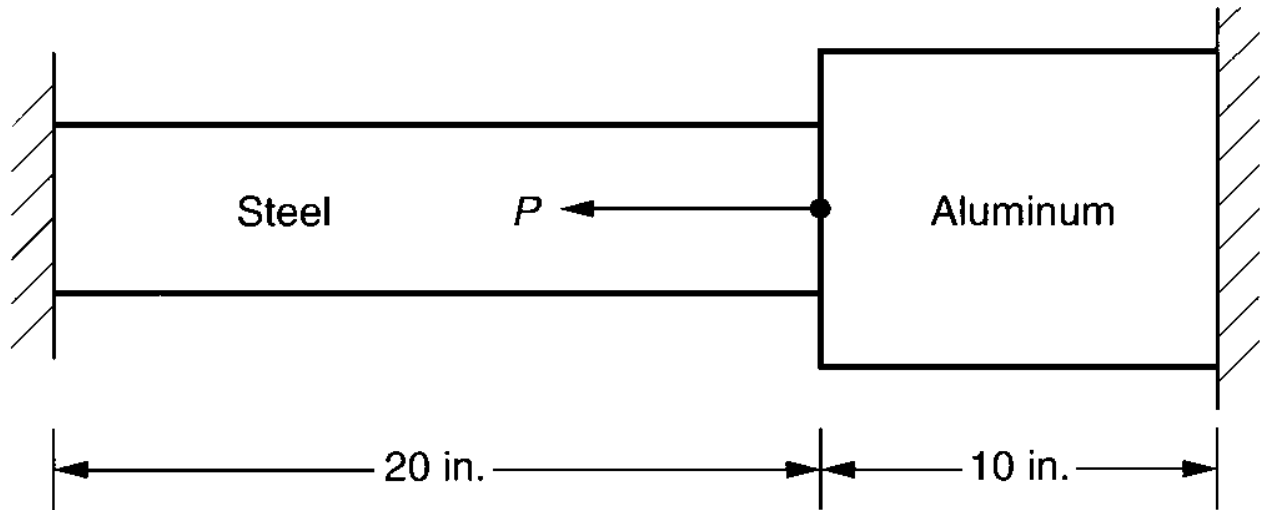
**Example 10.1.** Determine the stresses in the aluminum and steel segments of the composite bar of Fig. 10.4(a) when  $P = 7000$  lb. The cross-sectional areas of the steel and aluminum segments are  $2 \text{ in}^2$  and  $4 \text{ in}^2$ , respectively, and the moduli of elasticity are  $30 \times 10^6$  psi and  $10 \times 10^6$  psi, respectively.

**Solution.** The bar is statically indeterminate; therefore, the solution requires the use of the three mechanics concepts discussed in the previous paragraph.

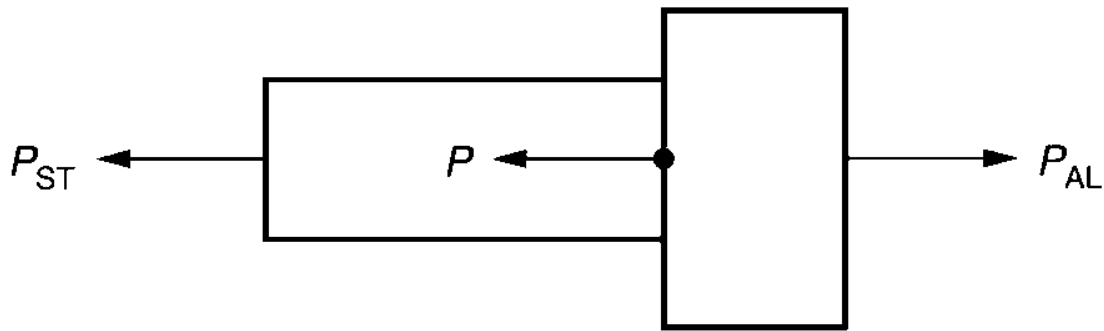
*Equilibrium.* The axial equilibrium equation is obtained from the free-body diagram of Fig. 10.4(b) as

$$-P_{ST} + P_{AL} - 7000 = 0 \quad (10:8)$$

**Figure 10.4** Statically indeterminate composite step-bar.



(a)



(b)

*Geometric compatibility.* The compatibility equation is obtained by noting that the total elongation of the bar is zero. Accordingly,

$$e = e_{ST} + e_{AL} = 0 \quad (10:9)$$

*Material behavior.* The steel and aluminum segments are assumed to behave in a linearly elastic manner, so their elongations are given by

$$e_{ST} = P_{ST} L_{ST} / (A_{ST} E_{ST}) \quad \text{and} \quad e_{AL} = P_{AL} L_{AL} / (A_{AL} E_{AL}) \quad (10:10)$$

Combining Eqs. (10.9) and (10.10) yields

$$\begin{aligned}
 P_{ST} &= \frac{1}{2} (L_{AL}=L_{ST})(E_{ST}=E_{AL})(A_{ST}=A_{AL})P_{AL} \\
 &= \frac{1}{2} (10=20)(30=10)(2=4)P_{AL} = \frac{1}{2} 3=4P_{AL} \quad (10:11)
 \end{aligned}$$

Solving Eqs. (10.8) and (10.11) simultaneously yields

$$P_{ST} = \frac{1}{2} 3000 \text{ lb} \quad \text{and} \quad P_{AL} = 4000 \text{ lb} \quad (10:12)$$

from which the stresses in the steel and aluminum are found as follows:

$$\begin{aligned}
 \sigma_{ST} &= \frac{1}{2} 3000=2 = \frac{1}{2} 1500 \text{ psi} = 1500 \text{ psi (compression)} \\
 \sigma_{AL} &= 4000=4 = 1000 \text{ psi (tension)}
 \end{aligned}$$

**Example 10.2.** Assuming that  $P = 0$  in Fig. 10.4(a), determine the stress in the steel and aluminum segments of the bar due to a temperature increase of  $10^\circ\text{F}$ . The *thermal expansion coefficients* for steel and aluminum are  $\alpha_{ST} = 6.5 \times 10^{-6}$  inches per inch per degree Fahrenheit (in./in./ $^\circ\text{F}$ ) and  $\alpha_{AL} = 13 \times 10^{-6}$  in./in./ $^\circ\text{F}$ .

**Solution.** Because free thermal expansion of the bar is prevented by the supports, internal stresses are induced in the two segments.

*Equilibrium.* The axial equilibrium equation is obtained from the free-body diagram of Fig. 10.4(b). Thus,

$$\frac{1}{2} P_{ST} + P_{AL} = 0 \quad (10:13)$$

*Compatibility.* The compatibility equation is obtained by noting that if the bar could expand freely, its total elongation  $\delta$  would be

$$\delta = \delta_{ST} + \delta_{AL} \quad (10:14)$$

where  $\delta_{ST}$  and  $\delta_{AL}$  denote the free thermal expansions of the separate segments. Because the net change in length of the bar is zero, internal strains are induced in the steel and aluminum such that the sum of the changes in lengths of the steel and aluminum segments must be equal to  $\delta$ . Therefore, the compatibility equation becomes

$$e_{ST} + e_{AL} \frac{1}{2} \delta = 0 \quad (10:15)$$

*Material behavior.* Assuming linear elastic behavior for both materials

$$e_{ST} = \frac{P_{ST} L_{ST}}{A_{ST} E_{ST}} \quad \text{and} \quad e_{AL} = \frac{P_{AL} L_{AL}}{A_{AL} E_{AL}} \quad (10:16)$$

Also, because

$$\phi_{ST} = \frac{P_{ST}}{E_{ST} L_{ST}} \quad \text{and} \quad \phi_{AL} = \frac{P_{AL}}{E_{AL} L_{AL}} \quad (10:17)$$

it follows that

$$\phi = (6.5 \times 10^6)(20)(10) + (13 \times 10^6)(10)(10) = 0.0026 \text{ in.} \quad (10:18)$$

Equations (10.13), (10.15), (10.16), and (10.18) yield

$$P_{ST} f_1 + (E_{ST} = E_{AL})(A_{ST} = A_{AL})(L_{AL} = L_{ST})g = (E_{ST} A_{ST} = L_{ST})\phi$$

or

$$P_{ST} f_1 + (30 \times 10^6)(2 \times 4)(10 \times 20)g = f[30 \times 10^6(2)] = 20g(0.0026)$$

Thus

$$P_{ST} = P_{AL} = 4457 \text{ lb} \quad (10:19)$$

The corresponding stresses in the steel and aluminum are compression and equal to

$$\frac{3}{4}_{ST} = 4457/2 = 2228 \text{ psi} \quad \text{and} \quad \frac{3}{4}_{AL} = 4457/4 = 1114 \text{ psi}$$

## 10.2 Torsion

---

Torsionally loaded bars occur frequently in industrial applications such as shafts connecting motor-pump and motor-generator sets; propeller shafts in airplanes, helicopters, and ships; and torsion bars in automobile suspension systems. Many tools or tool components possess a dominant torsional component such as screwdrivers and drill and router bits. (These tools also rely on an axial force component for their effectiveness.)

### Power Transmission

The specifications for a motor customarily list the power it transmits in horsepower (hp), and its angular speed in either revolutions per minute (rpm) or in cycles per second (Hz). To design or analyze a shaft, the **torque** that it is to transmit is required. Therefore, a relationship between horsepower, angular speed, and torque is required. In U.S. customary units and in the International System of Units (SI units) these relationships are

$$\begin{aligned} \text{hp} &= \frac{2\pi nT}{33,000} = \frac{2\pi fT}{5252} \quad (\text{U.S. customary units}) \\ &= \frac{2\pi fT}{9.549} = fT \quad (\text{SI units}) \end{aligned} \quad (10:20)$$

where  $f$  and  $n$  denote the angular speed in cycles per second and revolutions per minute, respectively, and  $T$  denotes the torque transmitted in Newton-meters (N·m) or inch-pounds (in.-lb), depending on the system of units used.

## Kinematics of Circular Shafts

The theory of circular shafts is based on the geometric assumption that a plane cross section simply rotates about the axis of the shaft and can be visualized as being composed of a series of **thin rigid disks** that rotate about the axis of the shaft.

To obtain a formula that expresses the rotation of one cross section relative to another infinitesimally close to it, consider a shaft of radius  $c$  and examine the angular deformations of an interior segment of radius  $r$  and length  $\Delta x$ . This portion of the bar is indicated in Fig. 10.5(a). Before twisting, line element AB is parallel to the shaft axis, and line element AC lies along a cross-sectional circle of radius  $r$ . The angle between these elements is 90 degrees. Due to twisting, AC merely moves to a new location on the circumference, but AB becomes A'B', which is no longer parallel to the shaft axis, as is indicated in Fig. 10.5(b). The shearing deformation  $e_r$  at radius  $r$  is

$$e_r = r\Delta\theta = \gamma_r \Delta x \quad (10:21)$$

where  $\gamma_r$  denotes the shearing strain between line elements A'B' and A'C', and  $\Delta\theta$  represents the angular rotation of the cross section at B relative to the cross section at A. In the limit, as  $\Delta x$  becomes infinitesimal, Eq. (10.21) becomes

$$\gamma_r = r d\theta = dx \quad (10:22)$$

Because a cross section is considered rigid, Eq. (10.22) indicates that the shearing strain varies linearly with distance from the center of the shaft. Consequently, because  $c$  denotes the outside radius of the shaft, the shearing strain at radius  $r$  is

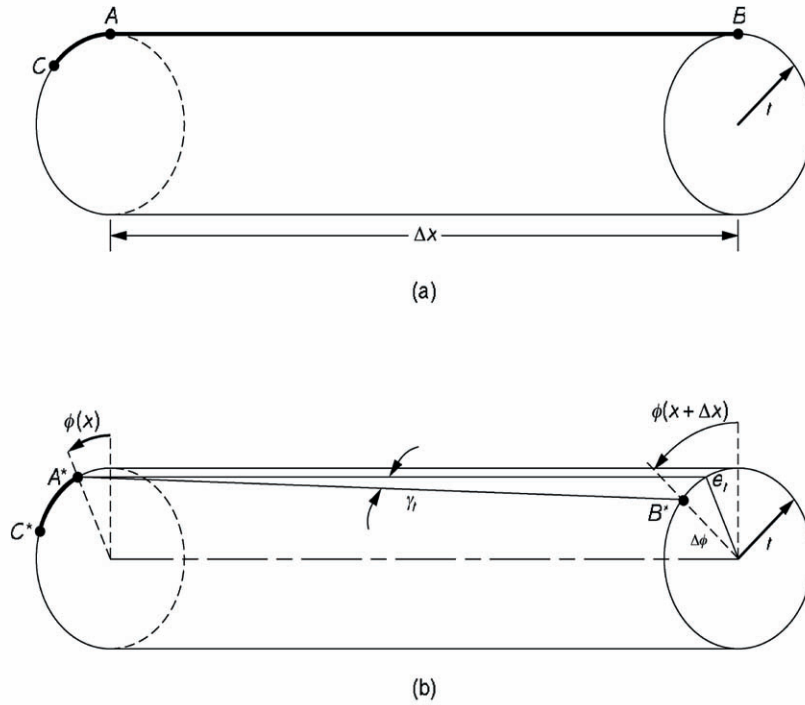
$$\gamma_r = (r/c)\gamma_c \quad (10:23)$$

## Equilibrium

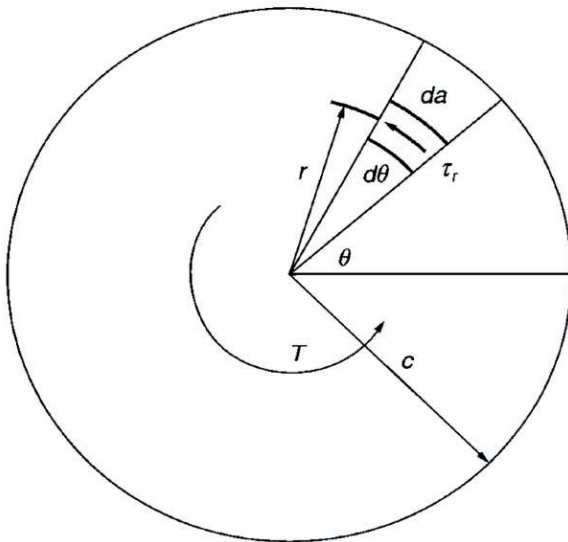
The shearing stress  $\tau_r$  that acts on a differential element of cross-sectional area  $da$  is shown in Fig. 10.6. A concentrated torque  $T$  that is equivalent to the torque produced by the distributed shearing stress  $\tau_r$  is

$$T = \int_{\text{area}} (\tau_r da)r \quad (10:24)$$

**Figure 10.5**



**Figure 10.6**



## Elastic Twisting of Circular Shafts

Explicit formulas for the angle of twist per unit length and for the shearing stress at any point  $r$  in a cross section of a circular shaft made from a linearly elastic material are obtained from Eqs. (10.22) and (10.24) and the stress-strain relation

$$\gamma_r = G \phi_r \quad (10:25)$$



in which  $G$  is the shearing modulus of elasticity. Common units for  $G$  are pounds per square inch (psi) or gigapascals (GPa). Accordingly,

$$T = \int_{\text{area}} (G \tau_r = r) r^2 da = G \int_{\text{area}} d\bar{A} = dx \int_{\text{area}} r^2 da$$

or

$$d\bar{A} = dx = T/JG \quad (10:26)$$

in which  $J$  is the polar moment of inertia of the cross-sectional area of the bar. Common units for  $J$  are inches to the fourth power (in<sup>4</sup>) or meters to the fourth power (m<sup>4</sup>), depending on the system of units used.

The shearing stress at radius  $r$  is obtained by combining Eqs. (10.22), (10.25), and (10.26). Thus,

$$\tau_r = Tr/J \quad (10:27)$$

Equations (10.26) and (10.27) provide the means needed to analyze the strength and stiffness of linearly elastic shafts with circular cross sections. These formulas remain valid for annular shafts for which the hollow and solid portions are concentric. Formulas for the polar moments of inertia  $J$  are

$$J = \begin{cases} \frac{1}{4} \pi d^4 & \text{(solid cross section)} \\ \frac{1}{4} \pi (d_o^4 - d_i^4) & \text{(annular cross section)} \end{cases} \quad (10:28)$$

where  $d_o$  and  $d_i$  denote external and internal diameters.

## Uniform Shaft

A **uniform shaft** is one for which the cross-sectional area, the shearing modulus of elasticity, and the applied torque do not change along its length. Because  $J$ ,  $G$ , and  $T$  are constants over the length  $L$ ; Eq. (10.26) integrates to give the angle of twist of one end relative to the other end as

$$\bar{A} = TL = JG \quad (10:29)$$

The shearing stress on any cross section at radial distance  $r$  is

$$\tau_r = Tr/J \quad (10:30)$$

## Nonuniform Shaft

A **nonuniform shaft** is one for which either  $J$ ;  $G$ ;  $T$ ; or a combination thereof changes abruptly along the length of the shaft. Three procedures are available to determine the angle of twist for circular shafts made from linearly elastic materials.

### Direct Integration

Equation (10.26) is integrated directly. Because the integrand  $T = JG$  can possess discontinuities at cross sections for which  $J$ ;  $G$ ; or  $T$  changes abruptly, the integration must be interpreted as a sum of several integrations. Discontinuities in  $J$ ;  $G$ ; and  $T$  can usually be detected by inspection. The polar moment of inertia  $J$  is discontinuous at abrupt changes in cross-sectional area,  $G$  is discontinuous at cross sections where the material changes abruptly, and the internal torque  $T$  is discontinuous at points where concentrated torques are applied.

### Discrete Elements

The shaft is divided into a finite number of segments for each of which  $T = JG$  is constant. Consequently, the shaft is perceived to be a series of connected uniform shafts for each of which Eq. (10.29) applies. Thus, if  $\hat{A}_i$  denotes the angle of twist of the  $i$ th segment, then the angle of twist for the shaft is

$$\hat{A} = \sum \hat{A}_i \quad (10:31)$$

### Superposition

The superposition principle applied to the twisting of circular shafts stipulates that the relative rotation of one cross section with respect to another cross section due to several torques applied simultaneously is equal to the algebraic sum of the relative rotations of the same cross sections due to each torque applied separately. If  $\hat{A}_{B=A}^0$ ;  $\hat{A}_{B=A}^{00}$ ;  $\dots$  denote relative angles of twist for each torque applied separately, then

$$\hat{A}_{B=A} = \hat{A}_{B=A}^0 + \hat{A}_{B=A}^{00} + \dots \quad (10:32)$$

Superposition of angles of twist requires that the torques be linearly related to the angles of twist that they produce, which in turn implies that the shearing stress must not exceed the proportional limit stress for the material involved. This requirement must be satisfied for each separate loading, as well as for the combined loading.

## Statically Indeterminate Circular Shafts

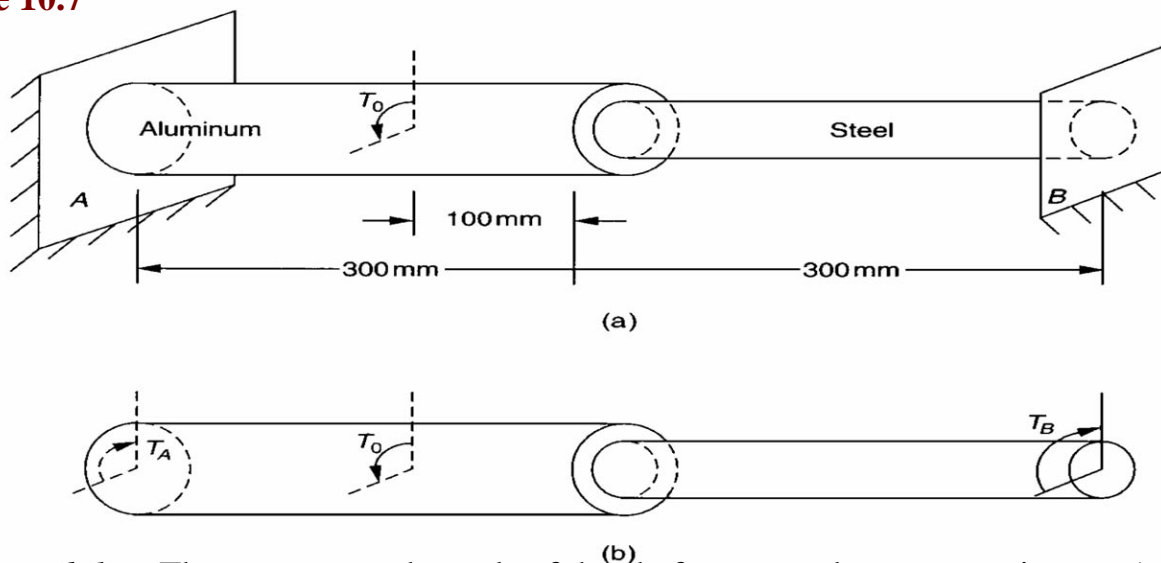
A shaft is statically indeterminate if the internal torque at a cross section cannot be determined from moment equilibrium about the axis of the shaft. In such cases an additional equation is obtained by requiring that angles of twist be compatible with the geometric constraints imposed on the shaft. As with axially loaded bars, three basic concepts of mechanics are involved in the solution of statically indeterminate shafts: equilibrium, geometric compatibility, and material behavior.

**Example 10.3.** The diameters of the aluminum and steel segments of the statically indeterminate step-shaft of Fig. 10.7(a) are 50 mm and 25 mm, respectively. Knowing that  $G_{AL} = 28 \text{ GPa}$ ,  $G_{ST} = 84 \text{ GPa}$ , and  $T_0 = 200 \text{ N} \cdot \text{m}$ , determine the maximum shearing stresses in the aluminum and in the steel.

**Solution.** *Equilibrium.* From Fig. 10.7(b), moment equilibrium about the axis of the shaft gives

$$T_A + T_B + T_0 = 0 \quad (10:33)$$

**Figure 10.7**



*Compatibility.* The supports at the ends of the shaft prevent the cross sections at A and B from rotating; hence, the required compatibility equation is

$$\hat{A}_{B=A} = 0 \quad (10:34)$$

and, with the aid of the superposition principle, it can be written as

$$\hat{A}_{B=A} = \hat{A}_{B=A}^0 + \hat{A}_{B=A}^{00} = 0 \quad (10:35)$$

Here  $\hat{A}_{B=A}^0$  and  $\hat{A}_{B=A}^{00}$  denote the relative angular rotations of the cross section at B with respect to the cross section at A due to the torques  $T_B$  and  $T_0$  acting separately.

To convert Eq. (10.35) into an algebraic equation involving the torques  $T_B$  and  $T_0$ , the discrete element procedure is used. First calculate the polar moments of inertia for the two segments:

$$\begin{aligned} J_{AL} &= \frac{1}{2} \pi (0.050)^4 = 0.613 \times 10^{-6} \text{ m}^4 \\ J_{ST} &= \frac{1}{2} \pi (0.025)^4 = 0.038 \times 10^{-6} \text{ m}^4 \end{aligned} \quad (10.36)$$

Using Eq. (10.29) for a uniform shaft, determine that

$$\begin{aligned} \phi_{B=A}^0 &= 0.3 T_B / J_{AL} (28 \times 10^9) + 0.3 T_B / J_{ST} (84 \times 10^9) = 111.47 \times 10^6 T_B / \text{m}^4 \\ \phi_{B=A}^{00} &= 0.2 T_0 / J_{AL} (28 \times 10^9) = 11.65 \times 10^6 T_0 / \text{m}^4 \end{aligned} \quad (10.37)$$

Consequently,

$$\phi_{B=A} = 111.47 T_B + 11.65 T_0 \times 10^6 = 0 \quad (10.38)$$

Equation (10.38) gives  $T_B$  and Eq. (10.33) gives  $T_A$ : Thus,

$$T_A = 179 \text{ N} \cdot \text{m} \quad \text{and} \quad T_B = 21 \text{ N} \cdot \text{m} \quad (10.39)$$

The maximum shearing stress in each material occurs at the most remote point on a cross section. Thus,

$$\begin{aligned} (\tau_{AL})_{\max} &= T_{AL} C / J_{AL} = 179 (0.025) / 0.613 \times 10^{-6} = 22.9 \text{ MPa} \\ (\tau_{ST})_{\max} &= T_{ST} C / J_{ST} = 21 (0.0125) / 0.038 \times 10^{-6} = 21.7 \text{ MPa} \end{aligned} \quad (10.40)$$

## Defining Terms

**Bar axis:** Straight line locus of centroids of cross sections along the length of a bar.

**Line element:** Imaginary fiber of material along a specific direction.

**Nonuniform bar:** A bar for which the cross-sectional area or the material composition changes abruptly along its length, or external forces are applied intermediate to its ends.

**Nonuniform shaft:** A bar of circular cross section for which the diameter or material composition changes abruptly along its length, or external twisting moments are applied intermediate to its ends.

**Thin rigid disk:** Imaginary circular cross section of infinitesimal thickness that is assumed to undergo no deformations in its plane.

**Torque:** Twisting moment.

**Uniform bar:** A bar of uniform cross-sectional area that is made of one material and is subjected to axial forces only at its ends.

**Uniform shaft:** A bar of uniform, circular cross-sectional area that is made of one material and is subjected to twisting moments only at its ends.

## References

- Bauld, N. R., Jr. 1986. Axially loaded members and torsion. In *Mechanics of Materials*, 2nd ed.
- Beer, F. P. and Johnston, E. R., Jr. 1981. Stress and strain—axial loading and torsion. In *Mechanics of Materials*.
- Gere, J. M. and Timoshenko, S. P. 1990. Axially loaded members and torsion. In *Mechanics of Materials*, 2nd ed.

## Further Information

Formulas for the twisting of shafts with the following cross-sectional shapes can be found in Bauld [1986]: thin-wall, open sections of various shapes; solid elliptical, rectangular, and equilateral triangular sections; open sections composed of thin rectangles; and circular sections composed of two different concentric materials. Also available in the same reference are formulas for the twisting of circular shafts in the inelastic range.

Karnopp, B. "Dynamics and Vibrations"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# III

## Dynamics and Vibrations

---



Noise, vibration, and harshness (NVH) testing for Chrysler automobiles is carried out in this hemi-anechoic chamber located at the Chrysler Technology Center in Auburn Hills, MI. The chamber is dedicated to finding sources of NVH using an attached chassis dynamometer which replicates real-world driving conditions.

A high-tech test dummy along with various test equipment is placed inside the vehicle. The test dummy has built-in microphones in its ears and rides in the vehicle as if it were driving on the road. The test dummy monitors the sounds of the road and engine as well as outside noises piped into the chamber.

This type of testing determines the inherent quietness and comfort of the vehicle and its ability to shield the cabin occupants from outside noise. The data gathered from this type of testing are used to improve interior sound and quality. (Photo courtesy of Chrysler Corporation.)

# III

## Dynamics and Vibrations

---

**Bruce Karnopp**

*University of Michigan*

- 12 **Dynamics of Particles: Kinematics and Kinetics** *B. Karnopp*  
Dynamics of Particles • Newton's Second Law • Moment of Momentum Relations • Momentum • Integrals of Newton's Second Law • Work-Energy Integral of Newton's Second Law • Conclusion
- 13 **Dynamics of Rigid Bodies: Kinematics and Kinetics** *A. A. Zeid and R. R. Beck*  
Kinematics of Rigid Bodies • Kinetics of Rigid Bodies
- 14 **Free Vibration, Natural Frequencies, and Mode Shapes** *D. A. Mendelsohn*  
Basic Principles • Single-Degree-of-Freedom Systems • Multiple-Degree-of-Freedom Systems • Continuous Systems (Infinite DOF)
- 15 **Forced Vibrations** *A. W. Leissa*  
Single-Degree-of-Freedom Systems • Multiple-Degree-of-Freedom Systems
- 16 **Lumped versus Distributed Parameter Systems** *B. A. Ovunc*  
Procedure of Analysis • Continuous Mass Matrix Method • Consistent Mass Matrix Method • Lumped Mass Matrix Method • Free Vibration of Frames • Forced Vibration • Practical Applications • Structures without Additional Effects • Structures with Additional Effects
- 17 **Applications of Structural and Dynamic Principles** *A. J. Kalinowski*  
Base Configuration Loaded Applications • Structural Configuration Loaded Applications • Additional Information
- 18 **Computer Simulation and Nomographs** *D. J. Inman*  
Nomograph Fundamentals • Models for Numerical Simulation • Numerical Integration • Vibration Response by Computer Simulation • Commercial Software for Simulation
- 19 **Test Equipment and Measuring Instruments** *T. W. Baird*  
Vibration and Shock Test Machines • Transducers and Signal Conditioners • Digital Instrumentation and Computer Control

ALL OF THE MECHANICALLY BASED DISCIPLINES of engineering (e.g., mechanical, aeronautical, civil, naval architecture) are grounded in three basic areas: dynamics of particles and rigid bodies, mechanics of deformable bodies, and thermodynamics and heat transfer. In this section dynamics and the related area of vibrations are discussed. The subject of dynamics has two major components: the geometry of motion (kinematics) and the dynamic equations of motion (kinetics). Three basic models are used in the study of dynamics:

- Mass particle (a point mass or a finite mass which does not rotate)
- Rigid body (an accumulation of mass in a rigid configuration)
- Deformable body (a continuous distribution of mass which changes shape under load)

In **Chapter 12** kinematics and kinetics of point masses (particles) are discussed. In **Chapter 13** this discussion is extended to rigid bodies. Both chapters discuss the measures of force, mass,



velocity, and acceleration and the concepts of momentum (linear and angular) and kinetic energy. In **Chapters 14** through **19** the various aspects of vibrations are discussed. Vibration is a subject of great importance in engineering. Any system with mass and elasticity is a "vibration waiting to happen." Under the right circumstances, such a system will vibrate, and invariably such vibration must be dealt with by the engineer. There are two basic concerns in vibration analysis: free motion and the response of a system to an excitation. In **Chapter 14** the ideas of natural frequencies and modal solutions are introduced. In **Chapter 15** the response to an excitation (forcing function) is discussed. In **Chapter 16** vibrations of continuous systems are discussed. Finally, in **Chapters 17** through **19**, examples of applications are given, computer methods are discussed, and test equipment and methods are presented. The fundamental ideas of dynamics and vibrations are extremely important in themselves. But they also form a framework for other, related areas such as automatic control and acoustics.

Anderson T. L. "Fracture Mechanics"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 11

## Fracture Mechanics

---

11.1 Fundamental Concepts

11.2 The Energy Criterion

11.3 The Stress Intensity Approach

11.4 Time-Dependent Crack Growth and Damage Tolerance

11.5 Effect of Material Properties on Fracture

Concluding Remarks

**Ted L. Anderson**

*Structural Reliability Technology*

Since the advent of iron and steel structures during the Industrial Revolution, a significant number of brittle fractures have occurred at stresses well below the tensile strength of the material. One of the most famous of these failures was the rupture of a molasses tank in Boston in January 1919 [Shank, 1953]. Over 2 million gallons of molasses were spilled, resulting in 12 deaths, 40 injuries, massive property damage, and several drowned horses.

The traditional strength-of-materials approach cannot explain events such as the molasses tank failure. In the first edition of his elasticity text, published in 1892, Love remarked that "the conditions of rupture are but vaguely understood." Designers typically applied safety factors of 10 or more (based on the tensile strength) in an effort to avoid these seemingly random failures.

Several centuries earlier, Leonardo da Vinci performed a series of experiments on iron wires that shed some light on the subject of brittle fracture. He found that the strength of the wires varied inversely with length. These data implied that flaws in the material controlled the strength; a longer wire corresponded to a larger sample volume and a higher probability of sampling a region containing a flaw. These results were only qualitative, however, and formal mathematical relationships between flaw size and failure stress were not developed until recently.

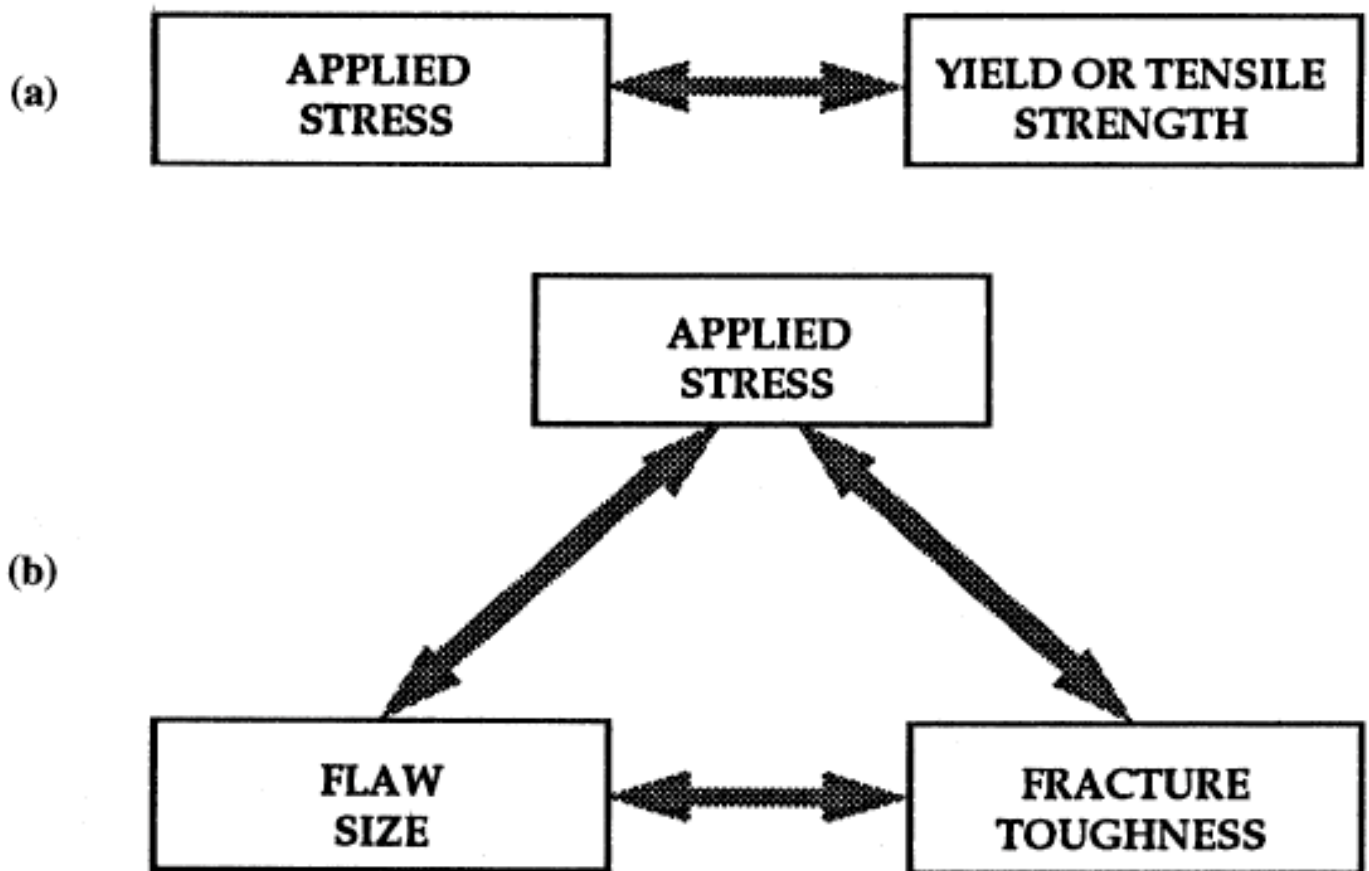
During World War II, a large number of Liberty ships and T2 tankers sustained brittle fractures [Bannerman and Young, 1946]. The need to understand the cause of these failures led to extensive research in the 1950s, which resulted in the engineering discipline known as **fracture mechanics**.

The field of fracture mechanics attempts to quantify the relationship between failure stress, flaw size, and material properties. Today, many segments of industry—including aerospace, oil and gas, and electric utilities—apply fracture mechanics principles in order to prevent catastrophic failures.

## 11.1 Fundamental Concepts

Figure 11.1 contrasts the fracture mechanics approach with the traditional approach to structural design and material selection. In the latter case the anticipated design stress is compared to the flow properties of candidate materials; a material is assumed to be adequate if its strength is greater than the expected applied stress. Such an approach may attempt to guard against brittle fracture by imposing a safety factor on stress, combined with minimum tensile elongation requirements on the material. The fracture mechanics approach [Fig. 11.1(b)] has three important variables, rather than two as in Fig. 11.1(a). The additional structural variable is flaw size, and **fracture toughness** replaces strength as the relevant material property. Fracture mechanics quantifies the critical combinations of these three variables.

**Figure 11.1** Comparison of the fracture mechanics approach to design with the traditional strength of materials approach. (a) The strength of materials approach. (b) The fracture mechanics approach.



Most fracture mechanics methodologies assume linear elastic behavior, although more advanced approaches incorporate nonlinear material behavior such as yielding. There are two alternative approaches to **linear elastic fracture analysis (LEFM)**: the energy criterion and the stress intensity approach, both of which are described below.

In addition to predicting the conditions for ultimate failure, fracture mechanics methodologies can also analyze time-dependent cracking mechanisms such as fatigue.

## 11.2 The Energy Criterion

The energy approach states that crack extension (i.e., fracture) occurs when the energy available for crack growth is sufficient to overcome the resistance of the material. The material resistance may include the surface energy, plastic work, or other type of energy dissipation associated with a propagating crack.

Griffith [1920] was the first to propose the energy criterion for fracture, but Irwin [1956] is primarily responsible for developing the present version of this approach: the **energy release rate**,  $G$ , which is defined as the rate of change in potential energy with crack area for a linear elastic material. At the moment of fracture,  $G = G_c$ , the critical energy release rate, which is a measure of fracture toughness.

For a crack of length  $2a$  in an infinite plate subject to a remote tensile stress (Fig. 11.2), the energy release rate is given by

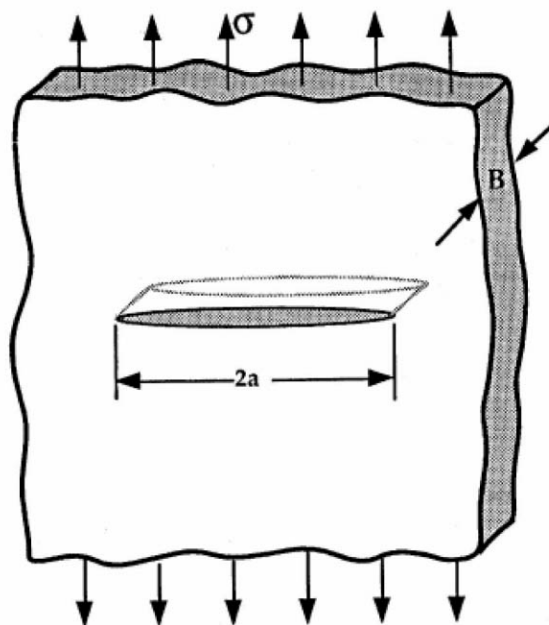
$$G = \frac{\pi \sigma^2 a}{E} \quad (11.1)$$

where  $E$  is Young's modulus,  $\sigma$  is the remotely applied stress, and  $a$  is the half crack length. At fracture,  $G = G_c$ , and Eq. (11.2) describes the critical combinations of stress and crack size for failure:

$$G_c = \frac{\pi \sigma_f^2 a_c}{E} \quad (11.2)$$

Note that for a constant  $G_c$  value, failure stress,  $\sigma_f$ , varies with  $1/\sqrt{a}$ . The energy release rate,  $G$ , is the driving force for fracture, while  $G_c$  is the material's resistance to fracture. To draw an analogy to the strength of materials approach of Fig. 11.1(a), the applied stress can be viewed as the driving force for plastic deformation, whereas the yield strength is a measure of the material's resistance to deformation.

**Figure 11.2** Through-thickness crack in an infinite plate subject to a remote tensile stress. In practical terms, "infinite" means that the width of the plate is  $\gg 2a$ .

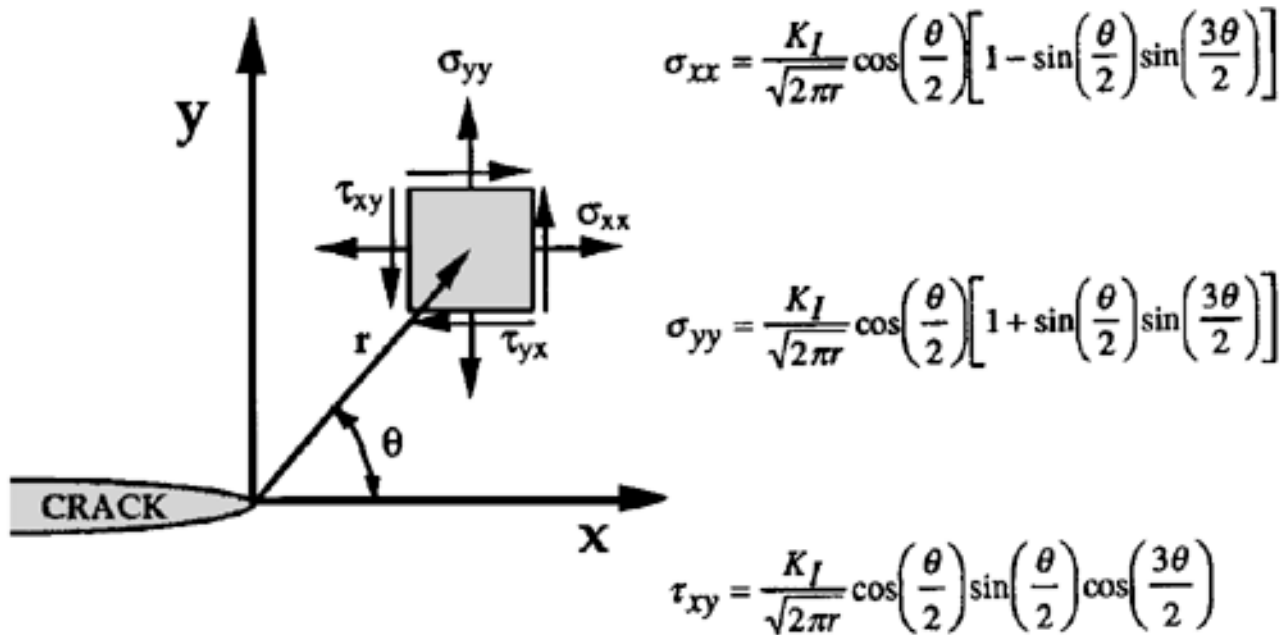


The tensile stress analogy is also useful for illustrating the concept of similitude. A yield strength value measured with a laboratory specimen should be applicable to a large structure; yield strength does not depend on specimen size, provided that the material is reasonably homogeneous. One of the fundamental assumptions of fracture mechanics is that fracture toughness ( $G_c$  in this case) is independent of the size and geometry of the cracked body; a fracture toughness measurement on a laboratory specimen should be applicable to a structure. As long as this assumption is valid, all configuration effects are taken into account by the driving force,  $G$ . The similitude assumption is valid as long as the material behavior is predominantly linear elastic.

## 11.3 The Stress Intensity Approach

Figure 11.3 schematically shows an element near the tip of a crack in an elastic material, together with the in-plane stresses on this element. Note that each stress component is proportional to a single constant,  $K_I$ . If this constant is known, the entire stress distribution at the crack tip can be computed with the equations in Fig. 11.3. This constant, which is called the *stress intensity factor*, completely characterizes the crack tip conditions in a linear elastic material [Irwin, 1957]. If one assumes that the material fails locally at some critical combination of stress and strain, then it follows that fracture must occur at a critical stress intensity,  $K_{IC}$ . Thus,  $K_{IC}$  is an alternate measure of fracture toughness.

**Figure 11.3** Stresses near the tip of a crack in an elastic material.



For the plate illustrated in Fig. 11.2, the stress intensity factor is given by

$$K_I = \sigma \sqrt{\pi a} \quad (11.3)$$

Failure occurs when  $K_I = K_{IC}$ . In this case,  $K_I$  is the driving force for fracture and  $K_{IC}$  is a measure of material resistance. As with  $G_c$ , the property of similitude should apply to  $K_{IC}$ . That is,  $K_{IC}$  is assumed to be a size-independent material property.

Comparing Eqs. (11.1) and (11.3) results in a relationship between  $K_I$  and  $G$ :

$$G = \frac{K_I^2}{E} \quad (11.4)$$

This same relationship obviously holds for  $G_c$  and  $K_{IC}$ . Thus, the energy and stress intensity approaches to fracture mechanics are essentially equivalent for linear elastic materials.

## 11.4 Time-Dependent Crack Growth and Damage Tolerance

Fracture mechanics often plays a role in life prediction of components that are subject to time-dependent crack growth mechanisms such as fatigue or stress corrosion cracking. The *rate* of cracking can be correlated with fracture mechanics parameters such as the stress intensity factor, and the critical crack size for failure can be computed if the fracture toughness is known. For example, Paris and Erdogan [1960] showed that the fatigue crack growth rate in metals could be described by the following empirical relationship:

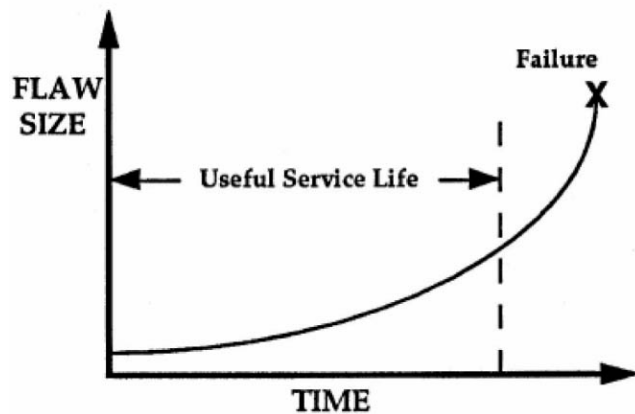
$$\frac{da}{dN} = C(\Delta K)^m \quad (11.5)$$

where  $da/dN$  is the crack growth per cycle,  $\Delta K$  is the stress intensity range, and  $C$  and  $m$  are material constants.

**Damage tolerance**, as its name suggests, entails allowing subcritical flaws to remain in a structure. Repairing flawed material or scrapping a flawed structure is expensive and is often unnecessary. Fracture mechanics provides a rational basis for establishing flaw tolerance limits.

Consider a flaw in a structure that grows with time (e.g., a fatigue crack or a stress corrosion crack) as illustrated schematically in Fig. 11.4. The *initial* crack size is inferred from **nondestructive examination (NDE)**, and the *critical* crack size is computed from the applied stress and fracture toughness. Normally, an *allowable* flaw size would be defined by dividing the critical size by a safety factor. The predicted service life of the structure can then be inferred by calculating the time required for the flaw to grow from its initial size to the maximum allowable size.

**Figure 11.4** The damage tolerance approach to design.



## 11.5 Effect of Material Properties on Fracture

---

Most early work was applicable only to linear elastic materials under quasistatic conditions, whereas subsequent advances in fracture research incorporated other types of material behavior. Elastic-plastic fracture mechanics considers plastic deformation under quasistatic conditions, whereas dynamic, viscoelastic, and viscoplastic fracture mechanics include time as a variable. Elastic-plastic, viscoelastic, and viscoplastic fracture behavior are sometimes included in the more general category of **nonlinear fracture mechanics**. The branch of fracture mechanics one should apply to a particular problem obviously depends on material behavior.

Consider a cracked plate (Fig. 11.2) that is loaded to failure. Figure 11.5 is a schematic plot of failure stress versus fracture toughness ( $K_{IC}$ .) For low-toughness materials, brittle fracture is the governing failure mechanism, and critical stress varies linearly with  $K_{IC}$  as predicted by Eq. (11.3). At very high toughness values, LEFM is no longer valid and failure is governed by the flow properties of the material. At intermediate toughness levels, there is a transition between brittle fracture under linear elastic conditions and ductile overload. Nonlinear fracture mechanics bridges the gap between LEFM and collapse. If toughness is low, LEFM is applicable to the problem; but if toughness is sufficiently high, fracture mechanics ceases to be relevant to the problem because failure stress is insensitive to toughness. A simple limit load analysis is all that is required to predict failure stress in a material with very high fracture toughness.



**Figure 11.5** Effect of fracture toughness on the governing failure mechanism.

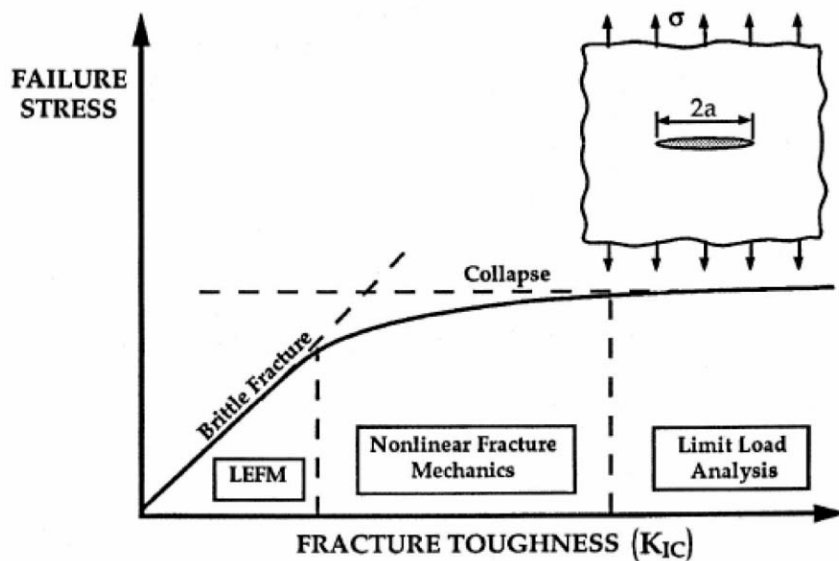


Table 11.1 lists various materials, together with the typical fracture regime for each material.

**Table 11.1** Typical Fracture Behavior of Selected Materials

Material	Typical Fracture Behavior
High-strength steel	Linear elastic
Low- and medium-strength steel	Elastic-plastic/fully plastic
Austenitic stainless steel	Fully plastic
Precipitation-hardened aluminum	Linear elastic
Metals at high temperatures	Viscoplastic
Metals at high strain rates	Dynamic-viscoplastic
Polymers(below $T_g$ )*	Linear elastic/viscoelastic
Polymers (above $T_g$ )*	Viscoelastic
Monolithic ceramics	Linear elastic
Ceramic composites	Linear elastic
Ceramics at high temperatures	Viscoplastic

Note: Temperature is ambient unless otherwise specified.

\* $T_g$ —Glass transition temperature.

## Concluding Remarks

Fracture is a problem that society has faced for as long as there have been human-made structures. The problem may actually be worse today than in previous centuries, because more can go wrong in our complex technological society. Major airline crashes, for instance, would not be possible without modern aerospace technology.

Fortunately, advances in the field of fracture mechanics have helped to offset some of the potential dangers posed by increasing technological complexity. Our understanding of how materials fail and our ability to prevent such failures has increased considerably since World War II. Much remains to be learned, however, and existing knowledge of fracture mechanics is not always applied when appropriate.

Although catastrophic failures provide income for attorneys and consulting engineers, such events are detrimental to the economy as a whole. An economic study [Duga *et al.*, 1983] estimated the cost of fracture in the U.S. in 1978 at \$119 billion (in 1982 dollars), about 4% of the gross national product. Furthermore, this study estimated that the annual cost could be reduced by \$35 billion if current technology were applied and that further fracture mechanics research could reduce this figure by an additional \$28 billion.

## Defining Terms

**Damage tolerance:** A methodology that seeks to prevent catastrophic failures in components that experience time-dependent cracking. Fracture mechanics analyses are used in conjunction with nondestructive examination (NDE) to ensure that any flaws that may be present will not grow to a critical size prior to the next inspection.

**Energy release rate:** The rate of change in stored energy with respect to an increase in crack area. Energy release rate is a measure of the driving force for fracture. A crack will grow when the energy available for crack extension is greater than or equal to the energy required for crack extension. The latter quantity is a property of the material.

**Fracture mechanics:** An engineering discipline that quantifies the effect of cracks and crack-like flaws on material performance. Fracture mechanics analyses can predict both catastrophic failure and subcritical crack growth.

**Fracture toughness:** A measure of the ability of a material to resist crack propagation. The fracture toughness of a material can be quantified by various parameters, including a critical stress intensity factor,  $K_{IC}$ , and a critical energy release rate,  $G_c$ .

**Linear elastic fracture mechanics (LEFM):** A branch of fracture mechanics that applies to materials that obey Hooke's law. LEFM is not valid when the material experiences extensive nonlinear deformation such as yielding.

**Nondestructive examination (NDE):** A technology that can be used to characterize a material without altering its properties or destroying a sample. It is an indispensable tool for fracture mechanics analysis because NDE is capable of detecting and sizing crack-like flaws.

**Nonlinear fracture mechanics:** An extension of fracture mechanics theory to materials that experience nonlinear behavior such as yielding.

## References

- Bannerman, D. B. and Young, R. T. 1946. Some improvements resulting from studies of welded Ship failures. *Welding J.* 25:xx–xx.
- Duga, J. J., Fisher, W. H., Buxbaum, R. W., Rosenfield, A. R., Burh, A. R., Honton, E. J., and McM The Economic Effects of Fracture in the United States. NBS Special Publication 647-2. United States Department of Commerce, Washington, DC.
- Griffith, A. A. 1920. The phenomena of rupture and flow in solids. *Philos. Trans.*, Series A. 221:163–198.
- Irwin, G. R. 1956. Onset of fast crack propagation in high strength steel and aluminum alloys. *Sagar* –305.
- Irwin, G. R. 1957. Analysis of stresses and strains near the end of a crack traversing a plate. *J. Appl.* –364.
- Love, A. E. H. 1944. *A Treatise on The Mathematical Theory of Elasticity*. Dover, New York.
- Paris, P. C. and Erdogan, F. 1960. A critical analysis of crack propagation laws. *J. Basic Eng.* 85:521–534.
- Shank, M. E. 1953. *A Critical Review of Brittle Failure in Carbon Plate Steel Structures Other than* –National Research Council, Washington, DC.

## Further Information

- Anderson, T. L. 1991. *Fracture Mechanics: Fundamentals and Applications*. CRC Press, Boca Raton, FL.
- Broek, D. 1986. *Elementary Engineering Fracture Mechanics*, 4th ed. Martinus Nijhoff, Dordrecht, The Netherlands.
- Ewalds, H. L. and Wanhill, R. J. H. 1984. *Fracture Mechanics*. Edward Arnold, London.
- Hertzberg, R. W. 1989. *Deformation and Fracture Mechanics of Engineering Materials*, 3rd ed. John Wiley & Sons, New York.
- International Journal of Fracture*, published bimonthly by Kluwer Academic, Dordrecht, The Netherlands.
- Kanninen, M. F. and Poplar, C. H. 1985. *Advanced Fracture Mechanics*. Oxford University Press, Oxford.
- Rolfe, S. T. and Barsom, J. M. 1989. *Fracture and Fatigue Control in Structures*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.

Karnopp, B. "Dynamics of Particles: Kinematics and Kinetics"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Dynamics of Particles: Kinematics and Kinetics

---

## 12.1 Dynamics of Particles

Cartesian Coordinates • Natural (Path) Coordinates • Cylindrical Coordinates • Spherical Coordinates • Kinematics of Relative Motion

## 12.2 Newton's Second Law

## 12.3 Moment of Momentum Relations

## 12.4 Momentum Integrals of Newton's Second Law

Impulse-Momentum and Angular Impulse-Moment of Momentum Relations

## 12.5 Work-Energy Integral of Newton's Second Law

The Work-Energy Relation for a Conservative Force

## 12.6 Conclusion

**Bruce Karnopp**

*University of Michigan*

## 12.1 Dynamics of Particles

---

The dynamics of particles consists of five main parts:

1. **Kinematics** of a point (the geometry of a point moving through space)
2. Newton's second law
3. Moment of momentum equation
4. Momentum integrals of Newton's second law
5. Work-energy integral of Newton's second law

The concept of a **particle** is an abstraction or model of the actual physical situation. The moon in motion about the earth might be modeled as a mass point. In fact, the motion of any finite body in which the rotation effects are not important can properly be described as a particle or point mass.

Although it is possible to derive all the fundamental equations in a purely vector format, in order to describe any particular problem in dynamics, it is crucial that a specific coordinate system be employed. The coordinate systems which will be considered in this chapter are the following: (a) Cartesian coordinates, (b) natural (path) coordinates, (c) cylindrical coordinates, (d) spherical coordinates, (e) relative motion.

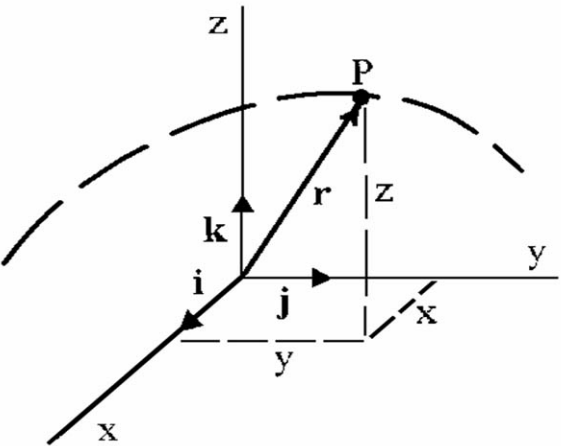
The fundamental equation that is encountered in dynamics is Newton's second law,  $\mathbf{F} = m\mathbf{a}$ , where  $\mathbf{F}$  is the total force acting on a particle and  $\mathbf{a}$  is the resulting acceleration. Thus the geometric problem of dynamics consists of finding the **position**,  $\mathbf{r}$ , the **velocity**,  $\mathbf{v}$ , and the **acceleration**,  $\mathbf{a}$ , of a point mass.

In order to use any coordinate system, the equations for the position  $\mathbf{r}$ , the velocity  $\mathbf{v}$ , and the acceleration  $\mathbf{a}$  as expressed in that coordinate system must be known. In order to achieve these results, the derivatives of the unit vectors of the coordinate system must be determined.

## Cartesian Coordinates

Consider the path of point  $P$  with respect to the Cartesian coordinate system shown in Fig. 12.1. The position, velocity, and acceleration are expressed in Table 12.1.

**Figure 12.1** Cartesian coordinates.



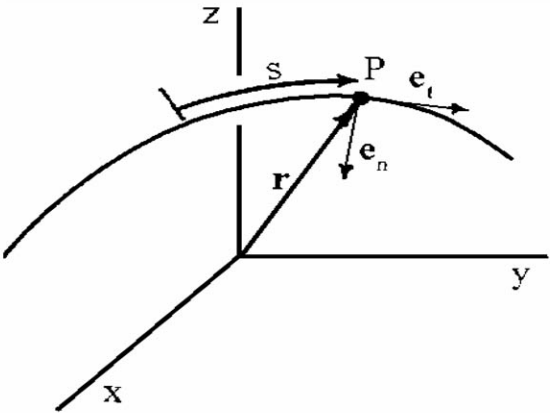
**Table 12.1** Equations of Cartesian Coordinates

$\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$
$\mathbf{v} = \dot{x}\mathbf{i} + \dot{y}\mathbf{j} + \dot{z}\mathbf{k}$
$\mathbf{a} = \ddot{x}\mathbf{i} + \ddot{y}\mathbf{j} + \ddot{z}\mathbf{k}$

## Natural (Path) Coordinates

Natural or path coordinates are useful to understand the intrinsic nature of the velocity and acceleration vectors. Natural coordinates are defined by the actual trajectory of the point  $P$  as it moves through space. Consider the trajectory as shown in Fig. 12.2.

**Figure 12.2** Natural coordinates.



The distance or arc length along the path (from some convenient starting position) is denoted by  $s$ . The velocity and acceleration of  $P$  are defined in terms of the path characteristics: the unit vector  $\mathbf{e}_t$  tangent to the path, the unit vector  $\mathbf{e}_n$  normal to the path, the radius of curvature,  $R$ , and the derivatives of the arc length with respect to time,  $\dot{s}$  and  $\ddot{s}$ . The quantity  $R$  is the radius of curvature, and  $\tau$  is the torsion of the curve. Table 12.2 lists the natural coordinate equations. Struik [1961] contains a proof of the derivatives of the unit vectors.

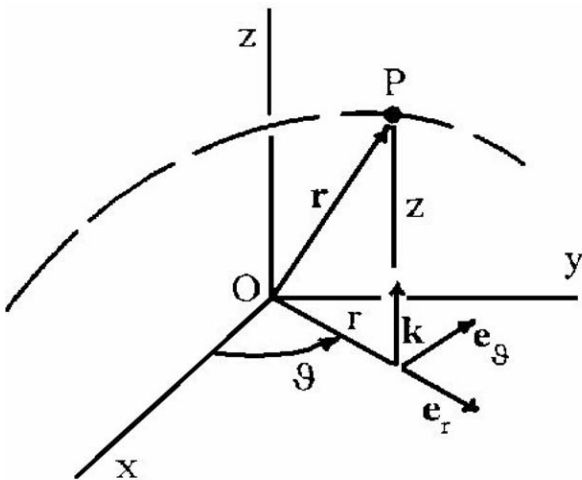
**Table 12.2** Equations of Natural Coordinates

Velocity and Acceleration	Derivatives of Unit Vectors
$\mathbf{v} = \dot{s}\mathbf{e}_t$	$\frac{d\mathbf{e}_t}{dt} = \frac{\dot{s}}{R}\mathbf{e}_n$
$\mathbf{a} = \ddot{s}\mathbf{e}_t + \frac{\dot{s}^2}{R}\mathbf{e}_n$	$\frac{d\mathbf{e}_n}{dt} = -\frac{\dot{s}}{R}\mathbf{e}_t + \dot{s}\tau\mathbf{e}_b$
	$\frac{d\mathbf{e}_b}{dt} = -\dot{s}\tau\mathbf{e}_n$

## Cylindrical Coordinates

Cylindrical coordinates are used when there is some symmetry about a line. If this line is taken to be the  $z$  axis, the coordinates appear as in Fig. 12.3. The parameters of cylindrical coordinates are introduced by dropping a line from the point  $P$  to the  $xy$  plane. The distance from the origin  $O$  to the intersection in the  $xy$  plane is denoted by the scalar  $r$ . Finally, the angle between the  $x$  axis and the line from  $O$  to the intersection is  $\vartheta$ . Thus the parameters that define cylindrical coordinates are  $\{r, \vartheta, z\}$ , and the unit vectors are  $\{\mathbf{e}_r, \mathbf{e}_\vartheta, \mathbf{k}\}$ . See Table 12.3.

**Figure 12.3** Cylindrical coordinates.

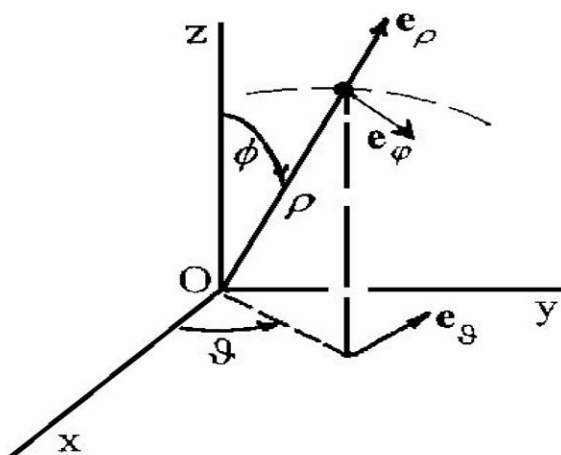


**Table 12.3** Equations of Cylindrical Coordinates

Position, Velocity, and Acceleration	Derivatives of Unit Vectors
$\mathbf{r} = r\mathbf{e}_r + z\mathbf{k}$	$\dot{\mathbf{e}}_r = \dot{\vartheta}\mathbf{e}_\vartheta$
$\mathbf{v} = \dot{r}\mathbf{e}_r + r\dot{\vartheta}\mathbf{e}_\vartheta + \dot{z}\mathbf{k}$	$\dot{\mathbf{e}}_\vartheta = -\dot{\vartheta}\mathbf{e}_r$
$\mathbf{a} = (\ddot{r} - r\dot{\vartheta}^2)\mathbf{e}_r + (r\ddot{\vartheta} + 2\dot{r}\dot{\vartheta})\mathbf{e}_\vartheta + \ddot{z}\mathbf{k}$	$\dot{\mathbf{k}} = 0$

## Spherical Coordinates

Spherical coordinates are particularly useful in problems with symmetry about a point. The coordinates are defined by the three parameters  $\rho$ ,  $\phi$ , and  $\vartheta$  and the corresponding unit vectors  $\mathbf{e}_\rho$ ,  $\mathbf{e}_\phi$ , and  $\mathbf{e}_\vartheta$ . Refer to [Fig. 12.4](#) and [Table 12.4](#).

**Figure 12.4** Spherical coordinates.**Table 12.4** Equations of Spherical Coordinates

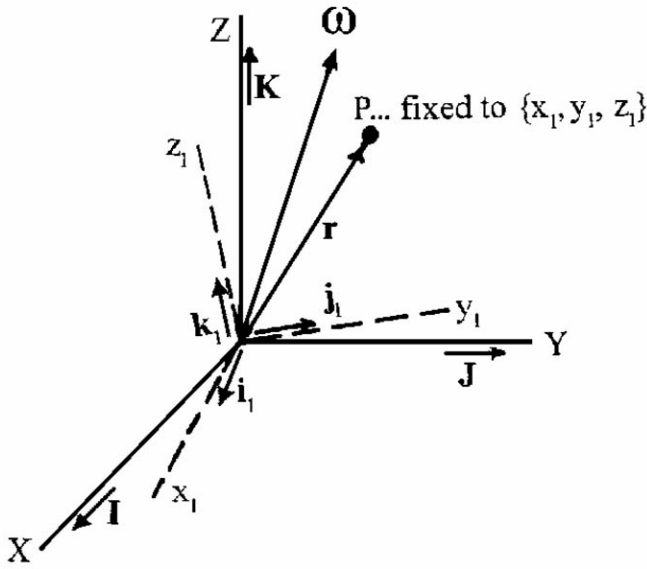
Position, Velocity, and Acceleration	Derivatives of Unit Vectors
$\mathbf{r} = \rho\mathbf{e}_\rho$	$\dot{\mathbf{e}}_\rho = \dot{\phi}\mathbf{e}_\phi + \dot{\vartheta}\sin\phi\mathbf{e}_\vartheta$
$\mathbf{v} = \dot{\rho}\mathbf{e}_\rho + \rho\dot{\phi}\mathbf{e}_\phi + \rho\dot{\vartheta}\sin\phi\mathbf{e}_\vartheta$	$\dot{\mathbf{e}}_\phi = -\dot{\phi}\mathbf{e}_\rho + \dot{\vartheta}\cos\phi\mathbf{e}_\vartheta$
$\mathbf{a} = (\ddot{\rho} - \rho\dot{\phi}^2 - \rho\dot{\vartheta}^2\sin^2\phi)\mathbf{e}_\rho$	$\dot{\mathbf{e}}_\vartheta = -\dot{\vartheta}\sin\phi\mathbf{e}_\rho - \dot{\vartheta}\cos\phi\mathbf{e}_\phi$
$+ (2\dot{\rho}\dot{\phi} + \rho\ddot{\phi} - \rho\dot{\vartheta}^2\sin\phi\cos\phi)\mathbf{e}_\phi$	
$+ (2\dot{\rho}\dot{\vartheta}\sin\phi + 2\rho\dot{\phi}\dot{\vartheta}\cos\phi + \rho\ddot{\vartheta}\sin\phi)\mathbf{e}_\vartheta$	



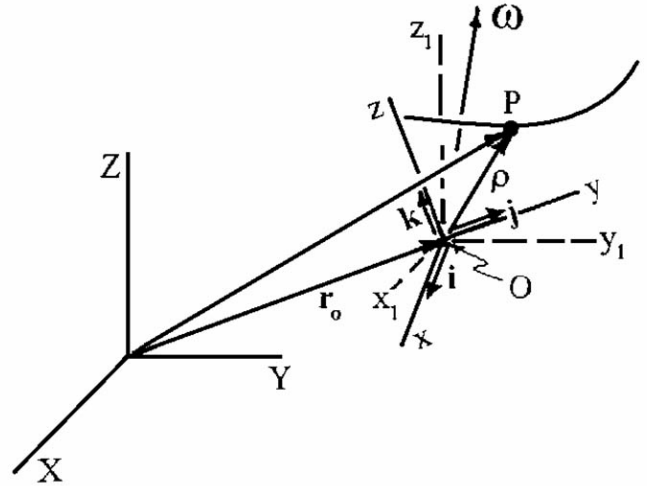
## Kinematics of Relative Motion

The equations of relative motion are used when it is convenient to refer the motion to a coordinate system that is in motion. In general, such a coordinate system cannot be used to write the equations of dynamics since it will not be an inertial reference frame. The crucial concept in this regard is the **angular velocity** vector,  $\omega$ . Consider [Fig. 12.5](#).

**Figure 12.5** Angular velocity.



**Figure 12.6** Relative motion.



The angular velocity vector  $\omega$  is introduced through the equation

$$\mathbf{v} = \omega \times \mathbf{r} \quad (12.1)$$

for any point  $P$  that is embedded in the moving frame  $\{x_1, y_1, z_1\}$ . Then

$$\frac{d\rho}{dt} = \frac{dx}{dt}\mathbf{i} + \frac{dy}{dt}\mathbf{j} + \frac{dz}{dt}\mathbf{k} + x\omega \times \mathbf{i} + y\omega \times \mathbf{j} + z\omega \times \mathbf{k}$$

or

$$\frac{d\rho}{dt} = \underbrace{\frac{dx}{dt}\mathbf{i} + \frac{dy}{dt}\mathbf{j} + \frac{dz}{dt}\mathbf{k}}_{\delta\rho/\delta t} + \underbrace{\omega \times (x\mathbf{i} + y\mathbf{j} + z\mathbf{k})}_{\omega \times \rho}$$

The first block of terms is the velocity as seen from the moving frame—that is, the velocity as it would appear to an observer whose feet are firmly planted in the moving frame  $\{x, y, z\}$ . It is convenient to denote this as  $\delta\rho/\delta t$ . The second block of terms is just  $\omega \times \rho$ . Thus,

$$\frac{d\rho}{dt} = \frac{\delta\rho}{\delta t} + \omega \times \rho \quad (12.2)$$

Equation (12.2) gives an operator equation for computing time derivatives with respect to a fixed or moving frame:

Equation (12.2) gives an operator equation for computing time derivatives with respect to a fixed or moving frame:

$$\frac{d}{dt} = \frac{\delta}{\delta t} + \omega \times \quad (12.3)$$

Now suppose the origin of the moving frame has a motion. Then if  $\mathbf{v}_o$  is the velocity of the moving origin (see Fig. 12.6),

$$\mathbf{v} = \mathbf{v}_o + \frac{\delta \rho}{\delta t} + \omega \times \rho \quad (12.4)$$

In order to interpret this, it is instructive to rearrange the terms:

$$\mathbf{v} = \underbrace{(\mathbf{v}_o + \omega \times \rho)}_{\text{Convective velocity}} + \underbrace{\frac{\delta \rho}{\delta t}}_{\text{Relative velocity}}$$

The *relative velocity* is that which is seen by an observer fixed to the moving frame. The *convective velocity* is the velocity of a fixed point that instantaneously shares the position of the moving point.

This process is repeated to determine the acceleration. That is, the operator of Eq. (12.3) is applied to Eq. (12.4) to get  $\mathbf{a} = d\mathbf{v}/dt$ . The result is arranged as follows:

$$\mathbf{a} = \underbrace{[\mathbf{a}_o + \omega \times (\omega \times \rho) + \underline{\omega} \times \rho]}_{\text{Convective acceleration}} + \underbrace{\frac{\delta^2 \rho}{\delta t^2}}_{\text{Relative acceleration}} + \underbrace{2\omega \times \frac{\delta \rho}{\delta t}}_{\text{Coriolis acceleration}} \quad (12.5)$$

Again, the *relative acceleration* is that which a moving observer in  $\{x, y, z\}$  would see. The *convective acceleration* is the acceleration of the fixed point of  $\{x, y, z\}$  that shares the instantaneous position of the moving point under consideration.

The equations of position, velocity, and acceleration are summarized in Table 12.5.

**Table 12.5** Equations of Relative Motion

$\mathbf{r} = \mathbf{r}_o + \rho$
$\mathbf{v} = (\mathbf{v}_o + \omega \times \rho) + \frac{\delta \rho}{\delta t}$
$\mathbf{a} = [\mathbf{a}_o + \omega \times (\omega \times \rho) + \underline{\omega} \times \rho] + \frac{\delta^2 \rho}{\delta t^2} + 2\omega \times \frac{\delta \rho}{\delta t}$

## 12.2 Newton's Second Law

In order to write Newton's second law for a particle,  $m$ ,

$$\mathbf{F} = m\mathbf{a} \quad (12.6)$$

the terms of the equation must be evaluated:

1. The force,  $\mathbf{F}$ , is obtained from a free-body diagram of the particle.
2. The mass,  $m$ , can be obtained from the weight of the particle:

$$\text{Weight} = mg \quad (12.7)$$

3. The acceleration is written in some convenient coordinate system (from the equations in section 12.1).

Any equation must ultimately be expressed in some unit system. The fundamental units of dynamics are force, mass, length, and time. The units for these quantities are shown in [Table 12.6](#). Conversion of units is shown in [Table 12.7](#).

**Table 12.6** Unit Systems Used in Dynamics

Unit System	Type	Force	Mass	Length	Time
English (large)	Gravitational	Pound (lb)	Slug $\text{lb} \cdot \text{s}^2/\text{ft}$	Foot (ft)	Second (s)
English (small)	Gravitational	Pound (lb)	$\text{lb} \cdot \text{s}^2/\text{in.}$	Inch (in.)	Second (s)
MKS—metric	Absolute	Newton (N)	Kilogram (kg)	Meter (m)	Second (s)
CGS—metric	Absolute	Dyne (dyn)	gram (g)	Centimeter (cm)	Second (s)
Metric—large	Gravitational	Kilogram (kg)	$\text{kg} \cdot \text{s}^2/\text{m}$	Meter (m)	Second (s)
Metric—small	Gravitational	Gram (g)	$\text{g} \cdot \text{s}^2/\text{cm}$	Centimeter (cm)	Second (s)

**Table 12.7** Conversion of Units

Force units:	1.0 lb	= 4.448 N
	1.0 lb	= $4.448 \cdot 10^5$ dyn
	1.0 lb	= 0.4536 (kgf)
	1.0 lb	= $4.536 \cdot 10^2$ (g force)
Length units:	1.0 in.	= 0.083 33ft
	1.0 in.	= 2.542.54 cm
	1.0 in.	= 0.0254 m
	1.0 ft	= 12 in.
	1.0 ft	= 30.48 cm
	1.0 ft	= 0.3048 m
Mass units:	$1.0 \text{ lb} \cdot \text{s}^2/\text{in.}$	= $12 \text{ lb} \cdot \text{s}^2/\text{ft}$
	$1.0 \text{ lb} \cdot \text{s}^2/\text{in.}$	= 1.2162 kg
	$1.0 \text{ lb} \cdot \text{s}^2/\text{in.}$	= $1.2162 \cdot 10^3$ g
1.0 slug =	$1.0 \text{ lb} \cdot \text{s}^2/\text{ft}$	= $0.083 \text{ 33 } \text{lb} \cdot \text{s}^2/\text{in.}$
	$1.0 \text{ lb} \cdot \text{s}^2/\text{ft}$	= 14.594 kg
	$1.0 \text{ lb} \cdot \text{s}^2/\text{ft}$	= $1.4594 \cdot 10^4$ g

## 12.3 Moment of Momentum Relations

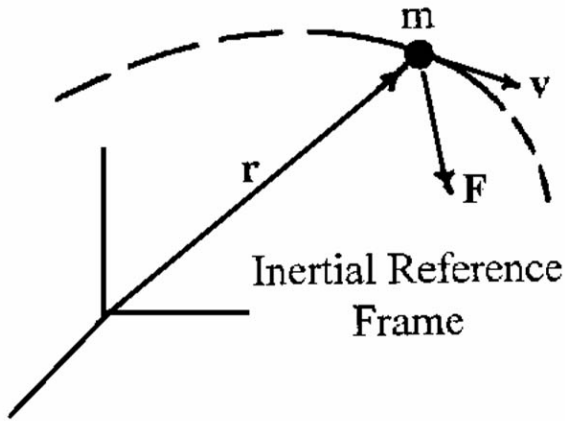
---

The moment of a force is determined by the (vector) *cross product*:

$$\mathbf{M}_o = \mathbf{r} \times \mathbf{F} \quad (12.8)$$

See Fig. 12.7.

**Figure 12.7** Moment of momentum and moment of force.



The *linear momentum* of  $m$  is

$$\mathbf{p} = m\mathbf{v} \quad (12.9)$$

and the *moment of momentum* (sometimes called the *angular momentum*) is

$$\mathbf{h}_o = \mathbf{r} \times \mathbf{p} = \mathbf{r} \times m\mathbf{v} \quad (12.10)$$

Computing the time derivative of  $\mathbf{h}$  gives

$$\frac{d\mathbf{h}_o}{dt} = \frac{d\mathbf{r}}{dt} \times m\mathbf{v} + \mathbf{r} \times m\mathbf{a}$$

The first term is  $\mathbf{v} \times m\mathbf{v}$ . This is the cross product of two vectors in the same direction. Thus this term is zero. From Eqs. (12.6) and (12.8), the remaining term is the moment of the force,  $\mathbf{F}$ , about the point  $O$ . Thus:

$$\mathbf{M}_o = \frac{d\mathbf{h}_o}{dt} \quad (12.11)$$

## 12.4 Momentum Integrals of Newton's Second Law

---

Newton's second law, Eq. (12.6), can be integrated over time or space. When the former is done, the result is called an *impulse* or an *angular impulse*. When the integration is performed over space, the result is the *work*. This will be demonstrated later.

### Impulse-Momentum and Angular Impulse-Moment of Momentum Relations

Recall Newton's second law,  $\mathbf{F} = m\mathbf{a}$ . Suppose we write  $\mathbf{a} = d\mathbf{v}/dt$ . Then

$$\int_{t_o}^{t_1} \mathbf{F} dt = m \int_{t_o}^{t_1} \frac{d\mathbf{v}}{dt} dt = m[\mathbf{v}(t_1) - \mathbf{v}(t_o)] \quad (12.12)$$

Equation (12.12) is called the *impulse change of linear momentum theorem*.

Similarly, taking Eq. (12.11) as the basis of the time integration gives

$$\int_{t_o}^{t_1} \mathbf{M}_o dt = \int_{t_o}^{t_1} \frac{d\mathbf{h}_o}{dt} dt = \mathbf{h}_o(t_1) - \mathbf{h}_o(t_o) \quad (12.13)$$

Equation (12.13) is called the *angular impulse change of angular momentum theorem*. Equations (12.12) and (12.13) are particularly interesting when the left-hand side is zero. Then we say that linear momentum is conserved or that the moment of momentum (angular momentum) is conserved.

Two important examples that utilize these conservation laws are collision problems [[Karnopp, 1974](#)] and central force motion problems [[Goldstein, 1959](#)].

## 12.5 Work-Energy Integral of Newton's Second Law

---

In deriving the momentum laws, Newton's second law is integrated over time. In the work-energy relation, the integration takes place over space. Recall Newton's second law,  $\mathbf{F} = m\mathbf{a} = m(d\mathbf{v}/dt)$ . An instantaneous quantity is the *power* of the force  $\mathbf{F}$ :

$$P = \mathbf{F} \cdot \mathbf{v} \quad (12.14)$$

The **power**,  $P$ , is a scalar quantity. The units of power are listed in [Table 12.8](#).

**Table 12.8** Units of Power

English:	ft · lb/s in. · lb/s 1.0 horsepower = 550 ft · lb/s
Metric:	N · m/s 1.0 watt = 1.0 N · m/s
Conversion:	1.0 N · m/s = 0.7376 ft · lb/s 1.0 ft · lb/s = 1.3557 N · m/s 1.0 horsepower = 746 watts 1.0 watt = 1.34048 · 10 <sup>-3</sup> hp

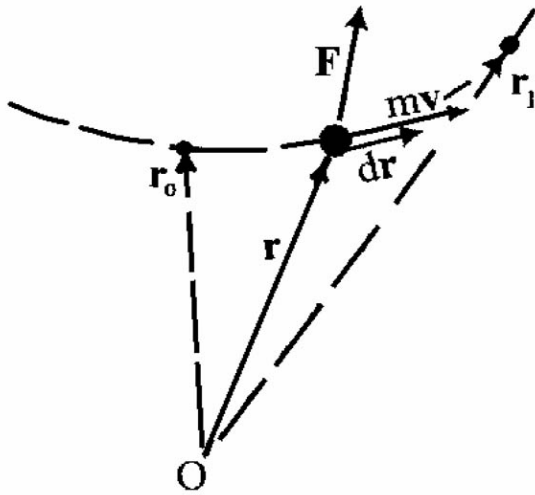
The *work* of a force is the time integral of the power of the force. Work is also a scalar quantity:

$$W = \int_{t_o}^{t_1} P \, dt \quad (12.15)$$

With Eq. (12.14), and recalling that  $\mathbf{v} = d\mathbf{r}/dt$  the work  $W$  (see Fig. 12.8) becomes

$$W = \int_{\mathbf{r}_o}^{\mathbf{r}_1} \mathbf{F} \cdot d\mathbf{r} \quad (12.16)$$

**Figure 12.8** Work of a force.



Equation (12.16) is what is called a *path* or *line integral*. That is, the value of the work is dependent, in general, on the particular path that is traversed between positions  $\mathbf{r}_o$  and  $\mathbf{r}_1$ .

Recall from Eqs. (12.6) and (12.14) that  $\mathbf{F} = m\mathbf{a} = m(d\mathbf{v}/dt)$  and  $P = \mathbf{F} \cdot \mathbf{v}$ . Inserting Eq. (12.13) into Eq. (12.14) gives

$$P = m \frac{d\mathbf{v}}{dt} \cdot \mathbf{v} \quad (12.16a)$$

Now consider that

$$\frac{1}{2} \frac{d}{dt} (m\mathbf{v} \cdot \mathbf{v}) = \frac{1}{2} \left[ m \frac{d\mathbf{v}}{dt} \cdot \mathbf{v} + m\mathbf{v} \cdot \frac{d\mathbf{v}}{dt} \right] = m \frac{d\mathbf{v}}{dt} \cdot \mathbf{v} \quad (12.16b)$$

Defining the *kinetic energy* of a particle to be

$$T = \frac{1}{2} m\mathbf{v} \cdot \mathbf{v} = \frac{1}{2} m\mathbf{v}^2 \quad (12.17)$$

Eqs. (12.16a) and (12.16b) give

$$P = \frac{d}{dt} T \quad (12.18)$$

That is, the power of the force  $\mathbf{F}$  equals the time rate of change of the kinetic energy  $T$ .

Finally, from Eq. (12.18), the work of  $\mathbf{F}$  in moving the particle from  $\mathbf{r}_o$  to  $\mathbf{r}_1$  equals the change in kinetic energy between  $\mathbf{r}_o$  and  $\mathbf{r}_1$ . The work-energy theorem is derived in [Table 12.9](#).

**Table 12.9** The Work-Energy Theorem

The work of a force $\mathbf{F}$ is	$W = \int_{\mathbf{r}_o}^{\mathbf{r}_1} \mathbf{F} \cdot d\mathbf{r}$
The kinetic energy of a particle is	$T = \frac{1}{2} m \mathbf{v} \cdot \mathbf{v} = \frac{1}{2} m v^2$
And the work-energy relation is	$W_{\mathbf{r}_o}^{\mathbf{r}_1} = T(\mathbf{r}_1) - T(\mathbf{r}_o)$

While Eq. (12.16) gives a way to compute the work of a force, there is a very special and important class of forces that give a very simple way of computing work. These are called *conservative forces*. A force is conservative if it can be derived from a *potential energy function* through differentiation.

$$\mathbf{F} = -\nabla V \quad (12.19)$$

where  $V$  is the potential energy function for  $\mathbf{F}$  and  $\nabla$  is the del operator. In Cartesian coordinates, Eq. (12.18) becomes

$$\mathbf{F} = -\frac{\partial V}{\partial x} \mathbf{i} - \frac{\partial V}{\partial y} \mathbf{j} - \frac{\partial V}{\partial z} \mathbf{k}$$

The general form for conservative forces, Eq. (12.19), is usually overly complex. Conservative forces are listed in [Table 12.10](#).

**Table 12.10** Conservative Forces

Force	Potential Energy
Gravity	$\mathbf{F} = -mg \mathbf{k}$ $V = mgz$
Universal gravitation*	$\mathbf{F} = -\frac{\gamma Mm}{r^2} \mathbf{e}_r$ $V = -\frac{\gamma Mm}{r}$
Spring force	$\mathbf{F} = -k\delta$ $V = \frac{1}{2} k\delta^2$

\* For motion about the earth  $\gamma M_e = 1.255 \cdot 10^3 \text{ mi}^3/\text{h}^2 = 5.2277 \cdot 10^3 \text{ km}^3/\text{h}^2$

## The Work-Energy Relation for a Conservative Force

Recall the equation for the work of a force,  $W = \int_{\mathbf{r}_o}^{\mathbf{r}_1} \mathbf{F} \cdot d\mathbf{r}$ . Suppose  $\mathbf{F}$  is conservative. Then, by

Eq. (12.19),  $\mathbf{F} = -\nabla V$ . And, finally, recall the equations of natural coordinates to write the expression for  $d\mathbf{r}$ :  $\mathbf{v} = (d\mathbf{r}/ds)(ds/dt) = (ds/dt)\mathbf{e}_t$ . Thus,

$$d\mathbf{r} = \frac{d\mathbf{r}}{ds} ds = \mathbf{e}_t ds$$

Finally, the work, by Eq. (12.16), becomes

$$W = \int_{r_o}^{r_1} -(\nabla V \cdot \mathbf{e}_t) ds$$

The term inside the parentheses is just the *directional derivative*  $dV/ds$ —that is, the derivative that is taken tangent to the path. Thus the work becomes

$$W = \int_{r_o}^{r_1} -\left(\frac{dV}{ds}\right) ds = \int_{r_o}^{r_1} -dV = -V(\mathbf{r}_1) + V(\mathbf{r}_o) \quad (12.20)$$

The crucial thing to note in Eq. (12.20) is that the work of a conservative force depends *only* on the end positions of the path. Thus the work-energy relation derived in [Table 12.9](#) becomes, in the case of conservative forces:

$$T(\mathbf{r}_1) + V(\mathbf{r}_1) = T(\mathbf{r}_o) + V(\mathbf{r}_o) = \text{constant}$$

For conservative and nonconservative forces:

$$[T(\mathbf{r}_o) + V(\mathbf{r}_o)] + W_{r_o}^{r_1} = [T(\mathbf{r}_1) + V(\mathbf{r}_1)]$$

where  $W_{r_o}^{r_1}$  is the work of the nonconservative forces.

Units of energy are shown in [Table 12.11](#).

**Table 12.11** Units of Energy

English:	ft · lb	
	in. · lb	
Metric:	N · m	= joule
	dyn · cm	= erg
Conversion		
1.0 joule =	1.0 N · m	= 10 <sup>7</sup> dyn · cm = 10 <sup>7</sup> erg
1.0 joule =	1.0 N · m	= 0.073 76 ft · lb
1.0 joule =	1.0 N · m	= 0.885 12 in. · lb
1.0 erg =	1.0 dyn · cm	= 7.376 · 10 <sup>-9</sup> ft · lb
1.0 erg =	1.0 dyn · cm	= 8.8512 · 10 <sup>-8</sup> in. · lb
	1.0 ft · lb	= 1.3557 N · m
	1.0 ft · lb	= 1.3557 · 10 <sup>7</sup> dyn · cm
	1.0 in. · lb	= 0.112 98 N · m
	1.0 in. · lb	= 0.112 98 · 10 <sup>7</sup> dyn · cm



The work-energy theorem is used when what is sought is the speed of a particle as a function of *position in space*. The impulse momentum theorems, on the other hand, will give the velocity as a function of *time*. Both relations are derived from Newton's second law and are called *first integrals*.

## 12.6 Conclusion

---

The notion of a mass point or particle forms the basis of Newtonian mechanics. Although many systems can be modeled as a point mass, others cannot. Rigid configurations of systems, deformable systems, and so forth all require more elaborate geometrical (kinematic) description. The kinetic equations (Newton's law, momentum, moment of momentum, etc.) must be expanded in these cases. Still, the equations for particle dynamics form the basis of these discussions.

### Defining Terms

**Acceleration:** The (vector) rate of change of velocity.

**Angular velocity:** The rate of change of orientation of a coordinate system.

**Kinematics:** The geometry of motion.

**Particle:** A point mass.

**Position:** The location of a point in space.

**Power:** The dot product of the force and the velocity.

**Velocity:** The (vector) rate of change of position.

### References

- Beer, F. P. and Johnston, E. R. 1984. *Vector Mechanics for Engineers: Statics and Dynamics*, 4th ed. McGraw-Hill, New York.
- Goldstein, H. 1959. *Classical Mechanics*. Addison-Wesley, Reading, MA.
- Hibbler, R. C. C. 1983. *Engineering Mechanics, Dynamics*, 3rd ed. Macmillan, New York.
- Karnopp, B. H. 1974. *Introduction to Dynamics*. Addison-Wesley, Reading, MA.
- Meriam, J. L. and Kraige, L. G. 1986. *Engineering Mechanics, Dynamics*, 2nd ed. John Wiley & Sons, New York.
- Struik, D. J. 1961. *Differential Geometry*. Addison-Wesley, Reading, MA.

### Further Information

- Synge, J. L. and Griffith, B. A. 1959. *Principles of Mechanics*. McGraw-Hill, New York.

Zeid, A.A., et. al. "Dynamics of Rigid Bodies: Kinematics and Kinetics"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Dynamics of Rigid Bodies: Kinematics and Kinetics

---

## 13.1 Kinematics of Rigid Bodies

Translation • Rotation • General Plane Motion: Euler Theorem • Instantaneous Center of Rotation in Plane Motion • Absolute and Relative Acceleration in Plane Motion • Space Motion

## 13.2 Kinetics of Rigid Bodies

Forces and Acceleration • Work and Energy

**Ashraf A. Zeid**

*Army High Performance Computing Research Center and Computer Sciences Corporation*

**R. R. Beck**

*U.S. Army Tank Automotive Research Development and Engineering Center*

## 13.1 Kinematics of Rigid Bodies

---

Kinematics is the study of the geometry of rigid body motion without reference to what causes the motion. **Kinematic analyses** are conducted to establish relationships between the position, **velocity**, and **acceleration** of rigid bodies or points on a rigid body.

The position and orientation of a body can be described by their distance from a perpendicular set of fixed axes called a *coordinate system*. The minimum number of independent or generalized coordinates needed to completely describe the position and orientation of a system of rigid bodies is equal to the number of *degrees of freedom* for the system. The number of *degrees of freedom* equals the number of nonindependent coordinates used to describe the position and orientation of each body of the system minus the number of constraints equations governing the system's motion. Therefore, the maximum number of independent coordinates needed to completely describe the position and orientation of a rigid body in space is six. Three independent equations are required to locate and describe the rigid body in translation with respect to time; the other three independent equations of motion are required to define its orientation and rotation in space with respect to time.

In general, the equations of motion of a rigid body are created relative to an inertial reference frame. The inertial reference frame is usually the rectangular set of Cartesian coordinate axes  $x, y, z$  with corresponding unit vectors  $i, j, k$ , as described previously.

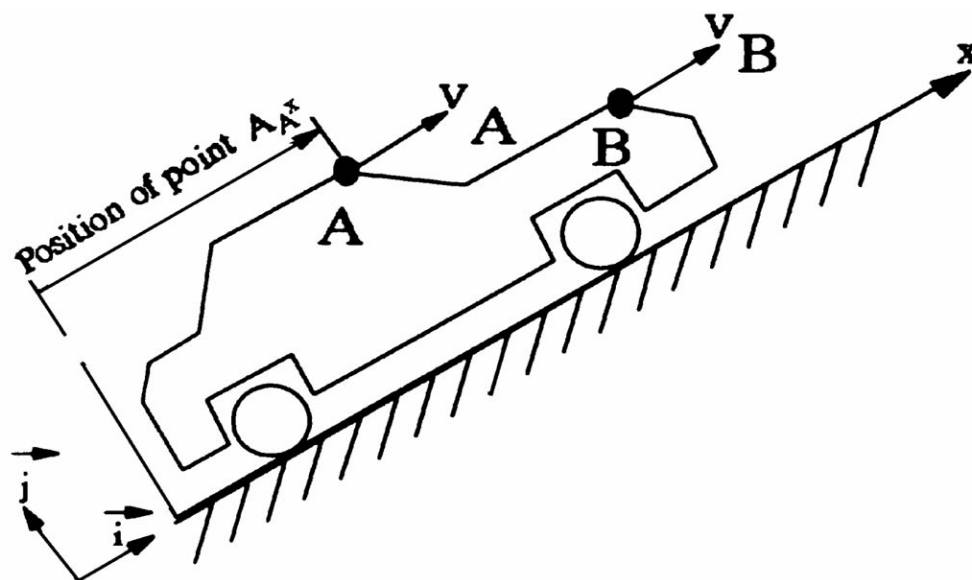
A rigid body is in *rectilinear translation* when a line that joins any two points on the body does

not rotate during motion. A rigid body is in *curvilinear translation* when all points of the body move on congruent curves. A *fixed-axis rotation* occurs when the line that connects any point on the body to the center of rotation rotates without any translation. When all points in a body move in parallel planes, the rigid body is in *general plane motion*. If no restriction is placed on the motion of the rigid body, it will move in *general space motion*. If the body is in general space motion and one of its points is pivoted, the body is in a *fixed-point rotation*.

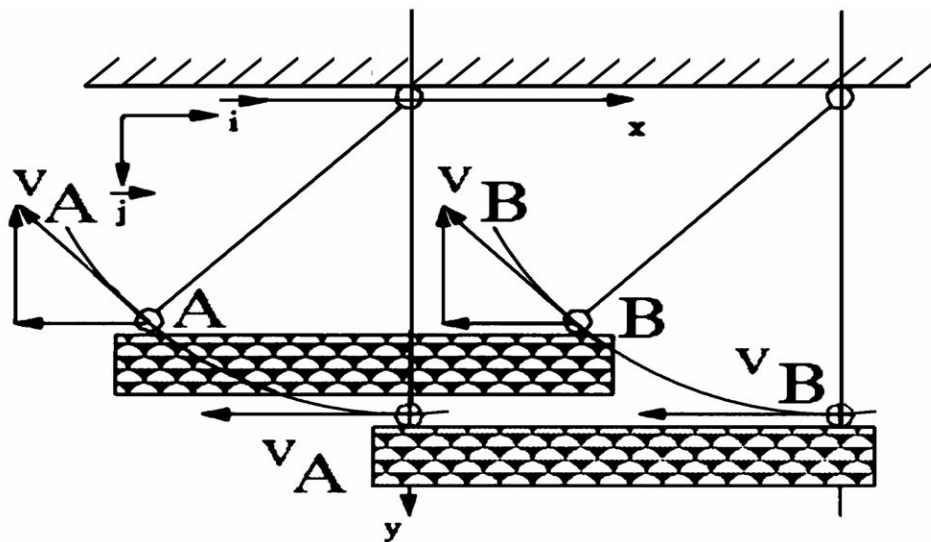
## Translation

All points on a rigid body in pure translation will have the same velocity and the same acceleration at any given instant. [Figures 13.1](#) and [13.2](#) show examples of two different types of translational motion and a possible choice of a fixed reference frame whose axes are denoted as  $x$  and  $y$  with corresponding unit vectors  $i$  and  $j$ , respectively.

**Fig. 13.1** Rectilinear translation.



**Fig. 13.2** Curvilinear translation.



When a body is undergoing rectilinear translation, as shown in Fig. 13.1, the velocities and accelerations of all points are identical in both magnitude and direction for all time.

$$\left. \begin{aligned} v_A &= v_B \\ a_A &= a_B \end{aligned} \right|_{\text{for all } t} \quad (13.1)$$

where  $\{A, B, \dots\}$  are arbitrary points on the body. In Fig. 13.2 the velocities of any two points  $A$  and  $B$  on the body are identical and parallel at any instant of time; however, unlike in rectilinear translation, the velocity and acceleration directions are not constant. For curvilinear translation the velocity equation holds at any instant of time but not necessarily throughout the entire motion:

$$\left. \begin{aligned} v_A &= v_B \\ a_A &= a_B \end{aligned} \right|_{t_1 \neq t_2} \quad (13.2)$$

## Rotation

The angular position of a body in pure rotation is completely defined by the angle between an arbitrary fixed reference line that passes through the center of rotation and any arbitrary line fixed to the body and passing also through the center of rotation, as shown in Fig. 13.3. The rotation angle  $\theta$  may be measured in degrees or radians, where

$$1 \text{ revolution} = 360 \text{ degrees} = 2\pi \text{ radians} \quad (13.3)$$

The rotation angular velocity  $\omega$  is defined as the rate of change of the angular position angle  $\theta$  with respect to time. It is expressed in radians per second (rps) or in revolutions per minute (rpm), as follows:

$$\omega = \frac{d\theta}{dt} \quad (13.4)$$

The rotational angular acceleration  $\alpha$  is the time rate of change of the angular velocity resulting in the following relationship:

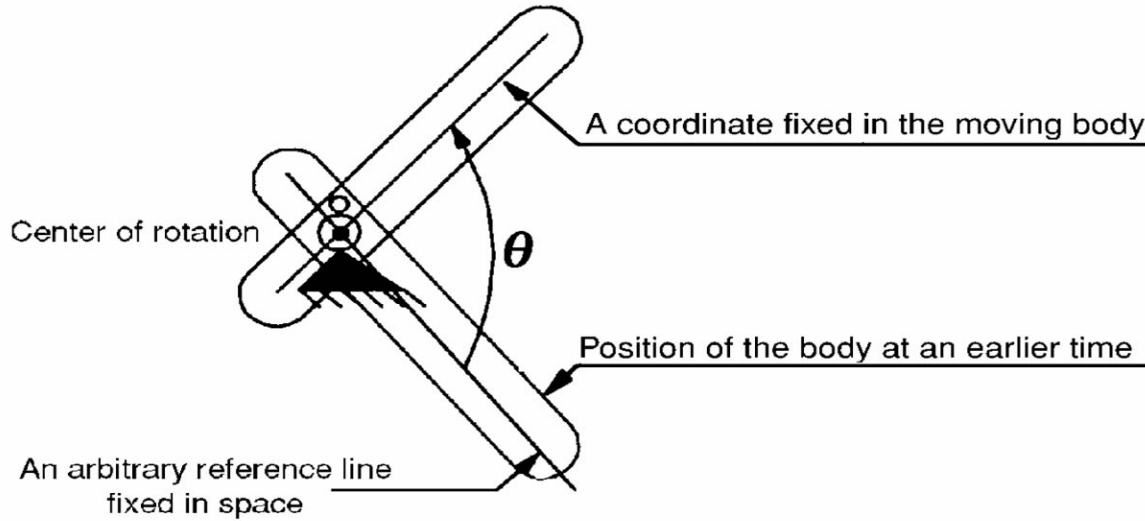
$$\alpha = \frac{d\omega}{dt} = \frac{d^2\theta}{dt^2} = \frac{d\omega}{d\theta} \frac{d\theta}{dt} = \omega \frac{d\omega}{d\theta} \quad (13.5)$$

In pure rotational motion, the relation between the *rotational position, velocity, and acceleration* are similar to pure translation. The angular velocity is the integral of the angular acceleration plus the initial velocity; the angular displacement is equal to the initial displacement added to the integral of the velocity. That is,

$$\begin{aligned} \omega &= \omega_0 + \alpha t \\ \theta &= \theta_0 + \omega t = \theta_0 + \omega_0 t + \frac{1}{2} \alpha t^2 \end{aligned} \quad (13.6)$$

In general, the angular velocity and angular acceleration are three-dimensional vectors whose three components are normally projected on a coordinate system fixed to the body that is translating and rotating in space, which is the general behavior of rigid bodies as discussed in later sections. For planar motion, two of the components of the angular velocity vector are equal to zero and the third component points always outward from the plane of motion.

**Fig. 13.3** A body in pure rotation.



Therefore, the position of any point  $B$  on a body in pure planar rotation is determined by the distance  $r_{B/A}$  of that point from the center of rotation  $A$  times the magnitude of the angle of rotation expressed in radians  $\theta$ . Thus the distance  $s$  that a point fixed on a rigid body travels during a rotation  $\theta$  is given by:

$$s = r_{B/A} \theta \quad (13.7)$$

Similarly, the linear velocity of that point will depend on the distance  $r_{B/A}$  and on the angular velocity  $\omega$  and will have a direction that is perpendicular to the line between the center of rotation and the point, as follows:

$$\vec{v} = \vec{\omega} \times \vec{r}_{B/A} \quad (13.8)$$

where  $\times$  indicates cross product. The angular acceleration of a point on a rigid body can be decomposed into a tangential and a normal component. The tangential component is the time rate of change of the linear velocity  $v$  and is in the direction of the linear velocity, namely, along the line perpendicular to the radius of rotation  $r_{B/A}$ .

$$\vec{a}_t = \frac{d\vec{v}}{dt} = \vec{\alpha} \times \vec{r}_{B/A} \quad (13.9)$$

The normal acceleration depends on the time rate of change of the velocity in the tangential direction and on the angle of displacement, which gives the equation

$$\vec{a}_n = \vec{\omega} \times \vec{\omega} \times \vec{r}_{B/A} \quad (13.10)$$

## General Plane Motion: Euler Theorem

General plane motion can be separated into a pure translation followed by a pure rotation about a point called the *center of rotation*. If we attach a coordinate system at point A, as shown in Fig. 13.4, the position of any point on the body can be described by the position vector of point A—namely,  $\vec{r}_A$ —added to the relative position of that point with respect to A—namely, the vector  $\vec{r}_{B/A}$ —all measured in the fixed coordinate system.

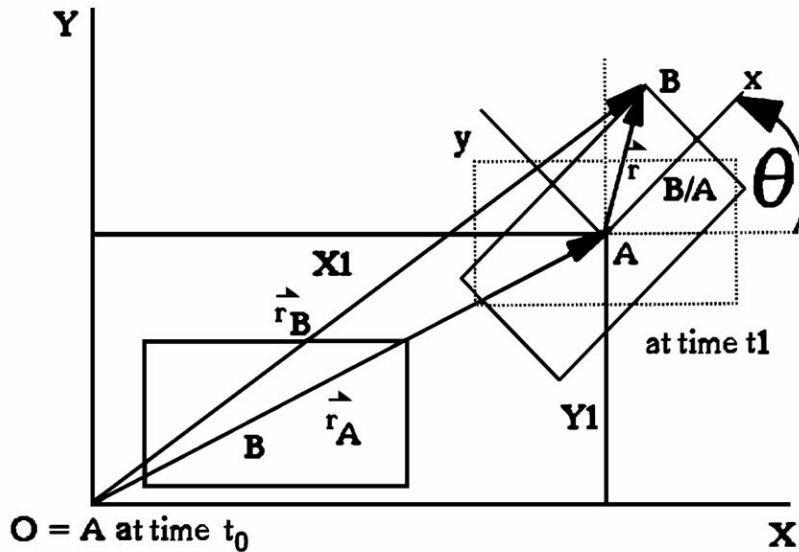
$$\vec{r}_B = \vec{r}_A + \vec{r}_{B/A} \quad (13.11)$$

Similarly, the velocity of a rigid body in general plane motion can be separated into a velocity due to pure translation  $\vec{v}_A$  together with a velocity due to pure rotation  $\vec{v}_{B/A}$ .

$$\vec{v}_B = \vec{v}_A + \vec{v}_{B/A} \quad (13.12)$$

where  $\vec{v}_{B/A} = \vec{\omega} \times \vec{r}_{B/A}$  and  $\vec{v}_A$  is the velocity vector of point A. The velocity vector  $\vec{v}_{B/A}$  is called the *relative velocity vector* of point B with respect to point A.

**Fig. 13.4** General motion of a rigid body in plane.



## Instantaneous Center of Rotation in Plane Motion

At any instant in time, a body in general plane motion has a point—which may be either outside or on the body—around which all points of the body appear to be rotating in pure rotation. This point, called *instantaneous center of rotation*, can be found if the velocity vector of any point on the body together with the angular velocity of the body are known. The instantaneous center of rotation will lie on a line perpendicular to the velocity vector and at a distance from that point that is equal to the magnitude of the velocity of the point divided by angular velocity  $\omega$  of the body.

Once the instantaneous center of rotation is found, the velocity of any point  $B$  on the body can be determined from the vector from that center to the point  $B$ ,  $\vec{r}_B$ , as follows:

$$\vec{v}_B = \vec{\omega} \times \vec{r}_B$$

The direction of the velocity vector will be perpendicular to the vector  $\vec{r}_B$ .

## Absolute and Relative Acceleration in Plane Motion

The angular acceleration of a point on a rigid body in plane motion also has a component due to translation and a component due to rotation; the latter component consists of a normal and a tangential component.

$$\vec{a}_B = \vec{a}_A + \vec{a}_{B/A}$$

$$\vec{a}_{B/A} = (\vec{a}_{B/A})_n + (\vec{a}_{B/A})_t \quad (13.13)$$

$$(\vec{a}_{B/A})_n = \omega \times \omega \times \vec{r}_{B/A}$$

$$(\vec{a}_{B/A})_t = \alpha \times \vec{r}_{B/A}$$

The acceleration of a point located by variable vector  $\vec{r}_{B/A}$  on a moving rigid body is given by the following relation:

$$\vec{a}_{B/A} = \vec{a}_A + \omega \times \omega \times \vec{r}_{B/A} + \alpha \times \vec{r}_{B/A} + 2\omega \times \frac{d\vec{r}_{B/A}}{dt} + \frac{d^2\vec{r}_{B/A}}{dt^2} \quad (13.14)$$

where the vector  $\vec{r}_{B/A}$  and its time derivative are measured in a fixed reference frame—namely, its components are  $[X_r, Y_r]$  as shown in Fig. 13.5. If the vector  $\vec{r}_{B/A}$  is known by its components in a body-fixed coordinate  $[x_r, y_r]$ , then they can be transformed to the inertial coordinates as follows:

$$\begin{aligned} X_r &= x_r \cos \theta - y_r \sin \theta \\ Y_r &= x_r \sin \theta + y_r \cos \theta \end{aligned} \quad (13.15)$$

This is a coordinate transformation and is orthogonal, that is, its transpose is equal to its inverse. In matrix form the transformation of coordinates in Eq. (13.15) can be written as follows:

$$\begin{bmatrix} X_r \\ Y_r \end{bmatrix} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x_r \\ y_r \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} x_r \\ y_r \end{bmatrix} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} X_r \\ Y_r \end{bmatrix} \quad (13.16)$$

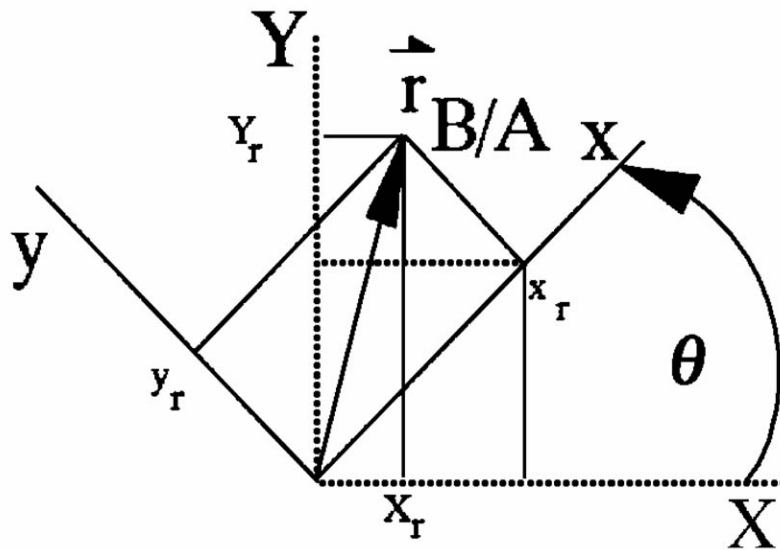
If we use the prime symbol to denote that the vector components are measured in a body-fixed coordinate, then Eq. (13.16) can be written in a more compact form as follows:

$$r_{B/A} = T_z r'_{B/A} \quad \text{and} \quad r'_{B/A} = T_z^T r_{B/A} \quad (13.17)$$

where  $T_z^T$  is the transpose of the rotation matrix around the  $z$  axis (which would point outward from the page).



**Fig. 13.5** Coordinate transformation in rotation.

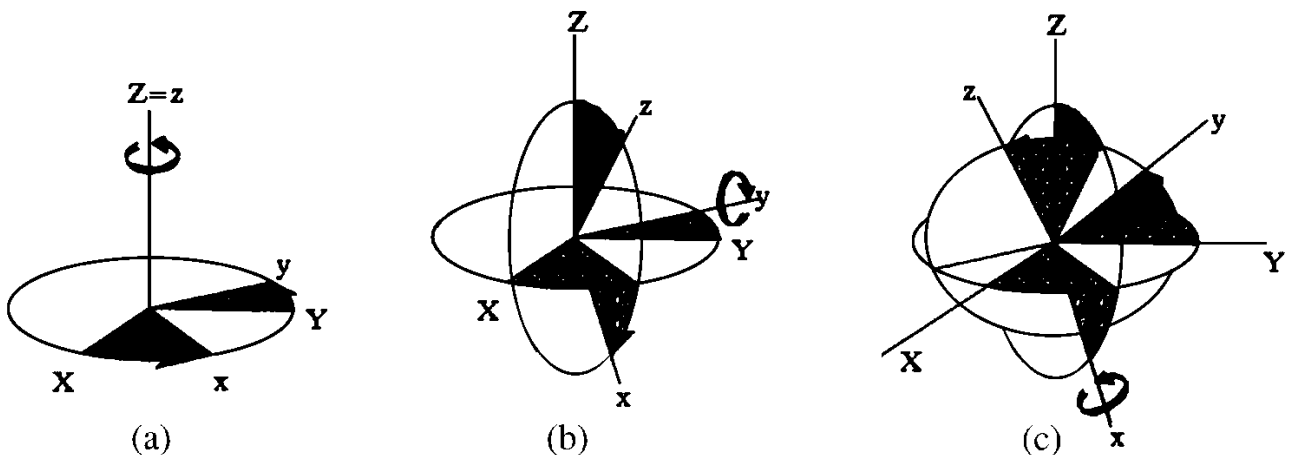


## Space Motion

Three angles, called *Euler angles*, may be used to describe the orientation of a rigid body in space. These angles describe three consecutive rotations around the three coordinates of the frame fixed in a moving body with respect to an inertial fixed frame. Twelve combinations of rotation sequences can be chosen; here we choose the rotation around the  $z$  axis,  $\psi$ , followed by a rotation around the body-fixed  $y$  axis,  $\theta$ , and finally a rotation around the body-fixed  $x$  axis,  $\phi$ .

Figure 13.6(c) shows the final position of a body which has a fixed coordinate system  $[x, y, z]$ . Originally, the body was oriented such that its fixed coordinate  $[x, y, z]$  corresponded to the inertial fixed coordinate system  $[X, Y, Z]$ . The body was then rotated by an angle  $\psi$  around  $z$ , as shown in Fig. 13.6(a), followed by a rotation of an angle  $\theta$  around  $y$ , as shown in Fig. 13.6(b), and finally by a rotation through an angle  $\phi$  around  $z$ .

**Fig. 13.6** Three consecutive rotations around the body-fixed axis (X;Y;Z).



If the components of a vector are known in the body-fixed coordinate system, then the components of that vector can be obtained in the inertial reference frame by multiplying the vector by a transformation matrix. This transformation is obtained from the sequential product of the three successive rotation matrices around axis  $z$ , then  $y$ , and then  $x$ , respectively. As an example, the transformation matrix for the rotation in the order shown in Fig. 13.6 is as follows:

$$T_{z,y,x} = \begin{bmatrix} \cos \theta \cos \psi & \cos \theta \sin \psi & -\sin \theta \\ -\cos \phi \sin \psi + \sin \phi \sin \theta \cos \psi & \cos \phi \cos \psi + \sin \phi \sin \theta \sin \psi & \sin \psi \cos \theta \\ \sin \phi \sin \psi + \cos \phi \sin \theta \cos \psi & -\sin \phi \cos \psi + \cos \phi \sin \theta \sin \psi & \cos \phi \cos \theta \end{bmatrix} \quad (13.18)$$

In order to transform any vector  $r'_{B/A}$  known by its components in a body-fixed coordinate system into the corresponding vector whose components are given in inertial fixed coordinates,  $r_{B/A}$ , and vice versa, the vector would be multiplied by the transformation matrix as follows:

$$r_{B/A} = T_{z,y,x} r'_{B/A} \quad \text{and} \quad r'_{B/A} = T_{z,y,x}^T r_{B/A} \quad (13.19)$$

where the superscript T denotes the transpose of the matrix. Because the transformation matrix is orthogonal its transpose is equal to its inverse, as shown by Eq. (13.19).

The time derivatives of the Euler angles can be obtained from the components of the angular rotation matrix  $\omega$  expressed in body coordinates. For the sequence of rotations shown in Fig. 13.6, the angular velocity vector can be expressed in terms of the rate of change of the Euler angles as follows. In Fig. 13.6(a) we have

$$\omega_x = 0$$

$$\omega_y = 0$$

$$\omega_z = \frac{d\psi}{dt}$$

In Fig. 13.6(b) we have

$$\omega_x = -\frac{d\psi}{dt} \sin \theta$$

$$\omega_y = \frac{d\theta}{dt}$$

$$\omega_z = \frac{d\psi}{dt} \cos \theta$$

Finally, in Fig 13.6(c) we have

$$\begin{aligned}
\omega_x &= -\frac{d\psi}{dt} \sin \theta + \frac{d\phi}{dt} \\
\omega_y &= \frac{d\psi}{dt} \cos \theta \sin \phi + \frac{d\theta}{dt} \cos \phi \\
\omega_z &= \frac{d\psi}{dt} \cos \theta \cos \phi - \frac{d\theta}{dt} \sin \phi \\
\frac{d\psi}{dt} &= (\omega_y \sin \phi + \omega_z \cos \phi) / \cos \theta \\
\frac{d\theta}{dt} &= \omega_y \cos \phi - \omega_z \sin \phi \\
\frac{d\phi}{dt} &= \omega_x + (\omega_y \sin \phi + \omega_z \cos \phi) \tan \theta
\end{aligned} \tag{13.20}$$

In vector form the above equation may be written as follows:

$$\omega = E \frac{d\Sigma}{dt} \quad \text{or} \quad \frac{d\Sigma}{dt} = E^{-1} \omega \tag{13.21}$$

where the matrix  $E$  is given by:

$$E = \begin{bmatrix} -\sin \theta & 0 & 1 \\ \cos \theta \sin \phi & \cos \phi & 0 \\ \cos \theta \cos \phi & -\sin \phi & 0 \end{bmatrix}, \quad \vec{\omega} = \begin{bmatrix} \omega_x \\ \omega_y \\ \omega_z \end{bmatrix},$$

$$\text{and} \quad E^{-1} = \begin{bmatrix} 0 & \frac{\sin \phi}{\cos \theta} & \frac{\cos \phi}{\cos \theta} \\ 0 & \cos \phi & -\sin \phi \\ 1 & \frac{\sin \phi \sin \theta}{\cos \theta} & \frac{\cos \phi \sin \theta}{\cos \theta} \end{bmatrix} \tag{13.22}$$

Note that the matrix  $E$  is not orthogonal, so its transpose is not equal to its inverse.

## 13.2 Kinetics of Rigid Bodies

---

### Forces and Acceleration

Kinetics is the study of the relation between the forces that act on a rigid body and the resulting acceleration, velocity, and motion as a function of the body mass and geometric shape. The acceleration of a rigid body is related to its mass and to the applied forces by D'Alembert's principle, which states that the external forces acting on a rigid body are equivalent to the effective forces of the various particles of the body.

In the case of a rigid body moving in a plane motion, the D'Alembert principle amounts to the vector equation  $\vec{F} = m\vec{a}$  together with the scalar equation of the moments  $M = I\alpha$ . In the particular case when a symmetric body is rotating around an axis that passes through its mass center—namely, centroidal rotation—the angular acceleration vector relates to the sum of moments by the equation  $M = I\alpha$ .

In general plane motion, the  $x$  and  $y$  components for the force vector, together with the moment equation, should be included in calculating the motion.

The **free-body diagram** is one of the essential tools for setting up the equations of motion that describe the kinetics of rigid bodies. It depicts the fundamental relation between the force vectors and the acceleration of a body by sketching the body together with all applied, reaction, and **D'Alembert force** and moment vectors drawn at the point where they are applied.

### Systems of Rigid Bodies in Planar Motion

Free-body diagrams can be used to set up, and in some cases to solve, problems that involve several rigid bodies interconnected by elements forcing them to a prescribed motion—for example, a motion that follows a curve or a surface. Such elements, called *kinematic joints*, can be rigid links with negligible masses or wires in tension, such as the ones used in pulley systems.

For planar motion three equations of motion are obtained by writing down the  $x$  and  $y$  components of the forces and acceleration with the equation of moments and the angular acceleration.

For a rigid body moving under a constraint, the free-body diagram is supplemented by a *kinematic analysis*, which provides the tangential and normal component of the acceleration. Rolling of a disc on a surface, which is a noncentroidal rotation, is an example of a constrained plane motion that belongs to this class of problems. The rolling can be with no sliding, with impending sliding, and with sliding. Rotation of gear pair and pulley also belong to this class of problems.

### Rotation of a Three-Dimensional Body about a Fixed Axis

If several bodies rotating each in their own plane are connected by a rigid shaft then each will exert a D'Alembert force,  $m\vec{a}$ , on the shaft. Their combined effect will be a vector force and a couple equal to the inertia  $I$  of the body times the angular acceleration  $\alpha$ .

If the body that rotates about a fixed axis is at rest and if the moments of the weights about the center of the rotating shaft is zero, we say that the system is *statically balanced*. When the body starts rotating, the moment due to D'Alembert forces,  $m\vec{a}$ , around the center of gravity of the system may not sum to zero; the system is *not dynamically balanced*. Rotating machinery strive to have their systems dynamically balanced to reduce the reaction forces at the bearings and consequently their wearing. Counterweights are added such that the total D'Alembert forces of the original bodies and the weights sum to zero.

In its most general case, the motion of a rigid body in space can be solved only through numerical integration, except for very few simple problems, such as **gyroscopic motion**.

## Work and Energy

The **kinetic energy** of a particle in translation is a scalar quantity measured in joules or ft-lb and can be simply defined as  $\frac{1}{2}mv^2$ .

The infinitesimal element of work,  $\Delta w$ , is defined as the product of the projection of the force vector on the path  $s$  of the body and the infinitesimal length  $ds$  of that path:  $(F \cos \beta) ds$ , where  $\beta$  is the angle that the force vector makes with the path  $ds$ .

The principle of work and energy states that the energy of a body is equal to the sum, over a certain displacement path, of the work done by all external forces that acted on the body and caused that displacement plus any initial kinetic energy that the body had at the beginning of the path.

For bodies in pure rotation the work of a couple is the product of the couple and the infinitesimal angle moved due to that couple. The summation of the work over all the angular displacement caused by that couple is the rotational energy and is also defined as  $\frac{1}{2}I\omega^2$ , where  $\omega$  is the angular velocity of the body.

The kinetic energy of a rigid body in general plane motion  $T$  is equal to the sum of kinetic energy in translation and the kinetic energy in rotation:

$$T = \frac{1}{2}(m\bar{v}^2 + \bar{I}\omega^2)$$

where  $\bar{v}$  is the velocity of the mass center  $G$  of the body and  $\bar{I}$  is the moment of inertia of the body about an axis through its mass center. This energy is identical to the kinetic rotational energy of the body if it is considered to be in pure rotation around its instantaneous center of rotation. In this case  $I$  would be the moment of inertia of the body around an axis that passes through the instantaneous center of rotation.

The principle of conservation of energy states that the sum of the potential and kinetic energy of a body acted upon by conservative forces—that is, nondissipative forces of friction or damping—remains constant during the time when these forces are applied.

Power is the product of the projection of the force vector on the velocity that resulted from this force. Power is measured in watt and horsepower units. The summation of power over a certain time interval is equal to the total energy stored in the body during that time.

### Kinetics of Rigid Bodies in Plane Motion: Impulse and Momentum for a Rigid Body

The principle of impulse and momentum for a rigid body states that the momenta of all the particles of a rigid body at time  $t_1$ , added to the impulses of external forces acting during the time interval from time  $t_1$  to time  $t_2$ , are equal to the system momenta at time  $t_2$ .

## Momentum of a Rigid Body in Plane Motion

**Translation, Rotation, and General Motion.** The momenta vector of a body in plane translation motion is the product of the mass and the velocity vector. For a rigid body in plane centroidal rotation, the linear momenta vector is equal to 0 since the mass center does not have any linear velocity. The sum of the couples of forces acting on the particle of that body gives the angular momentum, which is  $H_G = \bar{I}\omega$ .

In a general plane motion, the momentum is a vector with components along the  $x$ ,  $y$ , and  $\omega$  directions. The dynamic equations of motion of a rigid body in plane motion can be obtained from D'Alembert's principle as follows:

$$\begin{aligned}\frac{dm\vec{v}_x}{dt} &= (F_1)_x + (F_2)_x + \cdots \\ \frac{dm\vec{v}_y}{dt} &= (F_1)_y + (F_2)_y + \cdots \\ \frac{dI\alpha}{dt} &= M_1 + M_2 + \cdots + r_1 \times F_1 + r_2 \times F_2 + \cdots\end{aligned}\quad (13.23)$$

For a system of rigid bodies the linear momentum vector does not change in the absence of a resultant linear impulse. Similarly, the angular momentum vector does not change in the absence of an angular impulse.

**Space Motion.** The momentum vector of a rigid body moving in space has a linear component  $G$  and an angular component  $H$ . The linear component represents the D'Alembert principle as described by the following equations:

$$\begin{aligned}\frac{dm\vec{v}_x}{dt} &= (F_1)_x + (F_2)_x + \cdots \\ \frac{dm\vec{v}_y}{dt} &= (F_1)_y + (F_2)_y + \cdots \\ \frac{dm\vec{v}_z}{dt} &= (F_1)_z + (F_2)_z + \cdots\end{aligned}\quad (13.24)$$

where the velocity and the force vectors are expressed in their inertial  $[X, Y, Z]$  components. If these vectors are expressed in a body-fixed coordinate system, the time derivative should include the effect of the rotation vector, as in the case of the angular momentum. The angular momentum vector  $H$  is defined as follows:

$$\vec{H} = I\vec{\omega} \quad (13.25)$$

where  $H$  and  $\omega$  are expressed by their components in the body coordinate  $[x, y, z]$ . When a vector is expressed in a body coordinate its time derivative should include the effect of angular rotation. For this reason—and because normally the position vector of the point of application of a force  $F_i$  from a center of rotation  $A$ , denoted by  $\vec{r}_{Bi/A}$ , is known by its components in a body coordinate system—the equation stating that the time rate of change of the angular momentum is equal to the sum of the moments would be written as follows:

$$\frac{dI\omega}{dt} + \omega \times I\omega = T_{z,y,x}^T M_1 + T_{z,y,x}^T M_2 + \cdots + r'_{B1/A} \times T_{z,y,x}^T F_1 + r'_{B2/A} \times T_{z,y,x}^T F_2 + \cdots \quad (13.26)$$

where  $I$  is the matrix of inertia relative to a coordinate system fixed in the body and moving with it and the center of the coordinate system is located at point  $A$ . The forces and the moments are assumed to be known by their components in an inertial fixed coordinate system and  $\omega$  is the angular velocity of the body given by its components in the body-fixed coordinates.

If the body-fixed coordinate system is chosen along the principal axis of the body, then the equation of motion can be reduced to the Euler's equations as follows:

$$\begin{aligned} I_{xx}\dot{\omega}_x &= (I_{yy} - I_{zz})\omega_y\omega_z + \Sigma M_x \\ I_{yy}\dot{\omega}_y &= (I_{zz} - I_{xx})\omega_z\omega_x + \Sigma M_y \\ I_{zz}\dot{\omega}_z &= (I_{xx} - I_{yy})\omega_x\omega_y + \Sigma M_z \end{aligned} \quad (13.27)$$

In general, Eqs. (13.25) and (13.26) are solved by numerical integration, except for the cases where they are simplified, for example in gyroscopic motion. The Euler angles used in the transformation matrix  $T$  are obtained from the numerical integration of Eq. (13.20).

### Impulsive Motion and Eccentric Impact

The principle of conservation of momentum is useful in solving the problem of impacting bodies. If the colliding of two bodies is such that the collision point is on a line that joins their mass centers, then the collision is centroidal, the two bodies can be considered particles, and impulsive motion of particle dynamics can be used.

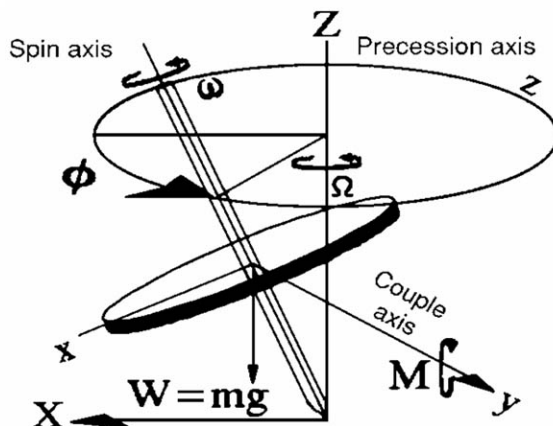
If the collision is noncentroidal, rotational motion will occur. In this case the projection of the velocity differential of the bodies' point of contact on the line normal to the contact surface after collision is equal to the same projection of the differential velocity prior to collision times the coefficient of restitution. This vector relation can be used to find the velocity after impact.

### Rotation Around a Fixed Point and Gyroscopic Motion

When a rigid body spins at a rate  $\omega$  about its axis of symmetry and is subjected to a couple of moment  $M$  about an axis perpendicular to the spin axis, then the body will precess at a rate  $\Omega$  about an axis that is perpendicular to both the spin and the couple axis. The rate of precession  $\Omega$  is equal to  $\Omega = M / I\omega$ .

A well-known example of gyroscopic motion is the motion of a top (see Fig. 13.7), in which the couple moment  $M$ —due to gravity—is expected to force the top to fall. However, the top does not fall and rather precesses around the  $y$  axis.

**Fig. 13.7** Gyroscopic motion.



## Defining Terms

**Acceleration:** The rate of change of the velocity vector. Absolute acceleration of a rigid body is the rate of change of the velocity vector of the mass center of the body. Relative acceleration is the acceleration of a point on a body due to the angular velocity of that body only.

**D'Alembert forces:** Force and moment vectors due to the linear and angular accelerations of the body.

**Free-body diagram:** An essential sketch used to solve kinetics problems that involves sketching the rigid body together with all internal, reaction, and external force vectors.

**Gyroscopic motion:** Describes the motion of a rigid body that is spinning with a very large angular velocity around one axis when the couple of a moment is applied on the second axis. The resultant motion, called *precession*, is an angular velocity around the third axis.

**Kinematic analysis:** Starts from the geometry of constraints and uses differentiation to find the velocity and acceleration of the constrained points on a rigid body.

**Kinetic energy:** The accumulation of work of forces on a rigid body between two instants of time; includes kinetic energy due to translation and kinetic energy due to rotation.

**Rotation:** Centroidal rotation is the motion of a rigid body around an axis that passes through its mass center. In noncentroidal rotation the body rotates around an axis that passes through a point not corresponding to its mass center and which may not be on the body; this point is called the instantaneous center of rotation.

**Translation:** Rectilinear translation occurs when the velocity of any two points on the body remain equal in direction and magnitude throughout the entire duration of the motion. Curvilinear translation occurs when the velocity vectors of any two points are equal at any instant of time but change from one instant to another.

**Velocity:** Absolute velocity is the rate of change of the position vector of a point on a body measured from a fixed reference coordinate. Relative velocity is the rate of change of the position vector of a point on a rigid body measured from a moving reference frame.

## References

- Meriam, J. L. and Kraige, L. G. 1992. *Engineering Mechanics*, 3rd ed. John Wiley & Sons, New York.
- Beer, F. P. and Johnston, E. R. 1987. *Mechanics for Engineers-Dynamics*, 4th ed. McGraw-Hill, New York.
- Crandal, S. H., Karnopp, D. C., Kurtz., E. F, Jr., and Pridmore-Brown, D. C. 1968. *Dynamics of Mechanical and Electromechanical Systems*. McGraw-Hill, New York.
- Haug, E. J. 1989. *Computer-Aided Kinematics and Dynamics of Mechanical Systems. Volume I: Basic Methods*. Allyn & Bacon, Boston.
- Nikravesh, P. 1988. *Computer-Aided Analysis of Mechanical Systems*. Prentice Hall, Englewood Cliffs, NJ.
- Shabana, A. A. 1994. *Computational Dynamics*. John Wiley & Sons, New York.

## Further Information

Detailed treatment of the subject can be found in Meriam and Kraige [1992] and Beer and Johnston [1987].

A classical presentation of the subject can be found in Crandal *et al.* [1968].

Computer-aided analysis of the kinematics and dynamics of constrained rigid bodies in space motion can be found in Haug [1989] and Nikravesh [1988].



Mendelsohn, D. A. "Free Vibration, Natural Frequencies, and Mode Shapes"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Free Vibration, Natural Frequencies, and Mode Shapes

---

## 14.1 Basic Principles

### 14.2 Single-Degree-of-Freedom Systems

Equation of Motion and Fundamental Frequency • Linear Damping

### 14.3 Multiple-Degree-of-Freedom Systems

### 14.4 Continuous Systems (Infinite DOF)

**Daniel A. Mendelsohn**

*Ohio State University*

## 14.1 Basic Principles

---

In its simplest form, mechanical vibration is the process of a mass traveling back and forth through its position of static equilibrium under the action of a *restoring force* or *moment* that tends to return the mass to its equilibrium position. The most common restoring mechanism is a spring or elastic member that exerts a force proportional to the displacement of the mass. Gravity may also provide the restoring action, as in the case of a pendulum. The restoring mechanism of structural members is provided by the elasticity of the material of which the member is made. **Free vibration** is a condition in which there are no external forces on the system.

*Cyclic* or *periodic* motion in time is described by the property  $x(t + \zeta) = x(t)$ , where  $t$  is time and  $\zeta$  is the *period*, that is, the time to complete one cycle of motion. The **cyclic frequency** of the motion is  $f = 1/\zeta$ , usually measured in cycles per second (Hz). The special case of periodic motion shown in Fig. 14.1 is *harmonic motion*,

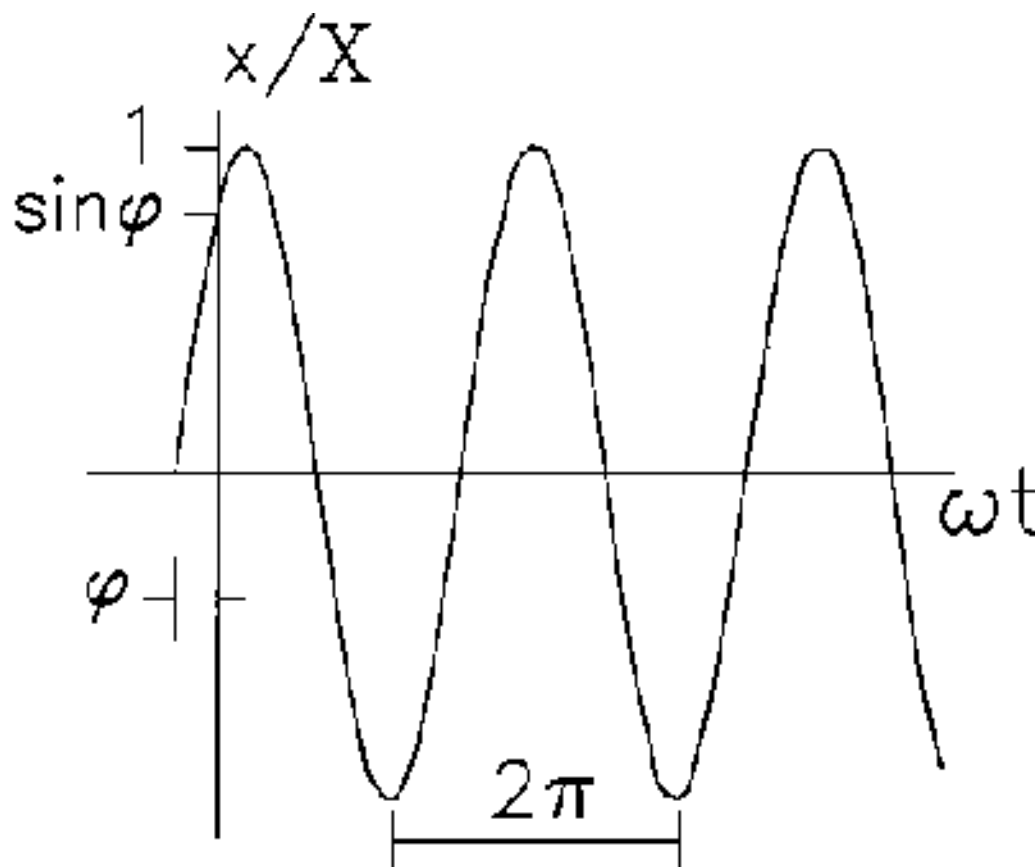
$$x(t) = A \sin(\omega t) + B \cos(\omega t) \quad (14:1a)$$

$$= X \sin(\omega t + \hat{A}) \quad (14:1b)$$

where  $\omega = 2\pi f$  is the **circular frequency**, typically measured in radians/s,

$X = (A^2 + B^2)^{1/2}$  is the *amplitude* of the motion, and  $\hat{A} = \tan^{-1}(B/A)$  is the *phase angle*. Many systems exhibit harmonic motion when in free vibration, but do so only at discrete **natural frequencies**. A vibrating system with  $n$  **degrees of freedom** (DOF) has  $n$  natural frequencies, and for each natural frequency there is a relationship between the amplitudes of the  $n$  independent motions, known as the **mode shape**. A structural elastic member has an infinite number of discrete natural frequencies and corresponding mode shapes. The **fundamental frequency** and associated mode shape refer to the smallest natural frequency and associated mode shape. The study of free vibrations consists of the determination of the natural frequencies and mode shapes of a vibrating system as a function of geometry, boundary conditions, mass (density) of the components, and the strength of the restoring forces or moments. Although the natural frequencies and mode shapes are valuable to know by themselves, they have perhaps their greatest value in the analysis of forced vibrations, as discussed in detail in the following chapter.

**Figure 14.1** Time history of undamped periodic or cyclic motion.



## 14.2 Single - Degree - of Freedom Systems

### Equation of Motion and Fundamental Frequency

The system shown in Fig. 14.2(a) consists of a mass,  $m$ , that rolls smoothly on a rigid floor and is attached to a linear spring of stiffness  $k$ . Throughout this chapter all linear (or longitudinal) springs have stiffnesses of dimension force per unit change in length from equilibrium, and all rotational (or torsional) springs have stiffnesses of dimension moment per radian of rotation from equilibrium (i.e., force times length). The distance of the mass from its equilibrium position, defined by zero stretch in the spring, is denoted by  $x$ . Applying Newton's second law to the mass in Fig. 14.3(a) gives the equation of motion:

$$-kx = m \frac{d^2x}{dt^2} \quad \Rightarrow \quad m \frac{d^2x}{dt^2} + kx = 0 \quad (14:2)$$

Alternatively, Lagrange's equation (with only one generalized coordinate,  $x$ ) may be used to find the equation of motion:

$$\frac{d}{dt} \left( \frac{\partial L}{\partial (\dot{x})} \right) - \frac{\partial L}{\partial x} = 0 \quad (14:3)$$

The Lagrangian,  $L$ , is the difference between the *kinetic energy*,  $T$ , and the *potential energy*,  $U$ , of the system. The Lagrangian for the system in Fig. 14.2(a) is

$$L = T - U = \frac{1}{2} m \dot{x}^2 - \frac{1}{2} kx^2 \quad (14:4)$$

Substituting Eq. (14.4) into Eq. (14.3) gives the same equation of motion as in Eq. (14.2). Using the harmonic form in Eq. (14.1) for  $x$ , Eq. (14.2) is satisfied if  $\omega$  takes on the value

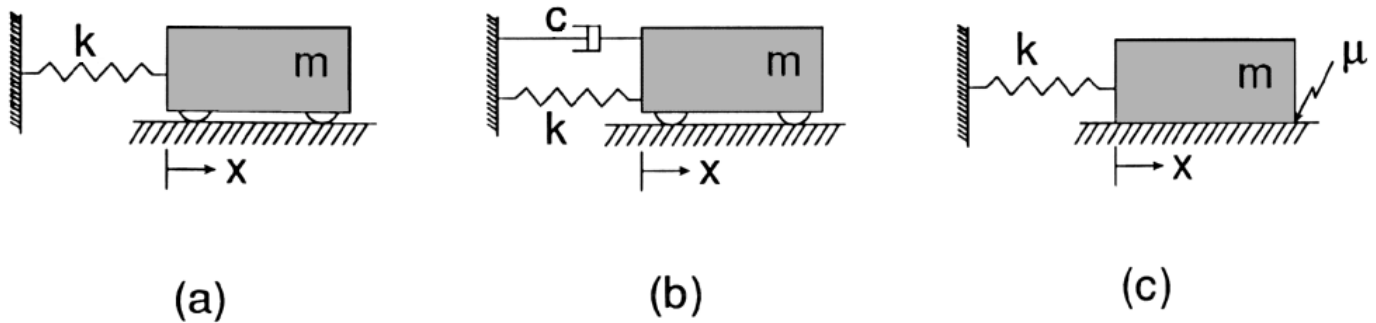
$$\omega = \sqrt{\frac{k}{m}} \quad (14:5)$$

which is therefore the natural frequency of the system. If the displacement and velocity are known at some time (say,  $t = 0$ ), then the constants in Eq. (14.1) may also be evaluated,

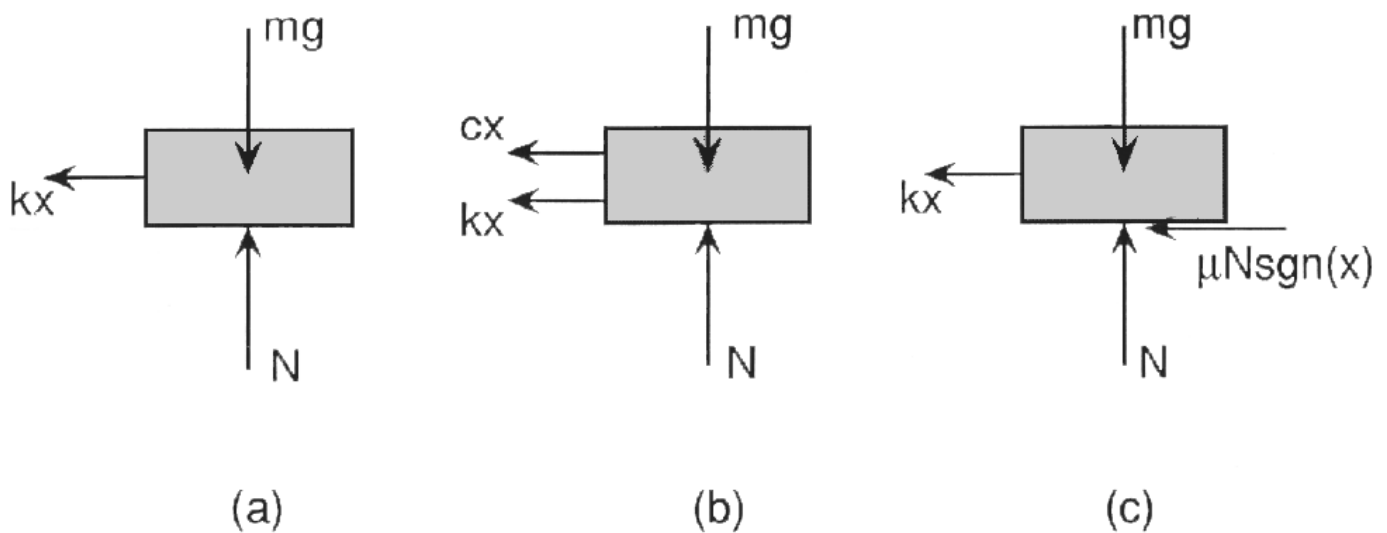
$$A = \frac{1}{\omega_0} \frac{dx}{dt}(0); \quad B = x(0) \quad (14:6)$$

and the corresponding displacement history is shown in Fig. 14.1.

**Figure 14.2** Typical one-degree-of-freedom system: (a) without damping, (b) with viscous damping, and (c) with frictional damping.



**Figure 14.3** Free-body diagrams of the single-degree-of-freedom systems in Figure 14.2.



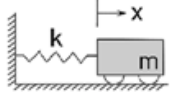
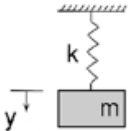
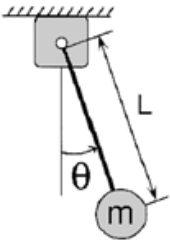
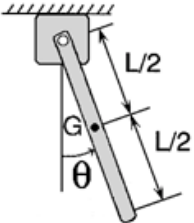
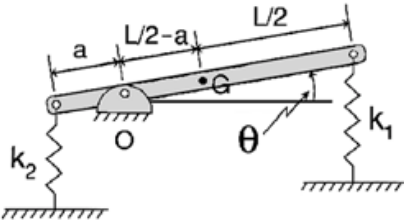
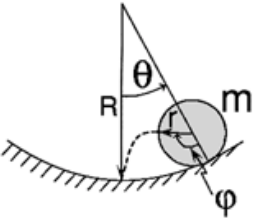
The natural frequency for conservative systems can also be found by the *energy method*. As the mass passes through equilibrium,  $U = 0$  and  $T = T_{\max}$ , while at its maximum displacement where the mass has zero velocity,  $T = 0$  and  $U = U_{\max}$ . Since the total energy is constant,  $\omega$  is the frequency for which  $T_{\max} = U_{\max}$ . Using Eq. (14.1a) and the system in Fig. 14.2(a), this principle gives

$$T_{\max} = \frac{1}{2}m(\omega X)^2 = \frac{1}{2}kX^2 = U_{\max} \quad (14:7)$$

which in turn gives the same result for  $\omega$  as in Eq. (14.5).

Table 14.1 contains the equation of motion and natural frequency for some single-DOF systems. Gravity acts down and displacements or rotations are with respect to static equilibrium. The mode shapes are of the form in Eq. (14.1) with  $\omega$  given in Table 14.1.

**Table 14.1** Equations of Motion and Natural Frequencies for some Single-DOF Systems

System	Equation of Motion	Natural Frequency
	$m \frac{d^2 x}{dt^2} + kx = 0$	$\sqrt{\frac{k}{m}}$
	$m \frac{d^2 y}{dt^2} + ky = 0$	$\sqrt{\frac{k}{m}}$
	$(mL^2) \frac{d^2 \theta}{dt^2} + (mgL)\theta = 0$	$\sqrt{\frac{g}{L}}$
 <p>bar has mass <math>m</math>, center of gravity <math>G</math></p>	$\left(\frac{1}{2}mL^2\right) \frac{d^2 \theta}{dt^2} + \left(mg\frac{L}{2}\right)\theta = 0$	$\sqrt{\frac{3g}{2L}}$
	$I_0 \frac{d^2 \theta}{dt^2} + [k_1 a^2 + k_2 (L - a)^2] \theta = 0$ $I_0 = \frac{1}{12}mL^2 + m\left(\frac{L}{2} - a\right)^2$	$\sqrt{\frac{k_1 a^2 + \left(\frac{k_2}{k_1}\right)^2 (L - a)^2}{\frac{1}{m} \left[\frac{1}{3}L^2 - aL + a^2\right]}}$
 <p><math>R\theta = r\phi</math></p>	$\frac{3}{2}mr^2 \left(\frac{R}{r} - 1\right) \frac{d^2 \theta}{dt^2} + (mgr)\theta = 0$ <p><math>r = \text{radius of mass } m</math></p>	$\sqrt{\frac{2g}{3(R - r)}}$

## Linear Damping

Figures 14.2(b) and 14.3(b) show an example of viscous damping caused by a dashpot of strength  $c$  (force per unit velocity) that acts opposite the velocity. Newton's second law then gives

$$-kx - c \frac{dx}{dt} = m \frac{d^2x}{dt^2} \Rightarrow m \frac{d^2x}{dt^2} + c \frac{dx}{dt} + kx = 0 \quad (14:8)$$

which has the solution

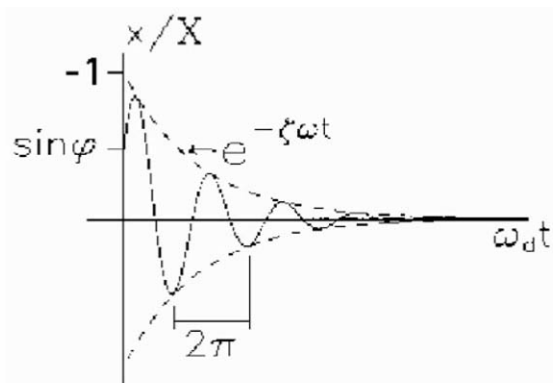
$$\begin{aligned} x(t) &= e^{-\zeta \omega_n t} [A \sin(\omega_d t) + B \cos(\omega_d t)] \\ &= X e^{-\zeta \omega_n t} \sin(\omega_d t + \hat{A}) \end{aligned} \quad (14:9)$$

where the *damped natural frequency*,  $\omega_d$ , *damping factor*,  $\zeta$ , and *critical damping coefficient*,  $c_c$ , are given by

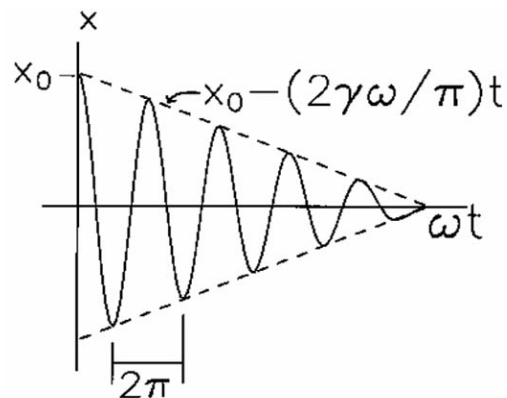
$$\omega_d = \omega_n \sqrt{1 - \zeta^2}; \quad \zeta = \frac{c}{c_c}; \quad c_c = 2m\omega_n = 2\sqrt{mk} \quad (14:10)$$

respectively. If  $c < c_c$ , then  $\omega_d$  is real and Eq. (14.9) represents exponentially damped oscillation, as shown in Fig. 14.4. If  $c \geq c_c$ , then the system is *supercritically damped* and decaying motion but no vibration occurs.

**Figure 14.4** Time history of viscously damped vibration.



**Figure 14.5** Time history of frictionally damped vibration.



The frictional effects in Figs. 14.2(c) and 14.3(c) are characterized by a Coulomb frictional force  $F = \mu N = \mu mg$ , where  $\mu$  is the coefficient of sliding friction. The equation of motion is then

$$m \frac{d^2x}{dt^2} + (\mu mg) \operatorname{sgn} \left( \frac{dx}{dt} \right) + kx = 0 \quad (14:11)$$

where  $\operatorname{sgn}(dx/dt)$  is equal to  $+1$  or  $-1$  for positive or negative values of  $dx/dt$ , respectively. This equation must be solved separately for each  $n$ th half period of the oscillation of frequency,  $\omega$ ,

$$x(t) = [x_0 - (\mu mg/k)] \cos(\omega t) + \operatorname{sgn} \left( \frac{dx}{dt} \right) \frac{\mu mg}{k} \quad (14:12)$$

where  $\frac{\mu mg}{k}$  is the minimum initial displacement to allow motion, and  $\omega$  is the undamped natural frequency, Eq. (14.5). Figure 14.5 shows  $x(t)$  for an initial displacement of  $x_0 = 20^\circ$ .

## 14.3 Multiple-Degree-of-Freedom Systems

For each DOF in an  $n$ -DOF system there is a *coordinate*,  $x_i$  ( $i = 1; 2; \dots; n$ ), which is a measure of one of the independent components of motion. The motion of the system is governed by  $n$ , generally coupled, equations of motion, which may be obtained by a Newtonian approach requiring complete free-body and acceleration diagrams for each mass. For systems with many DOFs this approach becomes very tedious. Alternatively, applying Lagrange's equations, with no damping present,

$$\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{x}_i} \right) - \frac{\partial L}{\partial x_i} = 0; \quad (i = 1; 2; \dots; n) \quad (14:13)$$

to the particular system yields the  $n$  equations of motion in the  $n$  unknown coordinates  $x_i$ ,

$$[M] \frac{d^2x}{dt^2} + [K]x = f_0g \quad (14:14)$$

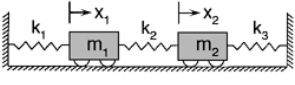
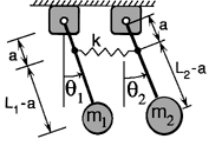
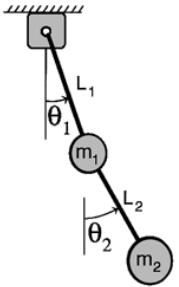


As stated before, the Lagrangian,  $L = T - U$ , is the difference between the kinetic and potential energies of the system.  $[M]$  and  $[K]$  are the  $n \times n$  *mass* and *stiffness matrices*, with elements  $m_{ij}$  and  $k_{ij}$ , which multiply the acceleration and displacement vectors of the masses, respectively. Writing  $x_i$  in the form of Eq. (14.1), Eq. (14.14) yields  $n$  homogeneous equations,  $[A]fXg = f0g$ , in the  $n$  amplitudes  $X_i$ . The elements of  $[A]$  are  $a_{ij} = k_{ij} - m_{ij}\omega_i^2$ . If a solution exists, the determinant of  $[A]$ , an  $n$ th order polynomial in  $\omega_i^2$ , must be zero. This yields the *frequency* or *characteristic equation*, whose  $n$  roots are the natural frequencies squared,  $(\omega_i)^2$ . Each mode shape may be written as a vector of  $n + 1$  amplitude ratios:

$$\frac{X_2}{X_1}, \frac{X_3}{X_1}, \dots, \frac{X_n}{X_1} \quad ; \quad (i = 1; 2; \dots; n) \quad (14.15)$$

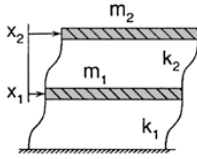
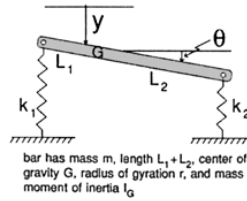
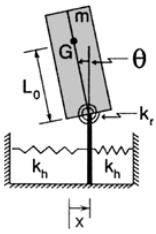
The ratios are found by eliminating one equation of  $[A]fXg = f0g$ , dividing the remaining  $n + 1$  equations by  $X_1$ , and solving. Then setting  $\omega = \omega_i$  gives the  $i$ th mode shape. Equations of motion, natural frequencies, and mode shapes for some two-DOF systems undergoing small amplitude vibrations are in Table 14.2. Gravity acts down and displacements and rotations are taken with respect to the position of static equilibrium.

**Table 14.2** Equations of Motion, Natural Frequencies, and Mode Shapes for some Two-DOF Systems

System	Equations of Motion	Natural Frequencies and Mode Shapes
	$\begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix} \begin{bmatrix} \frac{d^2 x_1}{dt^2} \\ \frac{d^2 x_2}{dt^2} \end{bmatrix} + \begin{bmatrix} k_1 + k_2 & -k_2 \\ -k_2 & k_2 + k_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\omega_1 = \sqrt{\frac{k}{2m_1}} \sqrt{A - B}, \quad \omega_2 = \sqrt{\frac{k}{2m_2}} \sqrt{A + B}$ $A = \frac{m_1(k_2 + k_3) + m_2(k_1 + k_2)}{m_1 k_2}$ $B = \sqrt{A^2 - 4 \frac{m_2}{m_1} \left[ \frac{(k_1 + k_2)(k_2 + k_3)}{k_2^2} - 1 \right]}$ $\left( \frac{X_2}{X_1} \right)_i = 1 + \frac{k_1}{k_2} - \frac{m_1}{k_2} \omega_i^2; \quad i = 1, 2$
	$\begin{bmatrix} m_1 L_1^2 & 0 \\ 0 & m_2 L_2^2 \end{bmatrix} \begin{bmatrix} \frac{d^2 \theta_1}{dt^2} \\ \frac{d^2 \theta_2}{dt^2} \end{bmatrix} + \begin{bmatrix} m_1 g L_1 + k a^2 & -k a^2 \\ -k a^2 & m_2 g L_2 + k a^2 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\omega_1 = \sqrt{\frac{g}{2L_1}} \sqrt{A - B}, \quad \omega_2 = \sqrt{\frac{g}{2L_1}} \sqrt{A + B}$ $A = 1 + \frac{L_1}{L_2} + \frac{k a^2}{m_1 L_1 g} \left( 1 + \frac{m_1 L_1^2}{m_2 L_2^2} \right)$ $B = \sqrt{A^2 - 4 \left[ 1 + \frac{k a^2}{m_1 L_1 g} \left( 1 + \frac{m_2 L_2}{m_1 L_1} \right) \right]}$ $\left( \frac{\Theta_2}{\Theta_1} \right)_i = -1 + \frac{m_1 L_1}{k a^2} (\omega_i^2 L_1 - g); \quad i = 1, 2$
	$\begin{bmatrix} m_1 L_1 & 0 \\ m_2 L_1 & m_2 L_2 \end{bmatrix} \begin{bmatrix} \frac{d^2 \theta_1}{dt^2} \\ \frac{d^2 \theta_2}{dt^2} \end{bmatrix} + \begin{bmatrix} (m_1 + m_2)g & -m_2 g \\ -k a^2 & -m_2 g \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\omega_1 = \sqrt{\frac{g}{2L_1}} \sqrt{A - B}, \quad \omega_2 = \sqrt{\frac{g}{2L_1}} \sqrt{A + B}$ $A = \left( 1 + \frac{m_1}{m_2} \right) \left( 1 + \frac{L_1}{L_2} \right)$ $B = \sqrt{A^2 - 4 \left( 1 + \frac{m_2}{m_1} \right) \left( \frac{L_1}{L_2} \right)}$ $\left( \frac{\Theta_2}{\Theta_1} \right)_i = 1 + \left( \frac{m_1}{m_2} \right) \left( 1 - \frac{\omega_i^2 L_1}{g} \right); \quad i = 1, 2$

(continues)

**Table 14.2** Equations of Motion, Natural Frequencies, and Mode Shapes for some Two-DOF Systems (*continued*)

System	Equations of Motion	Natural Frequencies and Mode Shapes
	$\begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix} \begin{bmatrix} \frac{d^2 x_1}{dt^2} \\ \frac{d^2 x_2}{dt^2} \end{bmatrix} + \begin{bmatrix} k_1 + k_2 & -k_2 \\ -k_2 & k_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\omega_1 = \sqrt{\frac{k_1}{2m_1}} \sqrt{A - B}, \quad \omega_2 = \sqrt{\frac{k_1}{2m_1}} \sqrt{A + B}$ $A = 1 + \frac{k_2}{k_1} + \frac{k_2}{k_1} \frac{m_1}{m_2}$ $B = \sqrt{A^2 - 4 \frac{k_2}{k_1} \frac{m_1}{m_2}}$ $\left( \frac{X_2}{X_1} \right)_i = 1 + \frac{k_1}{k_2} - \frac{m_1}{k_2} \omega_i^2; \quad i = 1, 2$
 <p>bar has mass <math>m</math>, length <math>L_1 + L_2</math>, center of gravity <math>G</math>, radius of gyration <math>r</math>, and mass moment of inertia <math>I_G</math></p>	$\begin{bmatrix} m & 0 \\ 0 & mr^2 \end{bmatrix} \begin{bmatrix} \frac{d^2 y}{dt^2} \\ \frac{d^2 \theta}{dt^2} \end{bmatrix} + \begin{bmatrix} k_1 + k_2 & k_2 L_2 - k_1 L_1 \\ k_2 L_2 - k_1 L_1 & k_1 L_1^2 + k_2 L_2^2 \end{bmatrix} \begin{bmatrix} y \\ \theta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $r = \text{Radius of gyration} = \sqrt{\frac{I_G}{m}}$	$\omega_1 = \sqrt{\frac{k_1}{2m}} \sqrt{A - B}, \quad \omega_2 = \sqrt{\frac{k_1}{2m}} \sqrt{A + B}$ $A = 1 + \frac{L_1^2}{r^2} + \frac{k_2}{k_1} \frac{L_2^2}{r^2} + \frac{k_2}{k_1}$ $B = \sqrt{A^2 - 4 \frac{k_2}{k_1} \frac{(L_1 + L_2)^2}{r^2}}$ $\left( \frac{Y}{\theta} \right)_i = \frac{mr^2 \omega_i^2 - (k_1 L_1^2 + k_2 L_2^2)}{k_2 L_2 - k_1 L_1}; \quad i = 1, 2$
	$\begin{bmatrix} m & -mL_0^2 \\ 0 & mr^2 \end{bmatrix} \begin{bmatrix} \frac{d^2 x}{dt^2} \\ \frac{d^2 \theta}{dt^2} \end{bmatrix} + \begin{bmatrix} k_h & 0 \\ k_h L_0 & k_r \end{bmatrix} \begin{bmatrix} x \\ \theta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $r = \text{Radius of gyration} = \sqrt{\frac{I_G}{m}}$	$\omega_1 = \sqrt{\frac{k_h}{2m}} \sqrt{A - B}, \quad \omega_2 = \sqrt{\frac{k_h}{2m}} \sqrt{A + B}$ $A = 1 + \frac{L_0^2}{r^2} + \frac{k_r}{k_h r^2}$ $B = \sqrt{A^2 - 4 \frac{k_r}{k_h r^2}}$ $\left( \frac{X}{\theta} \right)_i = \frac{mr^2 \omega_i^2 - k_r}{k_h L_0}; \quad i = 1, 2$

## 14.4 Continuous Systems (Infinite DOF)

The equations of motion of structural members made up of continuously distributed elastic or flexible materials are most easily obtained by a Newtonian analysis of a representative volume element. As an example, consider the longitudinal vibration of an elastic rod (Young's modulus  $E$ , density  $\rho$ ) of cross-sectional area  $A$ . A free-body diagram of a volume element  $A dx$ , with normal stresses  $[F_x](x)$  and  $[F_x + (\partial F_x / \partial x) dx](x)$  acting on the cross sections is shown in Fig. 14.6(a). A circular section is shown but the analysis applies to any shape of cross section. If  $u(x)$  is the displacement in the  $x$  direction of the cross section at  $x$ , then Newton's second law gives

$$\rho A dx + \frac{\partial F_x}{\partial x} dx = (\rho A dx) \frac{\partial^2 u}{\partial t^2} \quad (14.16)$$

Simplifying, letting  $dx$  go to zero, and noting uniaxial Hooke's law and the definition of the strain,  $\epsilon_x$ ,

$$\frac{\partial^3 u}{\partial x^3} = E \frac{\partial^2 u}{\partial x^2} = E \frac{\partial}{\partial x} \left( \frac{\partial u}{\partial x} \right) \quad (14:17)$$

Eq. (14.16) can be written as

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} \quad (14:18)$$

which is the equation of motion for standing modes of free vibration and for wave propagation along the rod at velocity  $c = \sqrt{E/\rho}$ . If  $u(x; t) = U(x)[A \sin(\omega t) + B \cos(\omega t)]$ , then Eq. (14.18) gives

$$\frac{d^2 U}{dx^2} + \omega^2 U = 0; \quad \omega^2 = \frac{1}{c^2} \frac{E}{\rho} \quad (14:19)$$

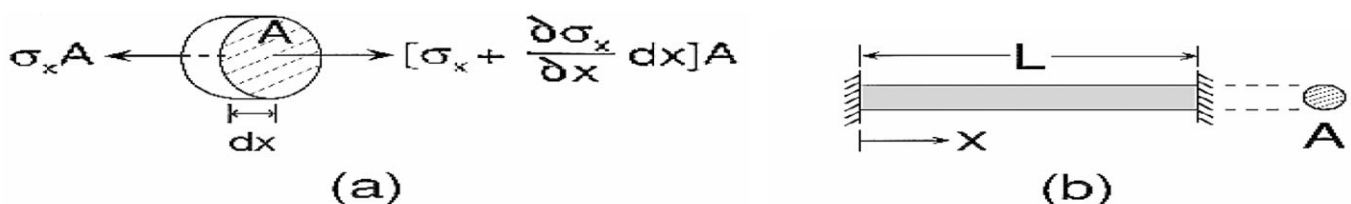
which has solution  $U(x) = C \sin(\omega x) + D \cos(\omega x)$ . Now as an example, consider the fixed-fixed bar of length  $L$  shown in Fig. 14.6(b) that has boundary conditions (BCs)  $U(0) = 0$  and  $U(L) = 0$ , which give, respectively,  $D = 0$  and either  $C = 0$ , which is not of interest, or

$$\sin(\omega L) = 0 \Rightarrow \omega = \omega_n = \frac{n\pi}{L} \Rightarrow \omega_n^2 = \frac{n^2 \pi^2}{L^2} \frac{E}{\rho} \quad (n = 1; 2; 3; \dots) \quad (14:20)$$








This is the frequency equation and the resulting infinite set of discrete natural frequencies for the fixed-fixed beam of length  $L$ . The mode shapes are  $U_n(x) = \sin(\omega_n x)$ .

The transverse motion  $y(x; t)$  of a taut flexible string (tension  $T$  and mass per unit length  $\rho$ ), the longitudinal motion  $u(x; t)$  of a rod (Young's modulus  $E$ ), and the torsional rotation  $\theta(x; t)$  of a rod of circular or annular cross section (shear modulus  $G$ ) all share the same governing equations (14.18) and (14.19), but with different  $c$  values:  $c^2 = T/\rho$ ,  $c^2 = E/\rho$ , and  $c^2 = G/\rho$ , respectively. Tables 14.3 and 14.4 contain frequency equations, nondimensional natural frequencies, and mode shapes for various combinations of BCs for a rod of length  $L$ . Only the fixed-fixed conditions apply to the string.

**Figure 14.6** Longitudinal vibration of a rod of circular cross section: (a) free-body diagram of representative volume element and (b) a clamped-clamped rod of length  $L$ .



**Table 14.3** Longitudinal and Torsional Vibration of a Rod

System	Frequency Equation	Natural Frequencies	Normalized Mode Shapes
	$\sin(\lambda_n L) = 0$	$(\lambda_n L) = n\pi$	$U_n(x) = \sin(\lambda_n x)$
	$\sin(\lambda_n L) = 0$	$(\lambda_n L) = (2n - 1)\pi/2$	$U_n(x) = \sin(\lambda_n x)$
	$\tan(\lambda_n L) = -\gamma(\lambda_n L)^1$	see Table 14.4	$U_n(x) = \sin(\lambda_n x)$
	$(\lambda_n L) \tan(\lambda_n L) = \gamma^2$	see Table 14.4	$U_n(x) = \sin(\lambda_n x)$
	$\sin(\lambda_n L) = 0$	$(\lambda_n L) = n\pi$	$U_n(x) = \cos(\lambda_n x)$
	$(\lambda_n L) \tan(\lambda_n L) = -\gamma^3$	see Table 14.4	$U_n(x) = \cos(\lambda_n x)$
	$\tan(\lambda_n L) = -\gamma(\lambda_n L)^4$	see Table 14.4	$U_n(x) = \cos(\lambda_n x)$

$$^1 \gamma_{\text{long}} = \frac{AE}{kL}; \quad \gamma_{\text{tor}} = \frac{I_p G}{kL}$$

$$^2 \gamma_{\text{long}} = \frac{\rho AL}{m}; \quad \gamma_{\text{tor}} = \frac{I_p \rho L}{I_0}$$

$$^3 \gamma_{\text{long}} = \frac{kL}{AE}; \quad \gamma_{\text{tor}} = \frac{kL}{I_p G}$$

$$^4 \gamma_{\text{long}} = \frac{m}{\rho AL}; \quad \gamma_{\text{tor}} = \frac{I_0}{I_p \rho L}$$

Notes:

“long” denotes longitudinal vibration and “tor” denotes torsional vibration

$A$  = cross-sectional area,  $L$  = rod length,  $E$  = Young’s modulus,  $G$  = shear modulus,  $k$  = force per length (long) or moment per radian (tor),  $\rho$  = mass per unit volume,  $I_p$  = polar moment of inertia of  $A$  about rod axis,  $I_0$  = mass moment of inertia of attached mass, the definition of  $\lambda_n$  is in the text.

**Table 14.4** Nondimensional Natural Frequencies  $(\lambda_n L)^1$ 

	$ \gamma $							
$n$	0	.5	1	2	5	10	100	$\infty$
For Longitudinal and Torsional Clamped/Spring and Free/Mass BCs <sup>2</sup>								
1	$\pi$	2.289	2.029	1.837	1.689	1.632	1.577	$\pi/2$
2	$2\pi$	5.087	4.913	4.814	4.754	4.734	4.715	$3\pi/2$
3	$3\pi$	8.096	7.979	7.917	7.879	7.867	7.855	$5\pi/2$
For Longitudinal and Torsional Clamped/Mass and Torsional Free/Spring BCs <sup>3</sup>								
1	0	.653	.860	1.077	1.314	1.429	1.555	$\pi/2$
2	$\pi$	3.292	3.426	3.644	4.034	4.306	4.666	$3\pi/2$
3	$2\pi$	6.362	6.437	6.578	6.910	7.228	7.776	$5\pi/2$
For Longitudinal Free/Spring BCs <sup>4</sup>								
1	$\pi$	2.975	2.798	2.459	1.941	1.743	1.587	$\pi/2$
2	$2\pi$	6.203	6.121	5.954	5.550	5.191	4.760	$3\pi/2$
3	$3\pi$	9.371	9.318	9.211	8.414	8.562	7.933	$5\pi/2$

<sup>1</sup>For the nonclassical boundary conditions in Table 14.3.

<sup>2</sup>See Table 14.3, cases 1 and 4.

<sup>3</sup>See Table 14.3, cases 2 and 3.

<sup>4</sup>See Table 14.3, case 3.

The transverse deflection of a beam,  $w(x; t)$ , is governed by the equation of motion,

$$\frac{\partial^4 w}{\partial x^4} = \frac{\mu}{EI} \frac{\partial^2 w}{\partial t^2} \quad (14:21)$$

which, upon substitution of  $w(x; t) = W(x)[A \sin(\omega t) + B \cos(\omega t)]$ , leads to

$$\frac{d^4 W}{dx^4} + \omega^4 W = 0; \quad \omega^4 = \frac{\mu A \omega^2}{EI} \quad (14:22)$$

This equation has the general solution

$W(x) = c_1 \sin(\omega x) + c_2 \cos(\omega x) + c_3 \sinh(\omega x) + c_4 \cosh(\omega x)$ . The frequency equation, natural frequencies, and normalized mode shapes are found by applying the BCs in the same manner as above. The results for various combinations of simply supported (SS:  $W = W'' = 0$ ), clamped (C:  $W = W' = 0$ ), and free (F:  $W'' = W''' = 0$ ) BCs for a beam of length  $L$  and flexural rigidity  $EI$  are given in Table 14.5.

## Defining Terms

**Cyclic and circular frequency:** The cyclic frequency of any cyclic or periodic motion is the number of cycles of motion per second. One cycle per second is called a hertz (Hz). The circular frequency of the motion is  $2\pi$  times the cyclic frequency and converts one cycle of motion into  $2\pi$  radians of angular motion. The circular frequency is measured in radians per second.

**Degree of freedom (DOF):** An independent motion of a moving system. A single mass rolling on a surface has one DOF, a system of two masses rolling on a surface has two DOFs, and a continuous elastic structure has an infinite number of DOFs.

**Free vibration:** The act of a system of masses or a structure vibrating back and forth about its position of static equilibrium in the absence of any external forces. The vibration is caused by the action of restoring forces internal to the system or by gravity.

**Fundamental frequency:** The smallest natural frequency in a system with more than one DOF.

**Mode shape:** The relationship between the amplitudes (one per DOF) of the independent motions of a system in free vibration. There is one mode shape for each natural frequency and it depends on the value of that natural frequency. For a continuous elastic structure the mode shapes are the shapes of the structure at its maximum deformation during a cycle of vibration.

**Natural frequency:** The frequency or frequencies at which a system will undergo free vibration. There is one natural frequency per DOF of the system. Natural frequencies depend on the geometry, the boundary conditions (method of support or attachment), the masses of the components, and the strength of the restoring forces or moments.

## References

- Clark, S. K. 1972. *Dynamics of Continuous Elements*. Prentice Hall, Englewood Cliffs, NJ.
- Den Hartog, J. P. 1956. *Mechanical Vibrations*, 4th ed. McGraw-Hill, New York.
- Gorman, D. J. 1975. *Free Vibration Analysis of Beams and Shafts*. John Wiley & Sons, New York.
- Leissa, A. W. 1993a. *Vibrations of Plates*. Acoustical Society of America, New York. (Originally issued by NASA, 1973.)
- Leissa, A. W. 1993b. *Vibrations of Shells*. Acoustical Society of America, New York. (Originally issued by NASA, 1973.)
- Magrab, E. B. 1979. *Vibrations of Elastic Structural Members*. Sijthoff and Noordhoff, Leyden, The Netherlands.
- Meirovitch, L. 1967. *Analytical Methods in Vibrations*. Macmillan, New York.
- Thomson, W. T. 1988. *Theory of Vibrations with Applications*. Prentice Hall, Englewood Cliffs, NJ.
- Timoshenko, S. P., Young, D. H., and Weaver, J. W. 1974. *Vibration Problems in Engineering*, 4th ed. John Wiley & Sons, New York.

## Further Information

There are several excellent texts that discuss the free vibrations of discrete systems (finite number of DOFs). In particular the books by Den Hartog [1956], Timoshenko *et al.* [1974], and Thomson [1988] are recommended.

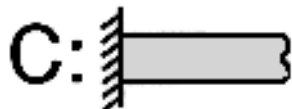
Extensive data for the natural frequencies of beams having elastic supports (translational or rotational), end masses, multiple spans, discontinuities in cross sections, axial tension or compression, variable thickness, or elastic foundations may be found in the monograph by

Gorman.

Other important structural elements are plates and shells. Plates are flat, whereas shells have curvature (e.g., circular cylindrical, elliptic cylindrical, conical, spherical, ellipsoidal, hyperboloidal). A summary of natural frequencies for plates obtained from 500 other references is available in the book on plate vibrations by Leissa [1993a]. Extensive frequency data for various shells taken from 1000 references is also available in the book on shell vibrations by Leissa [1993b].

**Table 14.5** Transverse Vibrations of a Beam

BCs	Frequency Equation	$\beta_1$	$\beta_2$	Asymptotic to	Normalized Mode Shape
C-F	$1 + \cos \beta \cosh \beta = 0$	1.875	4.694	$(2n + 1)\pi/2$	$(\cosh \lambda_n x - \cos \lambda_n x)$ $-\gamma_n(\sinh \lambda_n x - \sin \lambda_n x),$ $\gamma_n = \frac{\cosh \beta_n + \cos \beta}{\sinh \beta_n + \sin \beta}$
SS-SS	$\sin \beta = 0$	$\pi$	$2\pi$	$n\pi$	$\sin \lambda_n x$
C-SS	$\tanh \beta - \tan \beta = 0$	3.927	7.069	$(4n + 1)\pi/4$	$(\cosh \lambda_n x - \cos \lambda_n x)$ $-\gamma_n(\sinh \lambda_n x - \sin \lambda_n x),$ $\gamma_n = \frac{\cosh \beta_n - \cos \beta}{\sinh \beta_n - \sin \beta}$
F-SS	$\tanh \beta - \tan \beta = 0$	3.927	7.069	$(4n + 1)\pi/4$	$(\cosh \lambda_n x + \cos \lambda_n x)$ $-\gamma_n(\sinh \lambda_n x + \sin \lambda_n x),$ $\gamma_n = \frac{\cosh \beta_n - \cos \beta}{\sinh \beta_n - \sin \beta}$
C-C	$1 - \cos \beta \cosh \beta = 0$	4.730	7.853	$(2n + 1)\pi/2$	$(\cosh \lambda_n - \cos \lambda_n x)$ $-\gamma_n(\sinh \lambda_n x - \sin \lambda_n x),$ $\gamma_n = \frac{\sinh \beta_n + \sin \beta}{\cosh \beta_n - \cos \beta}$
F-F	$1 - \cos \beta \cosh \beta = 0$	4.730	7.853	$(2n + 1)\pi/2$	$(\cosh \lambda_n + \cos \lambda_n x)$ $-\gamma_n(\sinh \lambda_n x + \sin \lambda_n x),$ $\gamma_n = \frac{\sinh \beta_n + \sin \beta}{\cosh \beta_n - \cos \beta}$



Leissa, A. W. "Forced Vibrations"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



### 15.1 Single-Degree-of-Freedom Systems

### 15.2 Multiple-Degree-of-Freedom Systems

**Arthur W. Leissa**

*Ohio State University*

Consider a mechanical system that is subjected to external forces (or moments) that are periodic in time. The forces may arise in various ways. For example, forces may be applied directly to the system (mechanical connections, fluid pressure, electromechanical), or indirectly through a foundation (which may be represented by springs and dampers). Such exciting forces always occur in rotating bodies (e.g., electric motors, internal combustion engines, gas turbines), but can also have other sources (e.g., earthquake motions, wind gusts, acoustic excitations).

The frequency ( $\Omega$ ) of an exciting force is typically different from the natural frequencies ( $\omega_1, \omega_2, \omega_3, \dots$ ) of the system. However, if  $\Omega$  is close to *any* of the natural frequencies, the amplitude of the resulting motion may be very large. If  $\Omega$  *equals* one of the  $\omega_i$ , **resonance** exists. In this situation, if no damping were present, the amplitude would grow with time until the system failed due to excessive motion or stress. All physical systems have at least some damping, but the damping may be very small. In this situation, the amplitude of motion at resonance would remain finite, but could become very large—even excessive. When a system is excited, the responsive displacements are a combination (superposition) of all the mode shapes of free vibration. However, if  $\Omega$  is close to one of the  $\omega_i$ , the response is dominated by the mode shape corresponding to that  $\omega_i$ .

The most important reason to know the natural frequencies of free vibration is to avoid resonant situations. One seeks to change the mass or stiffness of the system to shift the natural frequencies away from the exciting frequencies. In typical situations, the largest resonant amplitudes occur at the lowest natural frequencies. Therefore, it is particularly important to know the smallest  $\omega_i$ . Free vibration mode shapes are also important because they enable one to determine *how* the system vibrates at or near resonance.

---

## 15.1 Single-Degree-of-Freedom Systems

Take the spring-mass system shown in Fig. 14.2(a) of Chapter 14 and add a horizontal exciting force  $F_o \sin \Omega t$  to the mass, where  $\Omega$  is the *exciting frequency*. From the free-body diagram of Fig. 14.3(a), the equation of motion is

$$m\ddot{x} + kx = F_o \sin \Omega t \quad (15.1)$$

The solution of Eq. (15.1) consists of the sum of two parts. One part is the *complementary* solution obtained by setting  $F_o = 0$ . This is the free, undamped vibration discussed in **Chapter 14**. The second part is the *particular* solution, due to  $F_o \sin \Omega t$ . This is

$$x = \frac{F_o/k}{1 - (\Omega/\omega)^2} \sin \Omega t \quad (15.2)$$

where  $\omega = \sqrt{k/m}$  is the *natural frequency*. Observing the amplitude of this motion in Eq. (15.2), one sees that if excitation begins with a small frequency ( $\Omega/\omega \ll 1$ ) and increases, the amplitude grows until, at  $\Omega/\omega = 1$ , it becomes (theoretically) infinite. This is resonance. As  $\Omega/\omega$  increases further, the amplitude diminishes. For large  $\Omega/\omega$ , it becomes very small.

If viscous damping is present, as represented in Fig. 14.2(b) of **Chapter 14**, the equation of motion is

$$m\ddot{x} + c\dot{x} + kx = F_o \sin \Omega t \quad (15.3)$$

Again the solution has two parts, one part being the free, damped vibration, and the other part being the forced motion. The free vibration part is given by Eq. (14.9) of **Chapter 14**. It decays with increasing time and eventually vanishes (i.e., it is *transient*). The forced vibration part is

$$x = A \sin \Omega t - B \cos \Omega t = C \sin(\Omega t - \phi) \quad (15.4a)$$

$$C = \sqrt{A^2 + B^2}, \quad \phi = \tan^{-1}(B/A) \quad (15.4b)$$

$$C = \frac{F_o/k}{[1 - (\Omega/\omega)^2]^2 + [2\zeta(\Omega/\omega)]^2} \quad (15.4c)$$

where  $\zeta = c/c_c$ ,  $c_c = 2\sqrt{mk}$  as in **Chapter 14**. This forced vibration is called the **steady state** vibration because it stays indefinitely, even after the transient, free vibration vanishes.

A graph of steady state amplitude versus forcing frequency is shown in Fig. 15.1. This graph is worthy of considerable study for it shows clearly what vibratory amplitudes exist at different forcing frequencies. The *nondimensional amplitude*  $C/\delta_{st}$  is used, where  $\delta_{st} = F_o/k$  is the **static deflection** that the mass would have if  $F_o$  were applied. For small  $\Omega/\omega$ , Fig. 15.1 shows that  $C/\delta_{st} = 1$ , regardless of the damping. The case discussed earlier with *no damping* ( $\zeta = 0$ ) is shown, although  $C/\delta_{st}$  is plotted positive for  $\Omega/\omega > 1$ . It is positive for all nonzero  $\zeta$  (and for all  $\Omega/\omega$ ), no matter how small. For no damping, the infinite amplitude at resonance is implied in Fig. 15.1. For small damping (e.g.,  $\zeta = 0.1$ ), the peak amplitude is several times the static deflection. If  $\zeta$  were only 0.01, the peak amplitude would be 50 times the static deflection.

**Figure 15.1** Displacement amplitude and phase angle resulting from an applied force, versus exciting frequency for various amounts of damping (one DOF).

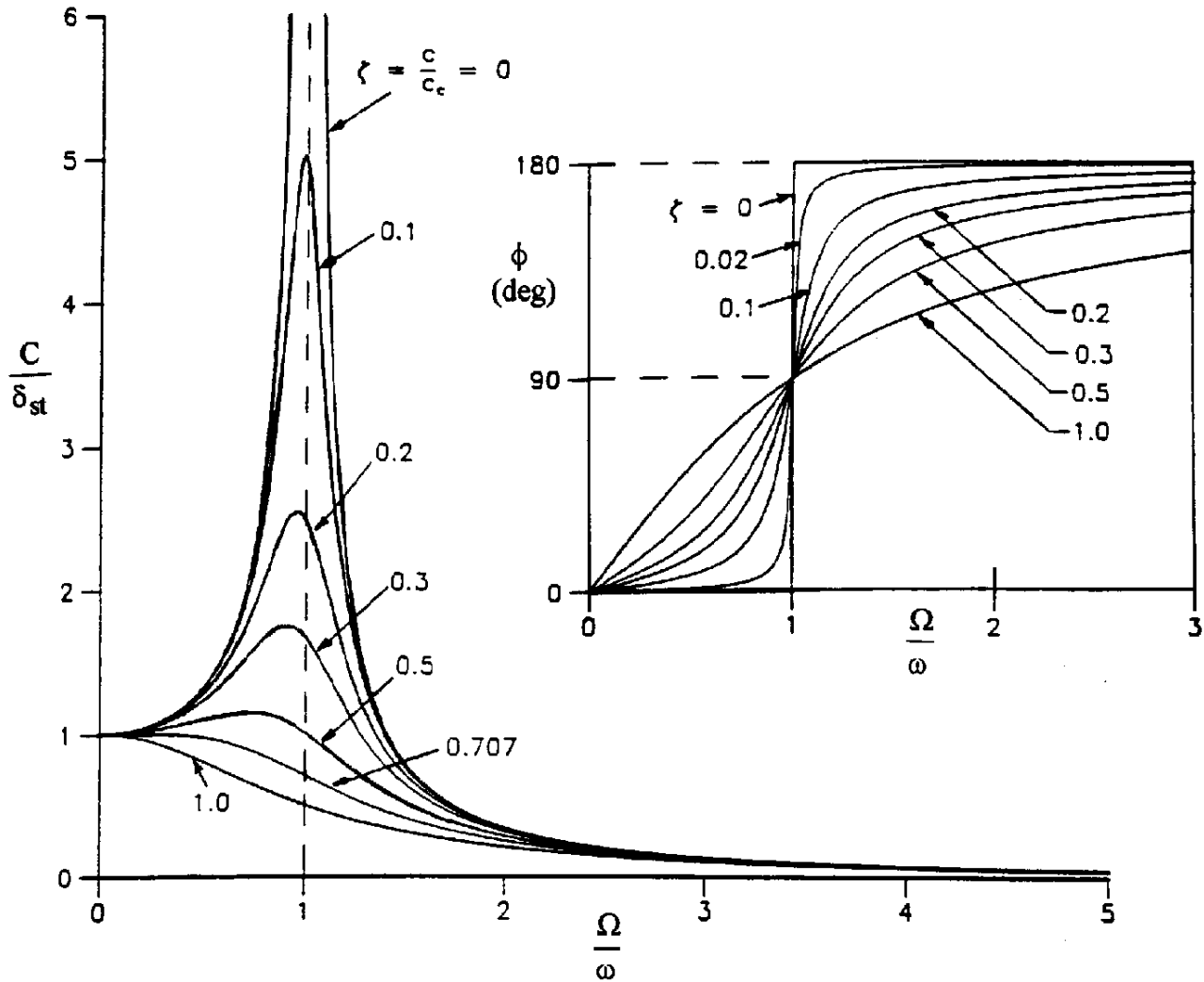


Figure 15.1 also shows the **phase angle**,  $\phi$ ; that is, the angle by which the motion *lags* the exciting force. For small  $\Omega/\omega$ , it is seen that the motion is essentially **in-phase** ( $\phi$  is nearly zero), whereas, for  $\Omega/\omega \gg 1$ , the motion is essentially **out-of-phase** ( $\phi$  is nearly  $180^\circ$ ). In the vicinity of resonance ( $\Omega/\omega = 1$ ),  $\phi$  changes rapidly as  $\Omega$  is varied, especially if the damping is small.

Suppose that, instead of applying an exciting force  $F_o \sin \Omega t$  directly to the mass in Fig. 14.2(b) of Chapter 14, the wall (or foundation) on the left side is given the vibratory displacement  $\delta_w \sin \Omega t$ . This motion causes forces to be transmitted through the spring and damper to the mass. One finds that the equation of motion is again Eq. (15.3), with  $F_o$  replaced by  $k\delta_w$ . Thus, Fig. 15.1 again describes the steady state vibratory amplitude of the mass, except that  $\delta_{st}$  is replaced by  $\delta_w$ . Now consider the *relative* displacement  $x_R = x - \delta_w \sin \Omega t$  between the mass and the wall. A free-body diagram yields the equation of motion:

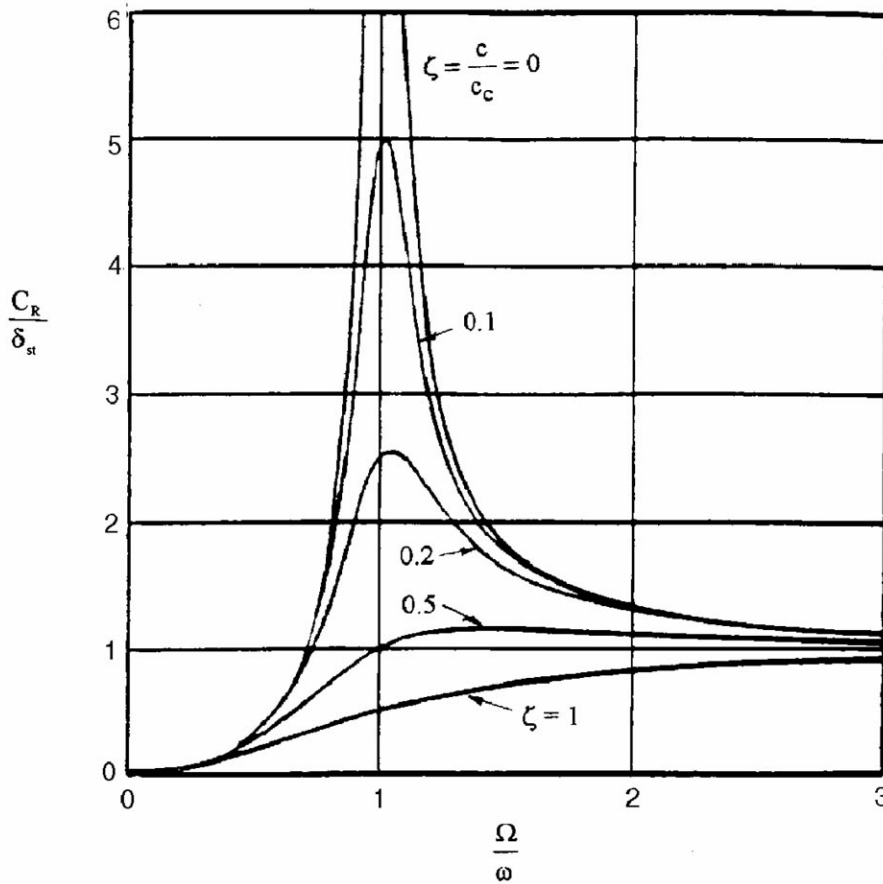
$$m\ddot{x}_R + c\dot{x}_R + kx_R = m\delta_w \sin \Omega t \quad (15.5)$$

which has a steady state solution in the form of Eq. (15.4a). The amplitude of the *relative* motion is found to be:

$$C_R = \frac{\delta_w (\Omega/\omega)^2}{[1 - (\Omega/\omega)^2]^2 + [2\zeta(\Omega/\omega)]^2} \quad (15.6)$$

A graph of the ratio of the amplitudes of relative displacement and wall displacement ( $C_R/\delta_w$ ) is shown in Fig. 15.2. The phase angle lag is the same as in Fig. 15.1. In Fig. 15.2, it is seen that at small excitation frequencies ( $\Omega/\omega$  almost zero), the relative displacement is nearly zero. But, at resonance ( $\Omega/\omega = 1$ ), large relative motion may occur, especially for small damping (small  $\zeta$ ). For  $\Omega/\omega \gg 1$ ,  $C_R/\delta_w$  is nearly unity, and the relative motion is  $180^\circ$  out of phase. This means that while the wall is shaking at a high frequency, the mass barely moves at all. This behavior is important in design when isolation from ground vibration is desired.

**Figure 15.2** Relative displacement resulting from foundation excitation (one DOF).



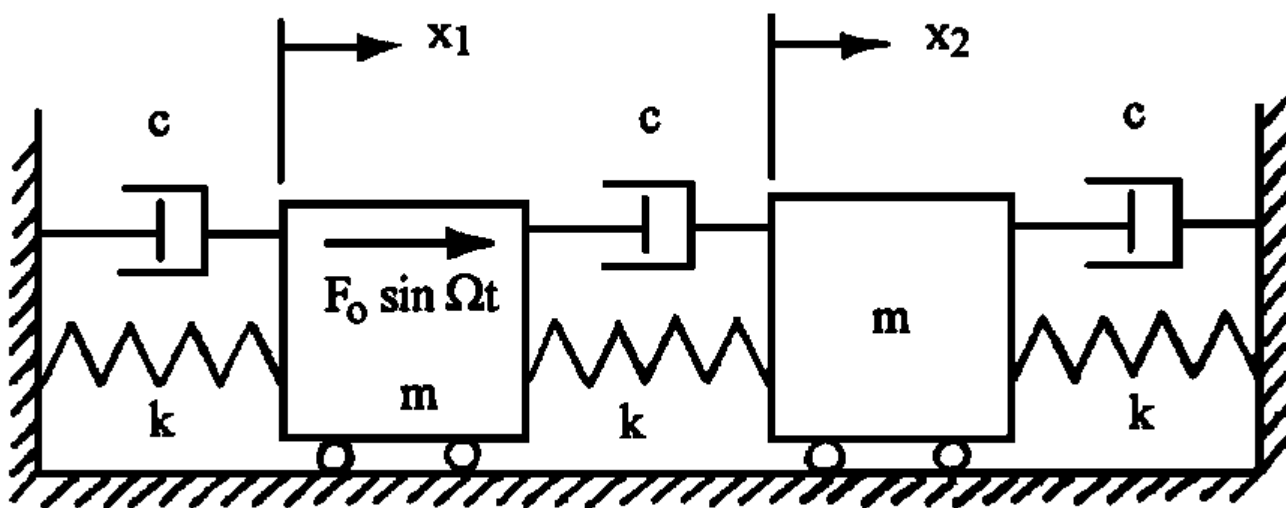
Other types of damping will exist in a typical mechanical system. These include (1) dry friction (e.g., the mass slides on a floor against opposing frictional forces); (2) structural (or material) damping (e.g., the spring material is not perfectly elastic, but dissipates energy during each cycle of vibratory motion); and (3) aerodynamic damping (e.g., the mass vibrates in air, instead of in a vacuum as the previously described models do). These other forms of damping may be approximated by an **equivalent viscous damping** with reasonable accuracy for many of the vibratory characteristics.

## 15.2 Multiple-Degree-of-Freedom Systems

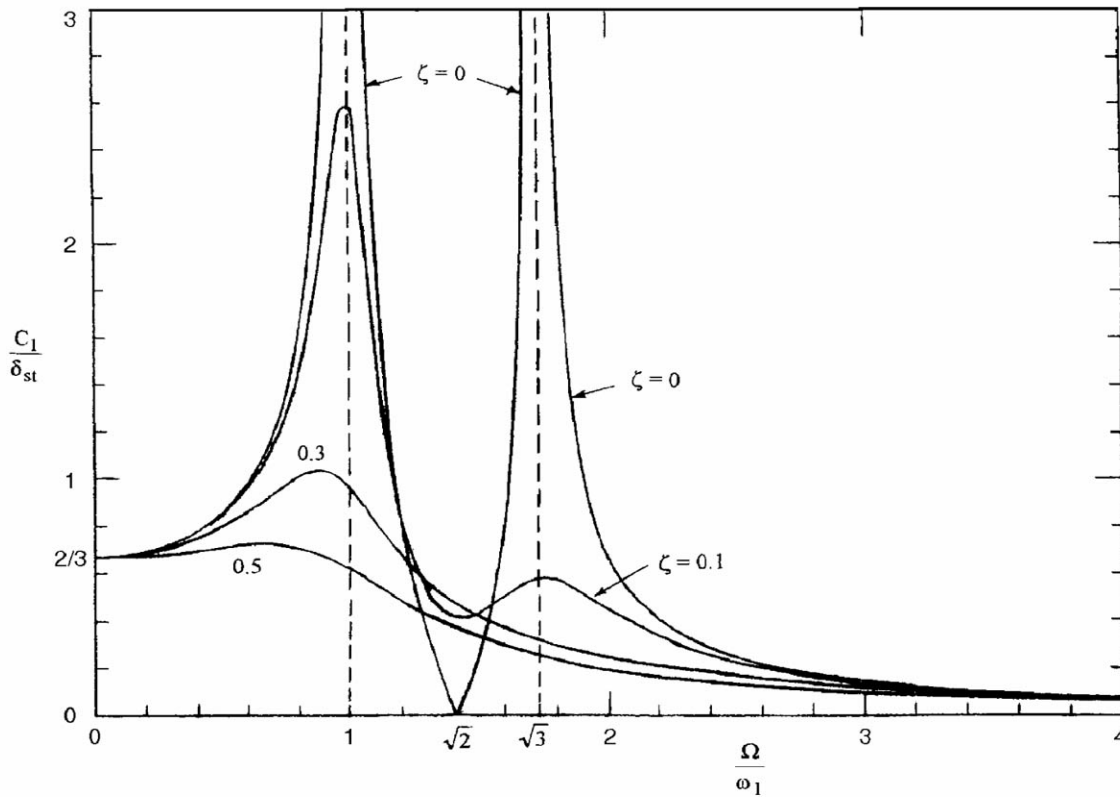
The characteristics described above (vibratory displacement and phase angle) behave similarly for systems having two or more degrees of freedom (DOF). That is, with small damping in the vicinity of a resonant frequency, the steady state displacement amplitude is large and the phase angle changes rapidly with changing  $\Omega$ . The primary difference is that, instead of having a single region of resonance, there are as many regions as there are DOF. A **continuous system** (e.g., string, rod, beam, membrane, plate, shell) has infinite DOF, with an infinite number of free vibration frequencies. Thus, with small damping, large amplitudes can occur in many ranges of exciting frequency for a given exciting force or moment. Fortunately, practical applications show that, typically, only the excitations near the lowest few natural frequencies are significant (although exceptions to this can be shown).

A *two-DOF system* is depicted in Fig. 15.3. The two equal masses are separated by equal springs (stiffnesses  $k$ ) and equal viscous dampers (damping coefficients  $c$ ). A force  $F_o \sin \Omega t$  is applied to one of the masses *only*. From free-body diagrams of each mass, one may obtain two differential equations of motion in the displacements  $x_1$  and  $x_2$ . The equations are coupled because of the spring and mass in the middle. Their steady state solution is found to be sinusoidal in time, with a common frequency  $\Omega$ , but different phase angles for each of the masses. A plot of the amplitude ( $C_1$ ) for the vibratory displacement ( $x_1$ ) of the mass to which the force is applied is seen in Fig. 15.4 for damping ratios  $\zeta = c/2\sqrt{mk} = 0, 0.1, 0.3$ , and  $0.5$ . With no damping, the amplitude is seen to become infinite at the two resonances ( $\Omega/\omega_1 = 1$  and  $\sqrt{3} = 1.732$ , where  $\omega_1 = \sqrt{k/m}$  is the smallest *natural frequency* of the system). Interestingly, for  $\zeta = 0$  and  $\Omega/\omega_1 = \sqrt{2} = 1.414$ , there is *no* motion of the mass to which the force is applied (although the other mass vibrates). This is an example of *vibration isolation*. By adding a second mass to a single-DOF system, the vibratory motion of the first mass may be eliminated at a certain exciting frequency. The added mass need not be equal. With small damping ( $\zeta = 0.1$ ), a large amplitude is observed in Fig. 15.4 at the first resonance, but a smaller one at the second resonance. For larger damping ( $\zeta = 0.3$ ), the second resonant peak essentially vanishes.

**Figure 15.3** Two-DOF mechanical system.

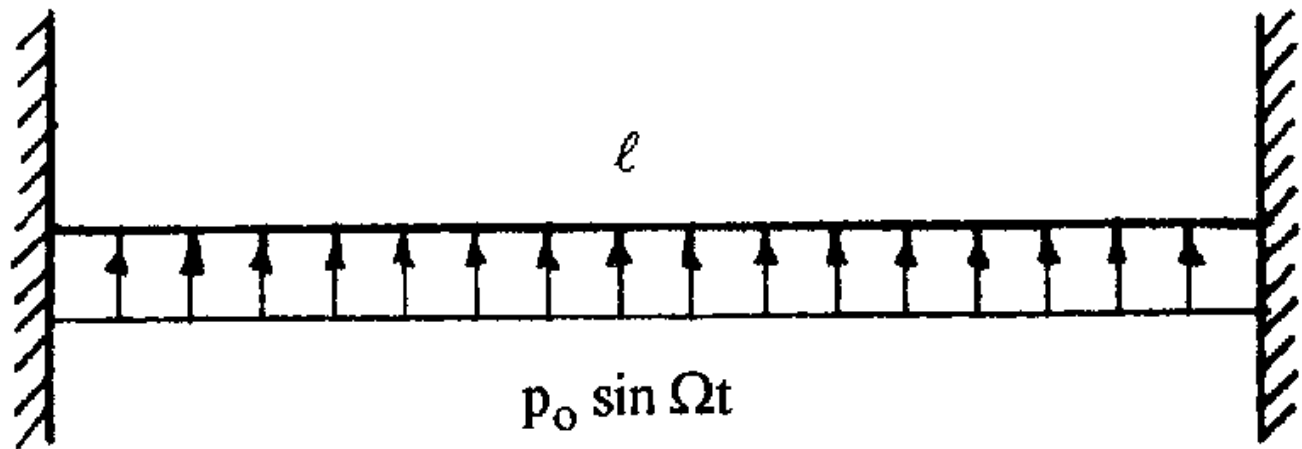


**Figure 15.4** Displacement amplitude versus exciting frequency (two DOF).

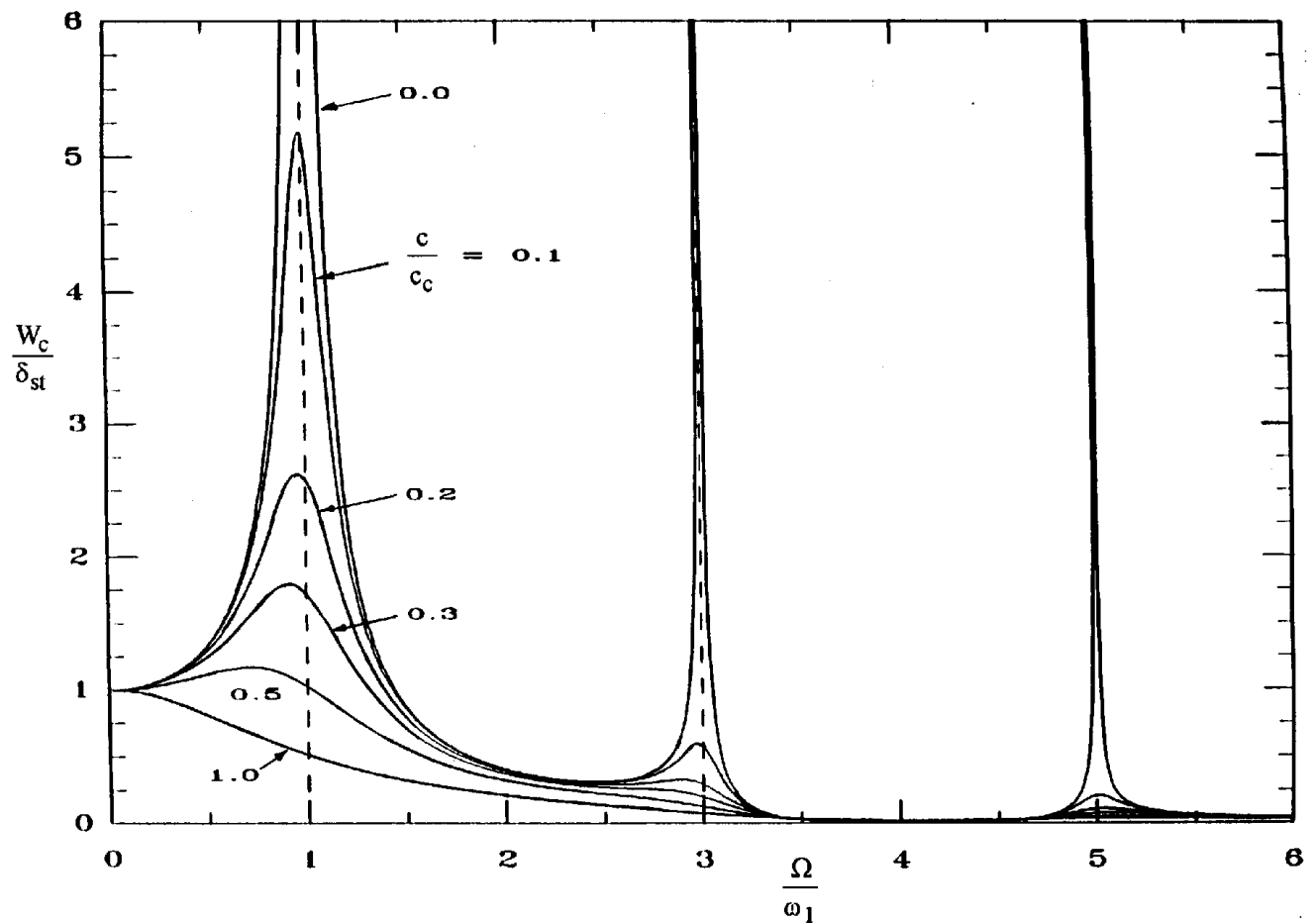


As an example of a **continuous system**, consider a string (or wire) of length  $l$  stretched with a tensile force ( $T$ ) between two rigid walls. It has uniform thickness and mass density ( $\rho$ , mass/length), and negligible bending stiffness. Let the string be subjected to a uniformly distributed loading ( $p$ , force/length) which varies sinusoidally in time ( $p = p_o \sin \Omega t$ ), as shown in Fig. 15.5. Considering only small amplitude transverse vibrations, the equation of motion is found to be a linear, second-order, *partial differential equation*. This may be solved exactly either in closed form or by taking the infinite sum of the displacement responses of the free vibration modes of the system (eigenfunction superposition). The amplitude of the transverse vibration of the center of the string ( $W_c$ ) is observed in Fig. 15.6. It is plotted in the nondimensional form  $W_c/\delta_{st}$ , where  $\delta_{st}$  would be the displacement if the pressure were *static* ( $\delta_{st} = p_o l^2 / 8T$ ). The abscissa is the frequency ratio ( $\Omega/\omega_1$ , where  $\omega_1$  is the first natural frequency of the system). If there is no damping ( $c = 0$ ), then infinite amplitudes (resonances) occur at the first, third, fifth, seventh, and so on natural frequencies. The natural frequencies are  $\omega_m = m\pi/l$ , where  $m = 1, 2, 3, \dots$ . The free vibration mode shapes are symmetric with respect to the center of the string for  $m = 1, 3, 5, \dots$ , and these are the modes that are excited by the symmetric loading. The antisymmetric modes ( $m = 2, 4, 6, \dots$ ) are not excited by it. In the vicinity of each resonance, the mode shape (a sine function along the length) for that natural frequency dominates. Away from resonances (e.g.,  $\Omega/\omega_1 = 2$ ), all symmetric modes are present. The width of each region of resonance decreases as the order of natural frequencies increases. Thus, for example, if  $W_c/\delta_{st}$  is to be less than 3, the range of unacceptable operating frequencies ( $\Omega/\omega_1$ ) is seen to be much smaller at the second resonance than at the first, and smaller yet at the third resonance. For small, uniformly distributed, viscous damping (e.g.,  $c/c_c = 0.1$ , where  $c_c = 2\pi\sqrt{T\rho/l^2}$  is the critical damping coefficient for the *first* mode), the amplitudes at the first three resonances are found to be  $W_c/\delta_{st} = 5.18, 0.59$ , and  $0.21$ .

**Figure 15.5** A string stretched between two walls (continuous system).



**Figure 15.6** Displacement amplitude versus exciting frequency for the vibrating string.





## Defining Terms

**Continuous system:** A system with continuously varying physical parameters (e.g., mass, stiffness, damping), having infinite degrees of freedom; as opposed to a discrete system, which has discontinuous parameters and finite degrees of freedom.

**Equivalent viscous damping:** Viscous damping which would yield a forced vibratory response the same as another form of damping.

**In-phase:** Vibratory displacement which follows in time an exciting force (or displacement).

**Out-of-phase:** Vibratory displacement which is opposite to the direction of the excitation.

**Phase angle:** The angle in a cycle of motion by which a displacement lags behind the exciting force (or displacement).

**Resonance:** Large amplitude motion which occurs when a forcing frequency is in the vicinity of a natural frequency of a system.

**Static deflection:** The limiting case of a forced vibratory displacement, when the exciting frequency is very small, so that dynamic (inertia) effects are negligible.

**Steady state:** The vibratory motion which persists after the transient effects die away or are neglected.

## References

- Den Hartog, J. P. 1956. *Mechanical Vibrations*, 4th ed. McGraw-Hill, New York.
- Leissa, A. W. 1978. On a direct method for analyzing the forced vibrations of continuous systems ha  
*J. Sound Vib.* 56(3):313–324.
- Leissa, A. W. 1989. Closed form exact solutions for the vibrations of continuous systems subjected t  
*J. Sound Vib.* 134(3):435–454.
- Leissa, A. W. and Chern, Y. T. 1992. Approximate analysis of the forced vibrations of plates. *J. Vib.*  
–111.
- Ruzicka, J. E. and Derby, T. F. 1971. *Influence of Damping in Vibration Isolation*. Shock and  
Vibration Information Center, Washington, DC.
- Snowdon, J. C. 1968. *Vibration and Shock in Damped Mechanical Systems*. John Wiley & Sons,  
New York.
- Thomson, W. T. 1988. *Theory of Vibration with Applications*. Prentice Hall, Englewood Cliffs,  
NJ.
- Timoshenko, S. P., Young, D. H., and Weaver, W., Jr. 1974. *Vibration Problems in*  
*Engineering*, 4th ed. John Wiley & Sons, New York.

## Further Information

Discussion of forced vibrations of one-DOF systems, including nonsinusoidal exciting forces, may be found in the excellent textbooks by: Den Hartog; Timoshenko, Young, and Weaver; and, Thomson.

For further information on equivalent viscous damping and its representation of other forms of damping, see the textbook by Thomson. Extensive graphs of amplitude versus frequency ratio for various types of damping are in the monograph by Ruzicka and Derby.

Vibration isolation in a two-DOF system is discussed very well in the textbook by Den Hartog.

Forced vibrations of rods and beams with material damping are thoroughly discussed in the monograph by Snowdon. Closed-form exact solutions for continuous systems [Leissa, 1989] and two useful approximate methods [Leissa, 1978; Leissa and Chern, 1992] are available in individual papers.



Ovunc, B. A. "Lumped versus Distributed Parameter Systems"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Lumped versus Distributed Parameter Systems

---

## 16.1 Procedure of Analysis

### 16.2 Continuous Mass Matrix Method

Member under Axial Displacement • Member under Bending along Its Major Moment of Inertia Axes •  
Dynamic Member Stiffness Matrix for Plane and Space Frames

### 16.3 Consistent Mass Matrix Method

### 16.4 Lumped Mass Matrix Method

### 16.5 Free Vibration of Frames

### 16.6 Forced Vibration

### 16.7 Practical Applications

### 16.8 Structures without Additional Effects

Single Beams • Frames

### 16.9 Structures with Additional effects

Single Beams • Frames

## Bulent A. Ovunc

*University of Southwestern Louisiana*

The lumped, consistent, and distributed (or continuous) mass methods are the main methods for dynamics and vibration analyses of structures. In the continuous mass method the equations of motion are satisfied at every point of the structure. In the consistent and lumped mass methods they are satisfied only at the joints of the structures. In consistent mass the displacements within members are assumed as static displacements. The lumped mass method considers the members as massless springs. The lumped and consistent mass methods are simple and fast; they are fairly approximate, but their accuracy decreases for structures subjected to the effects of the shear and rotatory inertia, member axial force, elastic medium, and so on. The continuous mass method provides accurate results under the assumptions made.

## 16.1 Procedure of Analysis

---

For frames, the general formulation is based on the vector of displacement  $\{a_o(y, t)\}$  at a time  $t$  and at a point  $y$  on the center line of its constitutive members. The same formulation is valid for the above mentioned three methods, only the vector of center line displacement  $\{a_o(y, t)\}$  depends on the assumptions made for each method.

For materially and geometrically linear frames, the vector  $\{a_o(y, t)\}$  is made of four independent components,

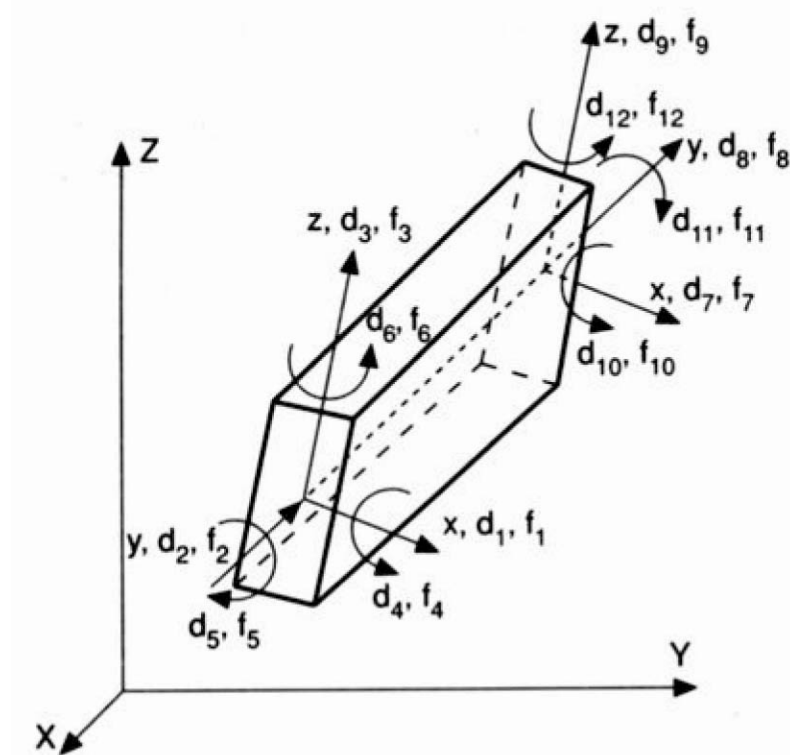
$$\{a_o(y, t)\}^T = [u(y, t) \quad v(y, t) \quad w(y, t) \quad \vartheta(\rho_\Theta, t)] \quad (16.1)$$

where  $u(y, t)$ ,  $v(y, t)$ , and  $w(y, t)$  are the displacements along the member axes  $x, y, z$ ;  $\vartheta(\rho_\Theta, t)$  is the twist rotation about  $y$  axis, and  $\rho_\Theta = (x^2 + y^2)^{1/2}$ . It is assumed that the material properties are independent of time, and that the external disturbances applied to a structure are proportional to a same-time variable function. Thus, the displacement function  $\{a_o(y, t)\}$ , can be written in separable variable form. The integration of the differential equations and elimination of the integration constants gives

$$\{a_o(y, t)\} = [N(y)]\{d(t)\} = [N(y)]\{d\}f(t) \quad (16.2)$$

where  $\{d(t)\} = \{d\}f(t)$ ,  $[N(y)]$ , and  $\{d\}$  are time-independent matrix-of-shape functions and vector-of-member displacements, and  $f(t)$  is a time-variable function of external disturbances. See [Fig. 16.1](#).

**Figure 16.1** Coordinate axes systems.



The strains  $\{\varepsilon\}$  at a point within the cross section of a member can be obtained from the center

line displacements  $\{a_o(y, t)\}$  [Przemieniecki, 1968] as

$$\{\varepsilon\} = [\partial]\{a_o(y, t)\} = [B(y)]\{d\}f(t) \quad (16.3)$$

where  $[B(y)] = [\partial][N(y)]$ , and  $[\partial]$  is the matrix of differential operators.

The stresses  $\{\sigma\}$  are determined from the stress-strain relationship as

$$\{\sigma\} = [E]\{\varepsilon\} = [E][B(y)]\{d\}f(t) \quad (16.4)$$

The expressions of the **strain energy**,  $U_i$ , and **kinetic energy**,  $\mathcal{K}$ , as well as the work done by damping forces,  $W_D$ , and by externally applied loads,  $W_e$ , are written as

$$\begin{aligned} U_i &= \frac{1}{2} \int_v \{\varepsilon\}^T \{\sigma\} dV = \frac{1}{2} (f(t))^2 \{d\}^T \left( \int_v [B(y)]^T [E] [B(y)] dV \right) \{d\} \\ W_D &= - \int_v c \{\dot{a}_o\}^T \{a_o\} dV = -c \dot{f}(t) f(t) \{d\}^T \left( \int_v [N(y)]^T [N(y)] dV \right) \{d\} \\ \mathcal{K} &= \frac{1}{2} \int_v m \{\dot{a}_o\}^T \{\dot{a}_o\} dV = \frac{1}{2} (\dot{f}(t))^2 \{d\}^T \left( m \int_v [N(y)]^T [N(y)] dV \right) \{d\} \\ W_e &= f_e(t) \int_y \{P_o\}^T \{a_o\} dy = f_e(t) \left( \int_y \{P_o\}^T [N(y)] dy \right) \{d\} \end{aligned} \quad (16.5)$$

where  $f_e(t)$  and  $\{P_o\}$  are the time-dependent and -independent parts of the vector of externally applied joint forces.

For a member, stiffness  $[k]$ , mass  $[m]$ , and damping  $[c]$  matrices are determined by substituting the strain energy  $U_i$ , **damping energy**  $W_D$ , kinetic energy  $\mathcal{K}$ , and **external energy**  $W_e$  into the Lagrangian dynamic equation [Ovunc, 1974],

$$\frac{\partial U_i}{\partial d_j} - \frac{\partial W_D}{\partial d_j} + \frac{d}{dt} \left( \frac{\partial \mathcal{K}}{\partial \dot{d}_j} \right) = \frac{\partial W_e}{\partial d_j} \quad (16.6)$$

which provides the equation of forced vibration for a member as

$$[k]\{d\}f(t) - 2\nu_o[m]\{d\}\dot{f}(t) + [m]\{d\}\ddot{f}(t) = \{P_o\}f_e(t) \quad (16.7)$$

where the damping matrix  $[c]$  is assumed to be proportional to mass matrix  $[m]$  and  $\nu_o$  is the damping coefficient.

For the free vibration,  $\{P\} = 0$ , the equation of motion is divided into two parts—time-independent and time-dependent:

$$([k] - \omega^2[m])\{d\} = 0 \quad (16.8)$$

$$\frac{d^2 f(t)}{dt^2} + 2\nu\omega \frac{df(t)}{dt} + \omega^2 f(t) = 0 \quad (16.9)$$

The time-dependent part,  $f(t)$ , Eq. (16.9), is the same for all four independent cases, and  $\nu = \nu_o\omega$  [Paz, 1993].

Although the lumped mass method was developed long before the continuous mass method, the continuous mass method is herein explained first. The consistent and lumped mass methods are presented as particular cases of continuous mass method.

## 16.2 Continuous Mass Matrix Method

For the materially and geometrically linear frames, an arbitrary vibration of its member is obtained by combining the four independent components: axial displacement, torsional rotation, and bending in two orthogonal planes.

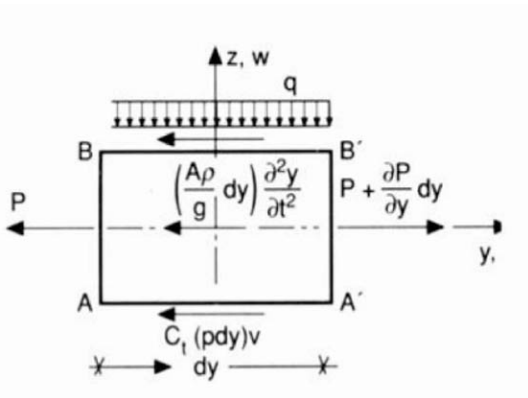
### Member under Axial Displacement

For members with constant section, the time-independent part of the differential equation of free vibration has been given [Ovunc, 1985]. (See Fig. 16.2.)

$$\frac{d^2 Y(y)}{dy^2} + \alpha^2 Y(y) = 0 \quad (16.10)$$

where  $\alpha^2 = (\omega^2 m - C_f p)/EA$  ; and where  $m = (A\rho + q)/g$ ,  $p$ , and  $C_f$  are the mass, the peripheral area, and the friction coefficient of the elastic medium per unit length of the member, respectively.

**Figure 16.2** Axial force member.



The time-independent part,  $f(t)$ , is the same for all four independent cases.

The time-independent part of axial displacement function  $Y(y)$  is obtained by integrating Eq. (16.10), through the elimination of the integration constants  $\{C\}$ , by the boundary conditions; thus one has

$$Y(y) = \{\phi_{ax}(y)\}^T [L]^{-1} \{d_{ax}\} = \{N_{ax}(y)\}^T \{d_{ax}\} \quad (16.11)$$

The nature of the shape function  $\{N_{ax}(y)\}$  depends on the sign of parameter  $\alpha^2$ .

$$\{N_{ax}(y)\}^T = (1/\sin \alpha l) \{(\sin \alpha l \cos \alpha y - \cos \alpha l \sin \alpha y) \quad \sin \alpha y\} \text{ for } \alpha^2 > 0$$

$$\{N_{ax}(y)\}^T = (1/\sinh \alpha l) \{(\sinh \alpha l \cosh \alpha y - \cosh \alpha l \sinh \alpha y) \quad \sinh \alpha y\} \text{ for } \alpha^2 > 0$$

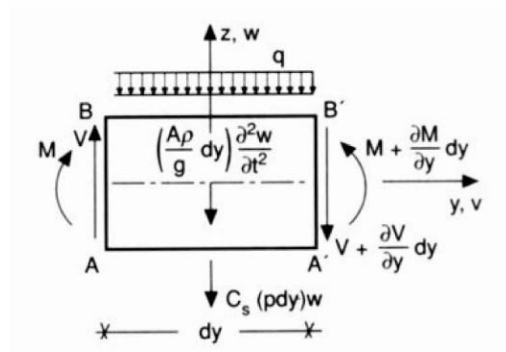
## Member under Bending along Its Major Moment of Inertia Axes

For members with constant section, the time-independent part of the differential equation of free vibration (see Fig. 16.3) has been given by Ovunc [1985] as

$$\frac{d^4 Z}{dy^4} + 2k^2 \frac{d^2 Z}{dy^2} - \beta^4 Z = 0 \quad (16.12)$$

where  $2k^2 = P/EI$ ,  $\beta^4 = (\omega^2 m - C_s p')/EI$ ; and where  $I$ ,  $C_s$ , and where  $p'$  are the moment inertia of the cross section, the subgrade coefficient of the elastic medium, and the projected area of the cross section.

**Figure 16.3** Bending member.



The shape function  $\{N_{bd}(y)\}$  due to deflection,

$$Z(y) = \{\phi_{bd}(y)\}^T [L]^{-1} \{d_{bd}\} = \{N_{bd}(y)\}^T \{d_{bd}\} \quad (16.13)$$

is obtained in a similar manner as in the case of axial displacement, Eq. (16.11). The nature of the shape function  $\{N_{bd}(y)\}$  and its component  $\{\phi_{bd}(y)\}^T$  depend on the parameters  $\alpha_1^2$  and  $\alpha_2^2$ , which are expressed in terms of  $k^2$  and  $\beta^4$ :

$$\alpha_1 = [(\beta^4 + k^4)^{1/2} + k^2]^{1/2}, \quad \alpha_2 = [(\beta^4 + k^4)^{1/2} - k^2]^{1/2}$$

Thus, for  $\beta^4 > 0$  and  $P > 0$  (compression is positive),

$$\{\phi_{bd}(y)\}^T = (\sin \alpha_1 y \quad \cos \alpha_1 y \quad \sinh \alpha_2 y \quad \cosh \alpha_2 y) \quad (16.14)$$

The above expression remains the same when the axial force  $P$  is tension, except  $\alpha_1$  and  $\alpha_2$  must be interchanged. For the combination of  $\beta^4 < 0$  and  $P < 0$  or  $P > 0$ , the expression of  $\{\phi_{bd}(y)\}$  can be determined in a similar manner.

The member stiffness matrices for the twist rotation and the bending in the  $Oxy$  plane are obtained by following similar steps as in the previous cases. The stiffness matrix for the space frame is determined by combining the stiffness matrices of all four independent cases.

## Dynamic Member Stiffness Matrix for Plane and Space Frames

The dynamic member stiffness matrix  $[k_{dyn}]$  for either a plane frame or a space frame is obtained by substituting either the shape functions for axial displacement and bending in  $Oyz$  plane [Eqs. (16.11) and (16.13)] or the shape functions of all the four independent cases, in the Lagrangian dynamic equation [Eq. (16.6)], integrating [Eq. (16.7)] one has

$$[k_{dyn}] = [k] - \omega^2 [m] \quad (16.15)$$

The continuous mass method has also been extended to frames with tapered members [Ovunc, 1990].

## 16.3 Consistent Mass Matrix Method

In the consistent mass matrix method, the deformations within a member are static deformations. The shape function for each independent component  $\{N(y)\}$  is a static displacement due to its corresponding independent cases. Thus, the shape functions for axial displacement,  $\{N_{ax}(y)\}$ , and for bending in the  $Oyz$  plane,  $\{N_{bn}(y)\}$ , are given as

$$\{N_{ax}(y)\}^T = ((1 - \eta) \quad \eta) \quad (16.16)$$

$$\{N_{bn}(y)\}^T = ((1 - 3\eta^2 + 2\eta^3) \quad (\eta - 2\eta^2 + \eta^3)\ell \quad (3\eta^2 - 2\eta^3) \quad (-\eta^2 + \eta^3)\ell) \quad (16.17)$$

where  $\eta = y/\ell$ . The shape functions for twist rotation and bending in the  $Oxy$  plane are obtained in similar manner.

The member stiffness  $[k]$  and mass  $[m]$  matrices are evaluated by substituting the shape functions in the Lagrangian dynamic equation [Eq. (16.6)]. Herein, the stiffness matrix  $[k]$  is a static stiffness matrix and the mass matrix  $[m]$  is a full matrix [Przemieniecki, 1968; Paz, 1993].

Moreover, the member stiffness matrix  $[k]$  and the mass matrix  $[m]$  for the consistent mass matrix method can be obtained as the first three terms of the power series expansion of the dynamic member stiffness matrix  $[k_{dyn}]$  for the continuous mass matrix [Paz, 1993].

## 16.4 Lumped Mass Matrix Method

---

The lumped mass method is obtained from the continuous mass matrix method by considering the limit when the mass of the members tend to zero. Thus, the shape functions [Eq. (16.17)] and the member stiffness matrix  $[k]$  are the same as in the consistent mass method. But the mass matrix  $[m]$  is diagonal [Paz, 1993].

## 16.5 Free Vibration of Frames

---

The **stiffness coefficients** for the frames are evaluated from those of its members as

$$K_{i,j} = \sum k_{r,s}, \quad M_{i,j} = \sum m_{r,s}, \quad P_i = \sum p_r$$

where  $r$  and  $s$  are the member freedom numbers corresponding to the  $i$  and  $j$  of the structure freedoms.

The equation of free vibration is obtained from those of members [Eq. (16.8)]

$$([K] - \omega^2[M])\{D\} = \{0\} \quad (16.18)$$

where  $\{D\}$  is the vector of the structure displacements.

The natural circular frequencies,  $\omega_i$ , and the corresponding modal shapes,  $\{D_i\}$ , are calculated from the equation of free vibration [Eq. (16.18)], also called the *frequency equation*.

## 16.6 Forced Vibration

---

The second-order differential ( $n$  simultaneous equations with  $n$  unknowns)

$$[M]\{\ddot{D}(t)\} + 2\nu_o[M]\{\dot{D}(t)\} + [K]\{D(t)\} = \{P(t)\} = \{P_o\}f_e(t) \quad (16.19)$$



is converted to  $n$  separate, second-order, single-variable differential equations through the two orthogonality conditions, which proves that each modal shape vector  $\{D_i\}$  is independent of the others. The  $i$ th participation factor  $\mathcal{A}_i$  is defined as the component of given forced vibration on the  $i$ th modal shape vector  $\{D_i\}$ . Thus, any arbitrary motion can be determined by considering the summation of its components  $\mathcal{A}_i$  on each modal shape vector  $\{D_i\}$  [Clough and Penzien, 1993; Paz, 1993; Ovunc, 1974].

If the time variable factor  $f_e(t)$  of the external disturbance [Eq. (16.2)] is periodic (pulsating), the forced vibration can be directly determined without the calculation of participation factors  $\mathcal{A}_i$  [Paz, 1993].

## 16.7 Practical Applications

For any structure the natural circular frequency  $\omega$  can be expressed in terms of a parameter  $C$  as [Paz, 1993]

$$\omega = C \sqrt{\frac{EI_j}{m_j \ell_j^4}} \quad (16.20)$$

where  $E$  is the Young's modulus, and  $I_j$ ,  $m_j$ , and  $\ell_j$  are the moment of inertia, mass, and span length of a selected member  $j$ .

Substituting the natural circular frequency  $\omega$  (its expression in terms of  $C$ ) into the frequency equation [Eq. (16.18)] gives

$$|-C^2[M] + [K]| = 0 \quad (16.21)$$

where the general terms  $M_{rs}$  and  $K_{rs}$  of the mass and the stiffness matrix are constant and expressed as

$$M_{rs} = M_{rs}/m_j \ell_j \quad \text{and} \quad K_{rs} = K_{rs} \ell_j / EI_j \quad (16.22)$$

It can be easily seen that the determinant of the frequency equation [Eq. (16.21)] is independent of the member characteristics ( $E$ ,  $m$ ,  $\ell$ ,  $I$ ) but depends on the parameter  $C$ . If, in a frame, the same characteristics of the members are multiplied by the same factor, the magnitude of the parameter  $C$  remains unchanged. However, a natural circular frequency  $w_i$  corresponding to  $C_i$  changes [Eq. (16.20)]. If the characteristics of some members change and those of the others remain constant, the parameter  $C_i$  is affected.

The advantages of one method over the others and the limits on their accuracy depend on the number and type of the members in the structures and whether the structure is subjected to additional effects.

The type of member depends on the ratio of the thickness  $t$  (of its cross section) to its span length  $\ell$  :  $\gamma = t/\ell$ .

- If the ratio  $\gamma = O(1/100)$  is in the order of  $1/100$ , the effect of bending is negligible. The

member is a very thin member, called *cable*.

- If the ratio  $\gamma = O(1/10)$ , axial force, torsion, and bending are affecting the member. The member is a thin member.
- If the ratio  $\gamma = O(1)$ , the member is considered a deep beam.

The inclusion of the variation of the width or the thickness of the member, the effect of member axial force, shear, rotatory inertia, vibration of the member within an elastic medium, and so on constitutes the additional effects.

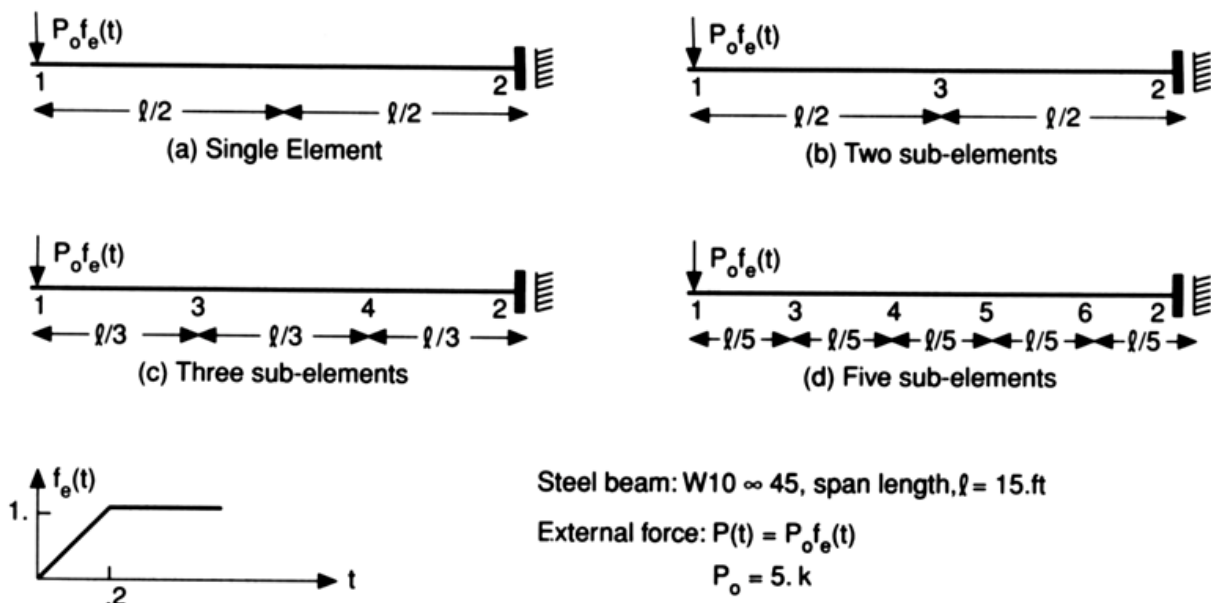
The dynamic responses of beams and frames are evaluated with or without the additional effects by lumped and consistent mass methods. The results are compared with those evaluated by continuous mass.

## 16.8 Structures without Additional Effects

### Single Beams

The data related to a cantilever beam composed of a single element or with two, three, or five subelements are given in Fig. 16.4. The dynamic responses of each beam have been determined by the lumped mass and the consistent mass matrix methods. The first two natural circular frequencies,  $\omega_1$  and  $\omega_2$ , and the vertical displacement  $d_{v11}$ , rotation  $d_{r11}$ , and shearing force  $V_{11}$  at the free end 1 as well as the bending moment  $M_{21}$  at the fixed end 2 due to the first mode  $\omega_1$  are furnished in Table 16.1 for the lumped mass method, for the consistent mass method, and for the continuous mass method.

**Figure 16.4** Cantilever beam formed by (a) a single member, (b) two, (c) three, (d) five subelements and subjected to a force  $P(t) = P_o(t)f_e(t)$ .



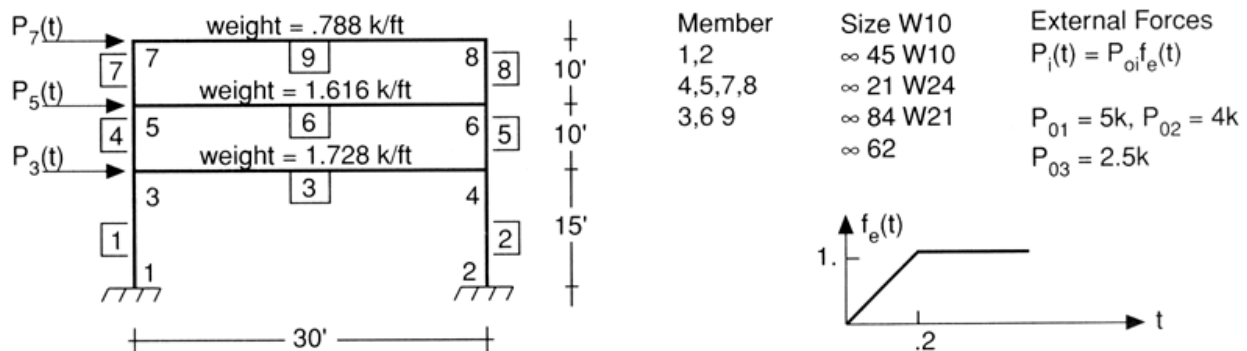
**Table 16.1** Comparison of Analyses

Member	$\omega_1$	$\omega_2$	$d_{v11} \cdot 10^{-1}$	$d_{r11} \cdot 10^{-2}$	$V_{11}$ (k)	$M_{21}$ (k/ft)
Analysis by Lumped Mass Method						
Single	66.2265	0.00	.51534	.0	9.490	71.714
Two	85.3437	439.6537	.39661	-.38314	1.515	30.179
Three	90.4627	510.7243	.37075	-.34954	1.061	29.104
Five	93.3437	560.500	.35681	-.33090	.625	28.548
Analysis by Consistent Mass Method						
Single	95.4062	925.5823	.35050	-.32200	2.000	26.180
Two	94.9687	595.0385	.34934	-.32070	1.317	28.052
Three	94.9687	592.2191	.34882	-.32023	.954	28.177
Five	94.9687	590.7352	.34831	-.32014	.611	28.190
Analysis by Continuous Mass Method						
Single	95.0762	595.8353	.34962	-.32080	1.312	28.025

In the lumped mass method, increasing the number of subelements improves the accuracies of the natural circular frequencies  $\omega_1$ ,  $\omega_2$ , and  $\omega_i$  and those of the vertical displacements,  $d_{v11}$ , and rotations,  $d_{r11}$ , when their magnitudes are compared to their magnitudes obtained by the continuous mass method. For the consistent mass method, the beam has better approximations when it is subdivided into two subelements.

## Frames

Three-story frames have been selected as examples to compare responses obtained by lumped, consistent, and continuous mass methods [Ovunc, 1980]. The data for the three-story frames are given in Fig. 16.5.

**Figure 16.5** Three-story steel frame.

The dynamic responses of the three-story frame for three different ratios of  $\alpha = I_b/I_{bo}$  are selected, where  $I_b$  and  $I_{bo}$  are new and actual moments of inertia of the beams. The moments of inertia  $I_{bo}$  are for thin members with a depth-to-span ratio  $\gamma$  in the order of 1/10. Only the sizes of the beams have been varied.

The frame's first two natural circular frequencies ( $\omega_1, \omega_2$ ), horizontal displacements ( $d_{h71}, d_{h72}$ ) at joint 7, and bending moments ( $M_{11}, M_{12}$ ) at joint 1 are given in Table 16.2.

**Table 16.2** Dynamic Responses of a Three-Story Frame

Method	$\alpha = I_b/I_{bo}$	$\omega_1$	$\omega_2$	$d_{h71}$ (ft)	$d_{h72}$ (ft)	$M_{11}$ (k/ft)	$M_{12}$ (k/ft)
Lumped mass	.01	2.9275	10.9188	.94994	-.02215	163.871	15.2073
	1.00	7.5657	22.3703	.11420	-.00105	81.711	1.2817
	100.00	8.2351	23.7597	.09400	-.00091	78.053	1.2048
Consistent mass	.01	2.9099	10.5168	.95120	-.02279	162.792	15.5520
	1.00	7.5652	22.3942	.11440	-.00106	81.672	1.2886
	100.00	8.2429	23.7951	.09382	-.00092	78.032	1.2148
Continuous mass	.01	2.1738	10.0834	1.19336	-.00236	228.538	39.703
	1.00	7.3964	21.8749	.11410	-.00106	81.849	1.317
	100.00	8.2402	23.7688	.09370	-.00091	78.110	2.216

The responses obtained by lumped mass and consistent mass methods are close to each other for any magnitude of  $\alpha$  (or  $\gamma$ ). However, they are roughly approximate compared to those obtained by the continuous mass method for low order of the ratio  $\alpha = O(.01)$  (or  $\gamma$ )—that is, when the beams are very thin. The responses obtained by lumped mass, consistent mass, and continuous mass are very close for the actual or higher order of the ratio  $\alpha \leq O(1)$  (or  $\gamma$ )—that is, for thin and deep beams.

## 16.9 Structures with Additional Effects

### Single Beams

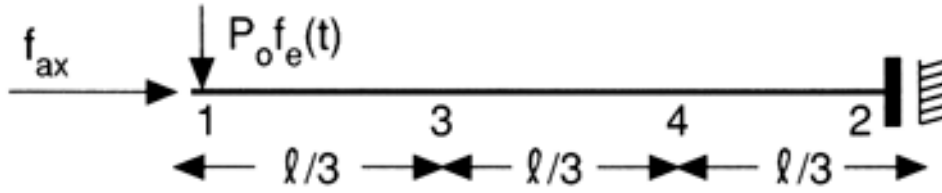
The cantilever beam subdivided into three subelements is subjected to a static axial force  $f_{ax}$ , at its free end 1 (Fig. 16.6). The effect of member axial force  $f_{ax}$  appears in the equations of free vibration as an additional matrix— $[N_{LM}]$  and  $[N_{CS}]$  for the lumped and consistent mass methods [Eq. (16.18)]:

$$([K] + [N_{LM}] - \omega^2[M_{LM}])\{D\} = \{0\} \quad (16.23)$$

$$([K] + [N_{CS}] - \omega^2[M_{CS}])\{D\} = \{0\} \quad (16.24)$$

In the continuous mass method, the member axial force  $f_{ax}$  appears in the argument of the trigonometric or hyperbolic functions [Eq. (16.14)].

**Figure 16.6** Beam subjected to axial force.



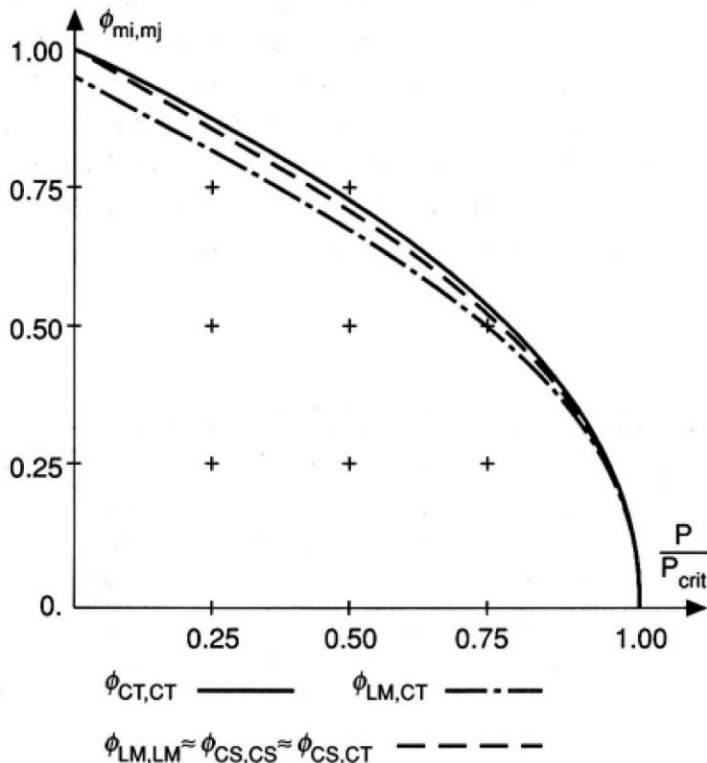
The dynamic responses of the cantilever beam (Fig. 16.6) have been computed by the lumped and consistent mass method by only changing the magnitude of the member axial force  $f_{ax}$  from zero to its critical value  $(f_{ax})_{crit}$  [Eqs. (16.23, 16.24)]. The same computations have been performed by using the continuous mass method [Eq. (16.14)].

The ratios for  $\phi$ ,

$$\phi_{mi,mj} = \omega_{1,mi} / \omega_{01,mj} \quad (16.25)$$

of the first natural frequency by method  $mi$ , (with the effect of member axial force) versus that of method  $mj$  (without the effect of member axial force) are plotted in Fig. 16.7. The index  $mi$  or  $mj$  designates  $LM$ ,  $CS$ , and  $CT$ —the lumped mass, consistent mass, and continuous mass methods.

**Figure 16.7** Effect of member axial force.



The comparison of the variations of the ratios  $\phi_{LM,LM}$  with  $\phi_{LM,CT}$  [Eq. (16.24)] exhibits rough approximation. The approximation involved in the variations of ratios  $\phi_{CS,CS}$  and  $\phi_{CS,CT}$  is very close to the actual one.

The comparison of the variations of the ratios  $\phi_{LM,CT}$  and  $\phi_{CS,CT}$  with  $\phi_{CT,CT}$  exhibits some degree of approximation.

## Frames

All the columns of the three-story steel frame are assumed to be subjected to a static axial force,  $f_{ax}$ , of the same magnitude.

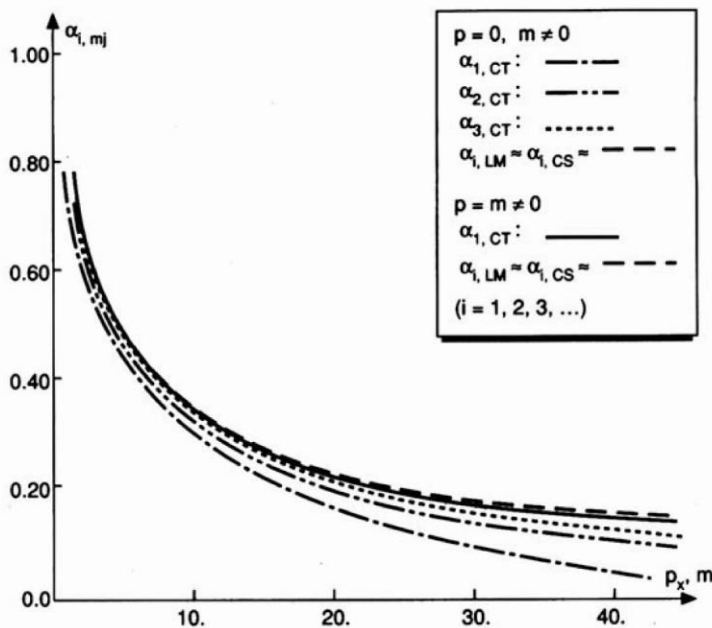
The dynamic responses of the three-story frame were evaluated by lumped, consistent, and continuous mass methods. Two different cases were considered. In the first case, only the magnitudes of the weights acting on the beams have been increased by a factor  $m$ . In the second case, the magnitudes of the member axial force,  $f_{ax}$ , and the weights acting on each beam have been increased by  $p_x$  and  $m$ , in such a way that both factors have the same magnitude,  $p_x = m$ .

For a same-method  $m_j$ , the ratio

$$\alpha_{i,mj} = (\omega_i / \omega_{oi})_{mj} \quad (16.26)$$

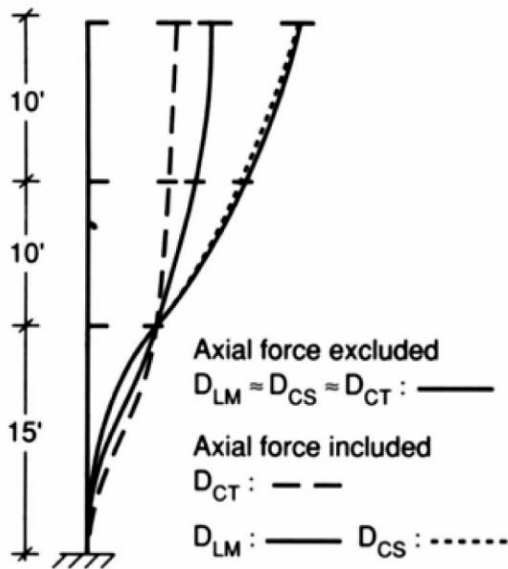
of the  $i$ th natural frequency  $\omega_i$ , (including the effect of member axial force and/or additional mass, only on the beams) versus the  $i$ th natural frequency  $\omega_{oi}$  (excluding all the additional effect) is plotted in Fig. 16.8. The ratio  $\alpha_{i,CT}$ —for first, second and third natural frequencies computed by continuous mass method—is also shown in Fig. 16.8.

**Figure 16.8** Additional effects on members.



The sways at the floor level  $D_{mj}$ , including and excluding the effect of the axial force  $f_{ax}$ , are computed by the lumped, consistent, and continuous mass matrix methods. The variations of the sway at the floor levels  $D_{mj}$  are plotted in Fig. 16.9.

**Figure 16.9** Sways at floor levels.



When the effects of member axial forces are excluded, the sways at the floor levels obtained by lumped, consistent, and continuous mass methods are almost the same. The effect of member axial force has shown small variations in the floor sways evaluated by lumped and consistent mass methods. But the variation in the sways at the floor levels computed by continuous mass method is large.

Although the first buckling mode for lumped and consistent mass methods occurs by increasing sways from lower to upper floors, for the continuous mass method the first buckling mode occurs between the base and the first floor. The relative displacement of second and third floors with respect to the displacement of the first floor tends to zero.

## Defining Terms

**Damping:** Results from the internal friction within the material or from system vibration within another material.

**Damping energy:** Work done by the internal friction within the material as a result of the motion.

**Kinetic energy:** Work done by a mass particle as a result of its motion.

**Stiffness coefficient  $K_{i,j}$ :** Force or moment in the direction of the first index ( $i$ ) required to maintain the equilibrium of the body due to a unit displacement or rotation in the direction of the second index ( $j$ ), while all the other specified displacements and rotations are equal to zero.

**Strain energy:** Work done by a particle due to its stress and strain.

**External energy:** Work done by an external force due to a displacement in its direction.

## References

- Clough, R. W. and Penzien, J. 1993. *Dynamic of Structures*. McGraw-Hill, New York.
- Ovunc, B. A. 1974. Dynamics of frameworks by continuous mass method. *Compt. Struct.* 4:1061–1089.
- Ovunc, B. A. 1980. Effect of axial force on framework dynamics. *Compt. Struct.* 11:389–395.
- Ovunc, B. A. 1985. Soil–Structure interaction and effect of axial force on the dynamics of offshore structures. *Compt. Struct.* 21:629–637.
- Ovunc, B. A. 1990. *Free and Forced Vibration of Frameworks with Tapered Members*, Struceng & –18, p. 341–346.
- Paz, M. 1993. *Structural Dynamics, Theory and Computations*. Van Nostrand Reinhold, New York.
- Przemieniecki, J. S. 1968. *Theory of Matrix Structural Analysis*. McGraw-Hill, New York.

## Further Information

- Paz, M. 1986. *Microcomputer Aided Engineering: Structural Dynamics*, Van Nostrand Reinhold, New York.
- Ovunc, B. A. 1972. The dynamic analysis of space frameworks by frequency dependent stiffness ma *International Association for Bridges and Structural Engineering*, vol. 32/2, Zurich, Switzerland –154.
- Ovunc, B. A. 1986. Offshore platforms subjected to wave forces. In *Recent Applications in Comput* –169.
- Ovunc, B. A. 1985. STDYNL, a code for structural systems. In *Structural Analysis Systems* (ed. Nik –238.
- Ovunc, B. A. 1992. *Dynamics of Offshore Structures Supported on Piles in Cohesionless Soil*. ASM –July 4, ASME PD, vol. 47-5, p. 11–18.
- Ovunc, B. A. 1990. *Vibration of Timoshenko Frames Including Member Axial Force and Soil-Struc* –6, p. 359–364.



Kalinowski, A. J. “Applications of Structural and Dynamic Principles”  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Applications of Structural and Dynamic Principles

---

## 17.1 Base Configuration Loaded Applications

Problem 1: Vehicle Suspension • Problem 2: Shock Isolation of Fragile Equipment

## 17.2 Structural Configuration Loaded Applications

Problem 3: Rotating Machinery Force Transmission • Problem 4: Free-Fall Shock

## 17.3 Additional Information

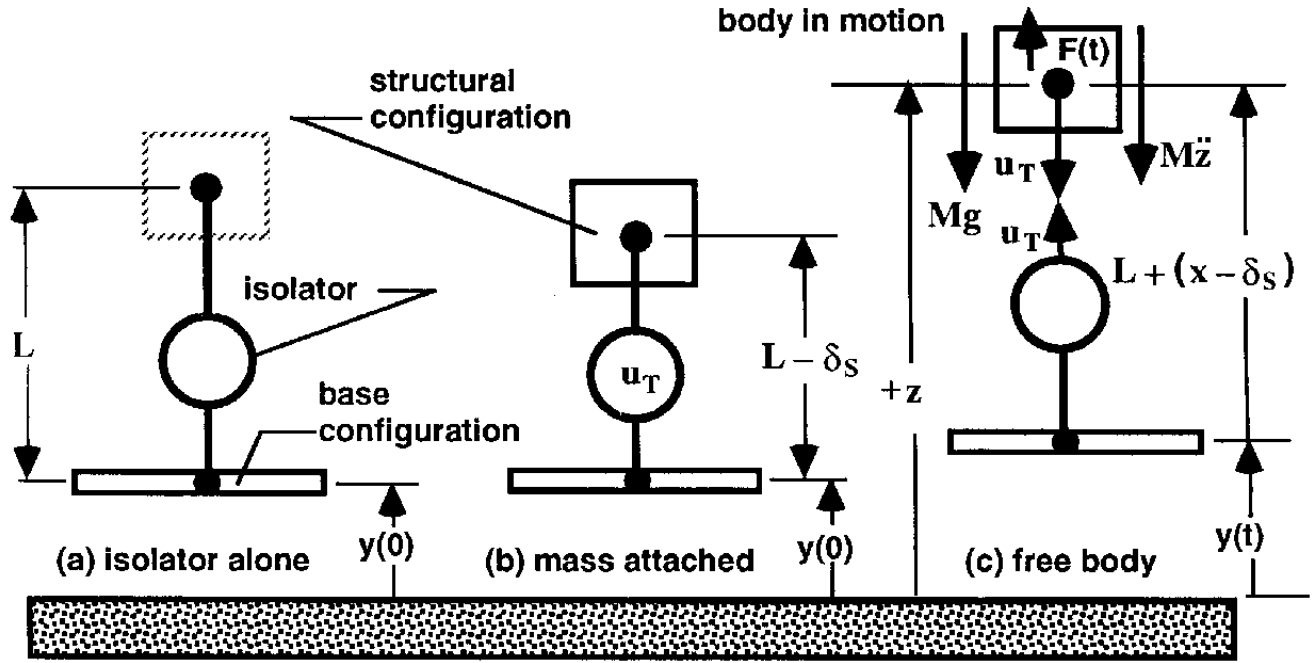
**Anthony J. Kalinowski**

*Naval Undersea Warfare Center*

In this chapter we will consider some practical applications of dynamics and vibrations, with specific emphasis on shock isolation and on vibration isolation. In both of these isolation situations, we are concerned with the transmission of interaction forces,  $u(t)$ , existing between two configurations [which are referred to here as the **base configuration (BC)** and the **structural configuration (SC)**], as illustrated in Fig. 17.1. This example system is a first-order representation of an idealized physical system that is general enough to represent *both* shock and vibration design situations. Most of the underlying physical principles impacting the design of either kind of isolation can be explained and illustrated with this simple one-degree-of-freedom model. The governing equations of motion can be generated from the three-part sequence shown in Fig. 17.1. In Fig. 17.1(a), the BC and SC are in the unloaded condition, and the two configurations are initially separated by a length  $L$ . This corresponds to the state where the model is lying horizontal relative to the vertical direction of the gravity field, or where the model is in the vertical position *before* the gravity field (acceleration of gravity  $g = 386 \text{ in./s}^2$ ) is taken into account. For vertically oriented models, it is convenient to write the equation of motion relative to an initially gravity-loaded model, as in Fig. 17.1(b), where the SC is shown at rest in a compressed state (the linear isolator has a preloaded compressive force of  $u_T = -K\delta_s$ , where  $\delta_s = Mg/K$  is defined *positively* as the **static deflection**). Next, Fig. 17.1(c) corresponds to a body in motion, and the governing equations of motion are obtained by constructing a free-body diagram of the SC and equating the sum of all vertical forces acting on the body to its mass,  $M$ , times its acceleration,  $d^2z/dt^2 \equiv \ddot{z}$  (where dot notation is used to refer to time differentiation from here on), resulting in the following dynamic equilibrium equation:

$$M\ddot{z} + u_T(x, \dot{x}) + Mg - F(t) = 0 \quad (17.1)$$

**Figure 17.1** Multipurpose single-degree-of-freedom system.



The relation between the total  $z$  displacement and the isolator stretch variable  $x$  is given by

$$z = y + x + (L - \delta_s) \quad (17.2)$$

and it follows that

$$\ddot{z} = \ddot{x} + \ddot{y} \quad (17.3)$$

In general, the isolator force,  $u_T(x, \dot{x})$ , could be a nonlinear function of the relative displacement  $x$  and relative velocity  $\dot{x}$ ; however, for the purposes of this introductory development, only linear isolators will be considered. Thus, the total isolator force acting on the mass  $M$  in Fig. 17.1(c) is given by

$$u_T(x, \dot{x}) = Kx + C\dot{x} - K\delta_s \equiv u(x, \dot{x}) - K\delta_s \quad (17.4)$$

where  $u(x, \dot{x})$  is the dynamic portion of the isolator force not including the static deflection force. Substituting Eqs. (17.2)–(17.4) into Eq. (17.1) gives

$$M\ddot{x} + C\dot{x} + Kx = F(t) - M\ddot{y}(t) \quad (17.5)$$

subject to initial conditions

$$x(t = 0) = x_0, \quad \dot{x}(t = 0) = \dot{x}_0 \quad (17.6)$$

It is noted that the  $Mg$  term and  $-K\delta_s$  cancel in the formation of Eq. (17.5).

The shock and vibration will take place about the static equilibrium position shown in Fig. 17.1(b); that is, when the vibrating body comes to rest (i.e.,  $x = 0$ ,  $\dot{x} = 0$ ), the isolator is still compressed an amount equal to the static deflection  $\delta_s$ . The general solution to Eq. (17.5) subject to initial conditions (17.6) can be obtained by several different methods; however, the method of Laplace transforms is used here because it readily applies to situations where (1) the boundary conditions are of the initial-value type, (2) the right-hand side is an arbitrary function of time, and (3) an equivalence between an impulse-loaded right-hand side (rapidly applied loading) and a suddenly applied initial-velocity problem with no right-hand side can be easily illustrated. Thus, taking the Laplace transform of Eq. (17.5) with respect to the Laplace transform variable  $s$  results in

$$x(s) = \frac{x_0 \cdot (sM + C)}{(s^2M + sC + K)} + \frac{\dot{x}_0 \cdot (M)}{(s^2M + sC + K)} + \frac{F(s) - M\ddot{y}(s)}{(s^2M + sC + K)} \quad (17.7)$$

Upon taking the inverse transform, this leads to the general solution for displacement and velocity:

$$x(t) = e^{-\eta t} \left( x_0 \cos(\omega_d t) + \frac{\dot{x}_0 + \eta x_0}{\omega_d} \sin(\omega_d t) \right) + \int_{\lambda=0}^{\lambda=t} \frac{e^{-\eta(t-\lambda)} \sin[\omega_d(t-\lambda)][F(\lambda) - M\ddot{y}(\lambda)]d\lambda}{M\omega_d} \quad (17.8)$$

$$\dot{x}(t) = -\eta x(t) + e^{-\eta t} [-x_0\omega_d \sin(\omega_d t) + (\dot{x}_0 + \eta x_0) \cos(\omega_d t)] + \int_{\lambda=0}^{\lambda=t} \frac{e^{-\eta(t-\lambda)} \cos[\omega_d(t-\lambda)][F(\lambda) - M\ddot{y}(\lambda)]d\lambda}{M} \quad (17.9)$$

In the applications to follow, it will be more convenient to work with the variables  $\omega_n \equiv \sqrt{K/M} = 2\pi f_n$ , corresponding to the **isolator natural frequency**, and  $\zeta \equiv C/(4\pi M f_n)$ , corresponding to the **critical damping ratio**. The variables  $\omega_d$  and  $\eta$  appearing in Eqs. (17.8) and (17.9) are called the **damped natural frequency** and **decay constant**, respectively, and can be expressed in terms of variables  $\omega_n$  and  $\zeta$  using  $\omega_d = \omega_n \sqrt{1 - \zeta^2}$  and  $\eta = \omega_n \zeta$ . These new variables have the following physical meanings:  $\omega_n$  corresponds to the free harmonic vibration (no driver present) of the isolator in the absence of damping;  $\omega_d$  corresponds to the damped free harmonic vibration;  $\zeta$  determines whether the system is *underdamped* ( $\zeta < 1$ ) or *overdamped* ( $\zeta > 1$ ), where in the former case the free motion oscillates harmonically with damped natural frequency  $\omega_d$  and in the latter case the free motion does not vibrate harmonically; and finally,  $\eta$  corresponds to the rate at which the underdamped system exponentially decays in time.

The solutions represented by Eqs. (17.8) and (17.9) will be used to evaluate the dynamic

responses in all the example problems to follow. The forms of the solutions are general and apply to either the situation where the drivers  $[F(t), \ddot{y}(t)]$  are given as an analytical expression or as a digital representation (e.g., earthquake responses). In the case of digital driver representations, the integrations in Eqs. (17.8) and (17.9) can easily be performed by numerical integration (e.g., Simpson's rule), and in the case of analytical driver representations, closed-form integrals can be obtained with the aid of integral tables or with the aid of symbolic evaluation packages such as Maple [Redfern, 1994], MATLAB's version of Maple [Sigmon, 1994], and Mathematica [Wolfram, 1991]. A computer program using MATLAB script language [Math Works, 1992] was used to generate the results presented here.

Given the displacement and velocity versus time from Eqs. (17.8) and (17.9), back-substituting  $x(t)$  and  $\dot{x}(t)$  into Eq. (17.4) gives the dynamic portion of the isolator force  $u(x, \dot{x})$ , which serves as the major ingredient for isolating the structural configuration from the base configuration. The total acceleration  $\ddot{z}$  of the SC can be obtained by substitution of  $u_T(x, \dot{x})$  into Eq. (17.1). We will refer to the single-degree-of-freedom model in Fig. 17.1(c) for the example problems considered here, and the physical significance of the ingredients of the model will be different for each usage. We will consider two types of applications:

*Base configuration loaded:* Here the base configuration has a prescribed displacement motion time history,  $y(t)$  (and therefore the base acceleration  $d^2y/dt^2 \equiv \ddot{y}$ ) is the basic input load, and the structural configuration has a zero-value *external forcing function*  $[F(t) = 0]$ . Some examples of this type problem are (1) *earthquake-resistant structures*, where the ground (base configuration) movement from fault slip motion excites buildings or bridges (structural configuration); (2) *vehicle suspension*, where the ground displacement road profile  $[y(\xi) \text{ vs. distance, } \xi, \text{ in the forward direction of the vehicle traveling at constant velocity } V \text{ as it moves parallel to the ground}]$  results in the ground acting as the equivalent base configuration, with prescribed displacement road profile shape  $y(\xi) = y(Vt)$  and corresponding *apparent base acceleration*  $d^2y/dt^2 = V^2 d^2y/d\xi^2$  measured perpendicular to the ground, and the vehicle car body (structural configuration) responds to these irregularities in the road; and (3) *electronic component isolation*, where the ship superstructure (base configuration) cabinet houses electronic components mounted in cabinets (structural configuration) that are subject to water-borne shock waves that impart known base motion accelerations,  $\ddot{y}$  (based on previously measured experimental data).

*Structural configuration loaded:* Here the structural configuration has a directly applied force time history,  $F(t)$ , and the base configuration has no motion (i.e.,  $y(t) = 0.0$ ). Some examples of this type of problem are: (1) *unbalanced rotating mass*, where an eccentric mass  $m_e$  with offset  $r_e$  is rotating at constant angular speed  $\omega$  (simple motor model with an unbalanced offset mass  $m_e$ ) and is mounted inside the motor housing (structural configuration) and the base configuration is taken as fixed ( $y(t) = 0$ ), where the explicit harmonic forcing function is  $F(t) = m_e r_e \omega^2 \cos(\omega t)$ ; and (2) *free-falling mass*, where a free-falling mass (structural configuration) is prevented from impacting a rigid surface (base configuration) by having an intermediate shock isolator break the fall of the object dropped from height  $H$  (e.g., a stunt motorcycle jumps off a ramp or a package containing fragile equipment drops), where the force on the structural configuration is  $F = -mg$  for  $t \geq 0$  and  $F = 0$  and  $t < 0$ , with initial

condition  $dx/dt = -\sqrt{2gH}$  at  $t = 0$  and  $x = 0$  at  $t = 0$ .

## 17.1 Base Configuration Loaded Applications

### Problem 1: Vehicle Suspension

In this example, a vehicle moving with horizontal constant velocity  $V$  passes over a roadway having a ground profile of  $y(\xi) = Y_0 \sin(k\xi)$ ,  $0 \leq \xi \leq N_c L$ ,  $k = 2\pi/L$ , where  $L$  is the period of the ground swell. Since the relation between horizontal distance traveled and time is given by  $\xi = Vt$ , the vertical BC base motion can be rewritten as

$$y(t) = Y_0 \sin(\omega t), \quad \ddot{y}(t) = -\omega^2 Y_0 \sin(\omega t), \quad \omega = 2\pi V/L \quad (17.10)$$

applied over the range  $0 \leq t \leq N_c L/V$  and  $\ddot{y}(t) = 0$  over the rest of the time duration. If the vehicle is four-wheeled, then a higher-degree-of-freedom model is needed to represent the response due to the rotational degrees of freedom; therefore, it is assumed that the vehicle is two-wheeled and is being towed while the load is balanced over the axle. Assume that all other forces acting are negligible; therefore,  $F(t) = 0.0$ . Consider a vehicle weighing  $W = 2000$  lb and traveling at a speed of  $V = 60$  mph that encounters a two-cycle ( $N_c = 2$ ) road swell of amplitude  $Y_0 = -1.01$  in. and period  $L = 20$  ft. Substituting the data into Eq. (17.10) results in a peak base acceleration of  $\ddot{y}_{\max} = A_0 = 2$  g's and a drive frequency of  $f = \omega/2\pi = 4.4$  Hz. Design a spring-damper isolator (i.e., find  $K$  and  $C$ ) that limits the vehicle (structural configuration) steady state vibration **acceleration transmission ratio**,  $T_A = \text{peak}|\ddot{z}/\ddot{y}_{\max}|$  to a value of 0.4 and limits the maximum relative displacement of the isolator  $x_{\max} = \pm \frac{3}{4}\delta_s$ .

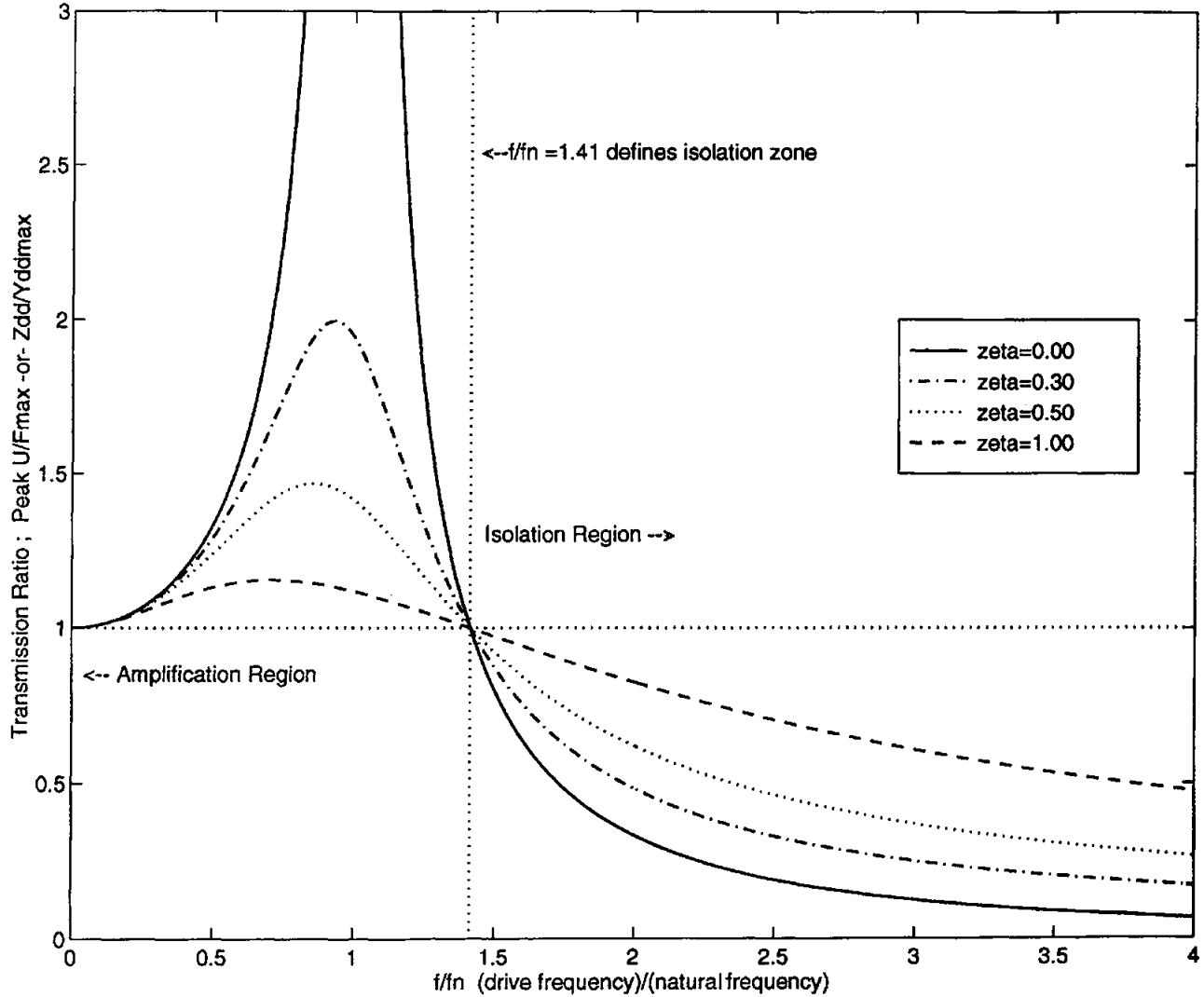
Before solving for  $K$  and  $C$ , it must be noted that isolator manufacturers often use a terminology other than explicitly stating these constants. Typically, in place of  $K$  and  $C$ , a natural frequency ( $f_n$ ) versus load ( $Mg$ ) curve is supplied and the acceleration transmission ratio  $T_A$  is given at resonance, as in [Barry 1993]. Later it will be shown how  $K$  and  $C$  can be back-calculated from  $f_n$  and  $T_A$ . For the present, however, attention is focused on finding the desired  $f_n$  as the first step. The key ingredient in this approach is to use the transmission ratio versus drive frequency ( $\beta \equiv f/f_n$ ) curve as shown in Fig. 17.2. This curve is obtained by substituting a solution of the form  $x(t) = \bar{A} \sin(\omega t) + \bar{B} \cos(\omega t)$  into Eq. (17.5), and solving for the  $\bar{A}$ ,  $\bar{B}$  constants, resulting in the following acceleration transmission ratio:

$$T_A = \sqrt{\frac{1 + (2\zeta\beta)^2}{(1 - \beta^2)^2 + (2\zeta\beta)^2}}, \quad \beta \equiv \omega/\omega_n = f/f_n \quad (17.11)$$

It is of interest to note that all the curves in Fig. 17.2 pass through the same frequency ratio,  $\beta = \sqrt{2}$ , and this special value forms the dividing line between isolation and amplification. Therefore, as a design strategy, in order to get isolation of the base configuration from the structural configuration, the isolator  $K$  is selected such that its natural frequency  $f_n$  results in  $\beta$

values  $> \sqrt{2}$ .

**Figure 17.2** Transmission for harmonic structure force or base acceleration.

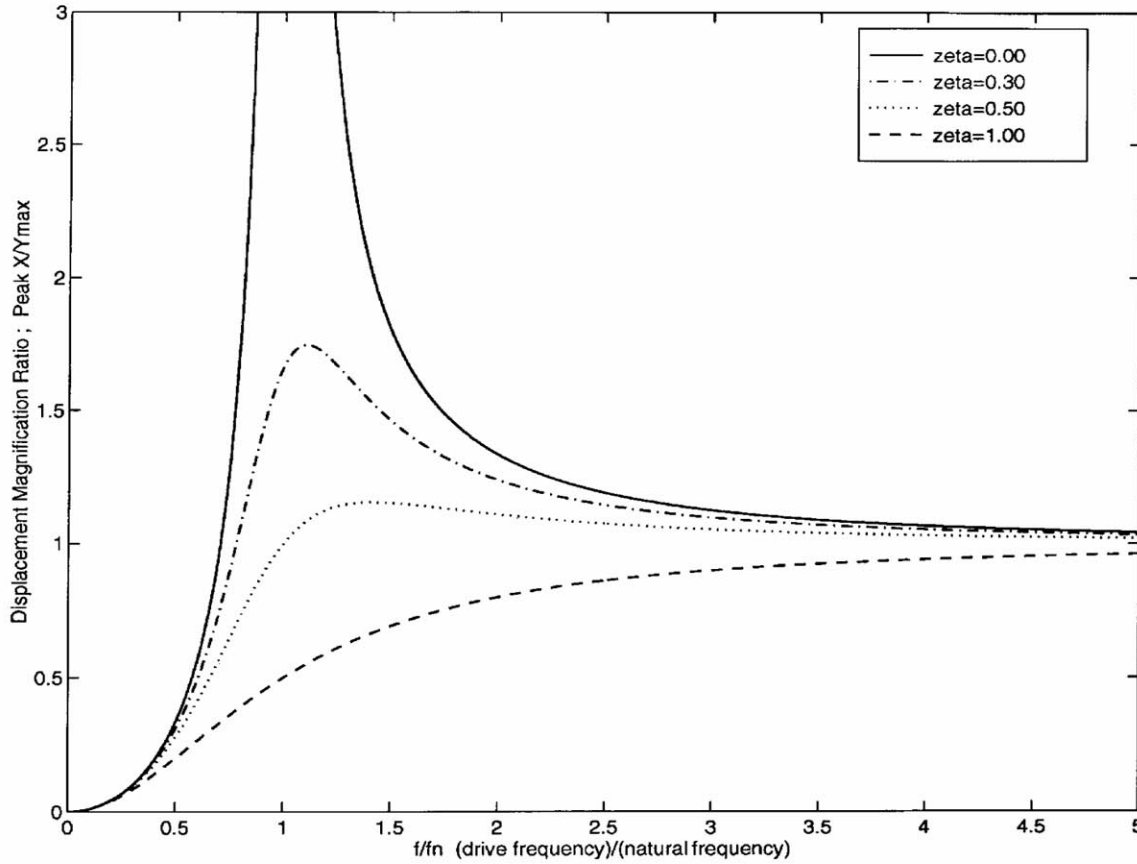


Another response ratio that applies when the relative displacements are of particular interest is the **displacement magnification ratio**, denoted as  $T_D = \text{peak}|x/y_{\max}|$ , which can be derived in a manner similar to the above relation, resulting in the expression

$$T_D = \frac{\beta^2}{\sqrt{(1 - \beta^2)^2 + (2\zeta\beta)^2}} \quad (17.12)$$

This is shown in [Fig. 17.3](#) plotted against the drive frequency parameter  $\beta$ .

**Figure 17.3** Magnification for harmonic base acceleration.



At this point the designer must decide whether to solve for the natural frequency  $f_n$  that gives the desired transmission ratio  $T_A$ , and live with the resulting displacement, *or* solve for the natural frequency  $f_n$  that gives the displacement magnification ratio  $T_D$ , and live with the resulting maximum acceleration  $\ddot{z}$ . It is noted that solving for  $\beta$  is equivalent to solving for  $f_n$ , since the drive frequency,  $f$ , is known. In this sample problem, it is decided that reducing peak acceleration has priority; therefore, with the aid of Eq. (17.11), we solve for the  $\beta$  that gives the desired  $T_A$ . Thus,

$$\beta = \sqrt{\frac{(a + 2T_A^2 - aT_A^2) + \sqrt{(a + 2T_A^2 - aT_A^2)^2 - 4T_A^2(T_A^2 - 1)}}{2T_A^2}}, \quad a = (2\zeta)^2 \quad (17.13)$$

where for light damping ( $\zeta < 0.05$ ),  $\beta$  can be approximated with

$$\beta \approx \sqrt{1 + 1/T_A} \quad (17.14)$$

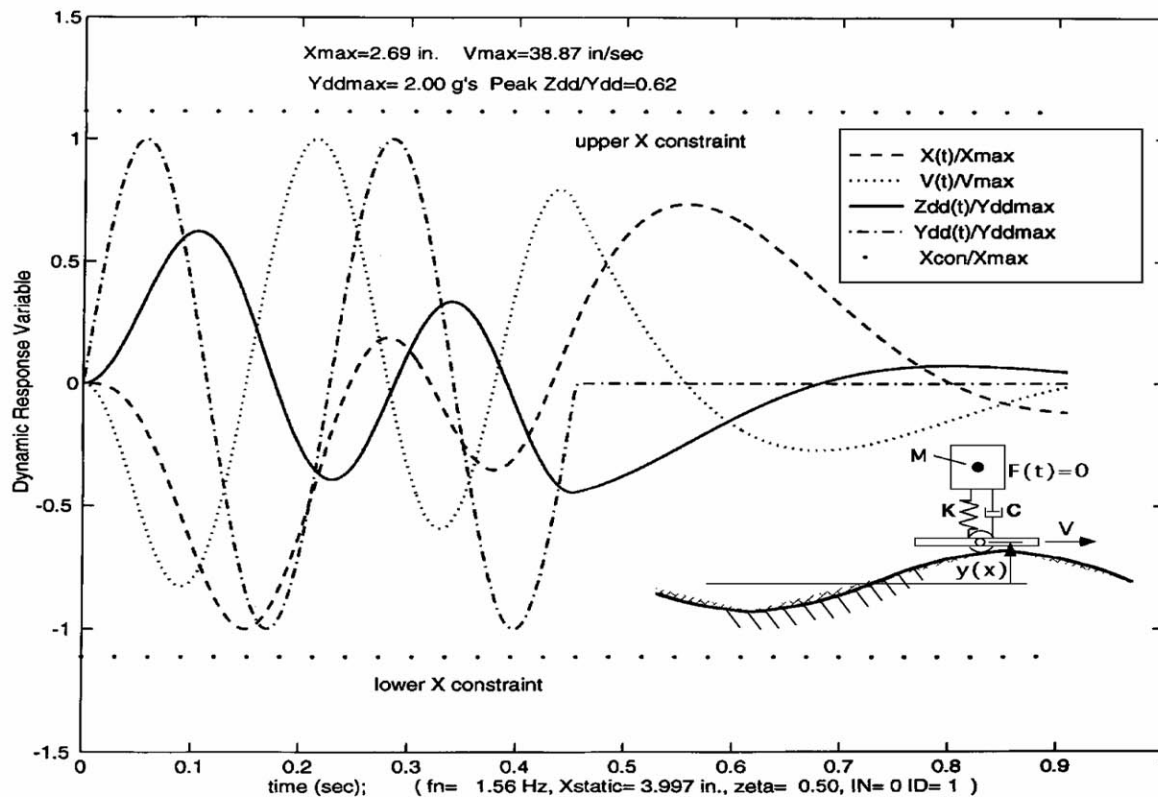
Before solving for  $\beta$  using Eq. (17.13) [or Eq. (17.14)], you must select a critical damping ratio  $\zeta$ .

At this point, a  $\zeta$  value can simply be selected according to whether light damping or heavy damping is desired, and a moderately heavy value of  $\zeta = 0.5$  is chosen in this example. Therefore, for the problem at hand, substituting  $\zeta = 0.50$  and  $T_A = 0.40$  into Eq. (17.13) gives  $\beta = 2.8131$ .



Upon substituting this value of  $\beta$  along with the drive frequency,  $f = 4.4$  Hz, into the  $\beta$  definition [i.e., the second equation in Eq. (17.11)], we can solve for the natural frequency,  $f_n = 1.564$  Hz, required to limit the acceleration transmission to  $T_A = 0.40$ . As the final step, the actual spring constant  $K$  and damping constant  $C$  can easily be back-calculated from the  $f_n$  and  $\zeta$  values in the following manner. In the sample problem, the 2000 lb load is divided equally over two isolators, so each one must carry a mass of  $M = 2000/(2 \times 386) = 2.5906$  lb-in./s<sup>2</sup>. Therefore,  $K = (2\pi f_n)^2 M = 250.2$  lb/in., and  $C = 4\pi M \zeta f_n = 25.46$  lb-s/in. for each of the two isolators. The dynamic response for important variables such as  $x(t)$ ,  $\dot{x}(t)$ ,  $\ddot{y}(t)$ , and  $\ddot{z}(t)/\ddot{y}_{\max}$  (i.e., the SC transient acceleration transmission ratio) is computed by evaluating Eqs. (17.8), (17.9), and (17.1) over the two cycles of input and for an equally long coastdown time duration after the road profile has become flat again. Upon observing the Fig. 17.4 solution, it is observed that the deflection stays within the maximum allowable constraint space of  $x_{\text{con}} = \pm 0.75\delta_s = \pm 3.00$  in.; however, the transient portion of the peak acceleration ratio  $\ddot{z}(t)/\ddot{y}_{\max} = 0.62$  overshoots the target steady state solution of  $T_A = 0.40$ . The isolator design employed the steady state solution; therefore, it is not unusual that the transient could exceed the steady state limit. It is further noted that by the end of the second response cycle, the acceleration ratio is already approaching the 0.40 target. When the road profile turns flat, the acceleration ratio rapidly tails off toward zero, due to the large damping value. To compensate for the overshoot in transient acceleration, a larger  $\beta$  (and thus a smaller  $f_n$ ) can be used on a second pass through the design process [e.g., enter  $T_A = 0.40 \times (0.40/0.62)$  into Eq. (17.13) and repeat the design process, where the peak  $\ddot{z}(t)/\ddot{y}_{\max}$  lowered from 0.62 to 0.45].

**Figure 17.4** Roadway base acceleration transmission to structure.



In building an isolator from scratch, one can simply design it to have the physical properties  $K$  and  $C$  directly. However, if isolators are to be selected from off-the-shelf stock, a few comments are in order regarding the selection process. Manufacturers often supply a set of load ( $Mg$ ) versus  $f_n$  curves for each isolator in a class of isolators. Upon entering such a set with a 1000 lb load, the isolator having a natural frequency nearest the target value, say  $f_n = 1.564$  Hz, is selected. Typically there won't be an isolator corresponding to the exact  $f_n$  desired at the operating load. Thus, picking an isolator with  $f_n$  on the low side will make the steady state transmission on the low side. In some cases the manufacturer must be contacted directly to get damping data, and in other cases damping information is given indirectly via a stated transmissibility,  $T_A$ , at resonance. The  $\beta$  location of the peak resonance  $T_A$  value in Fig. 17.2 in terms of the damping ratio  $\zeta$  is given by

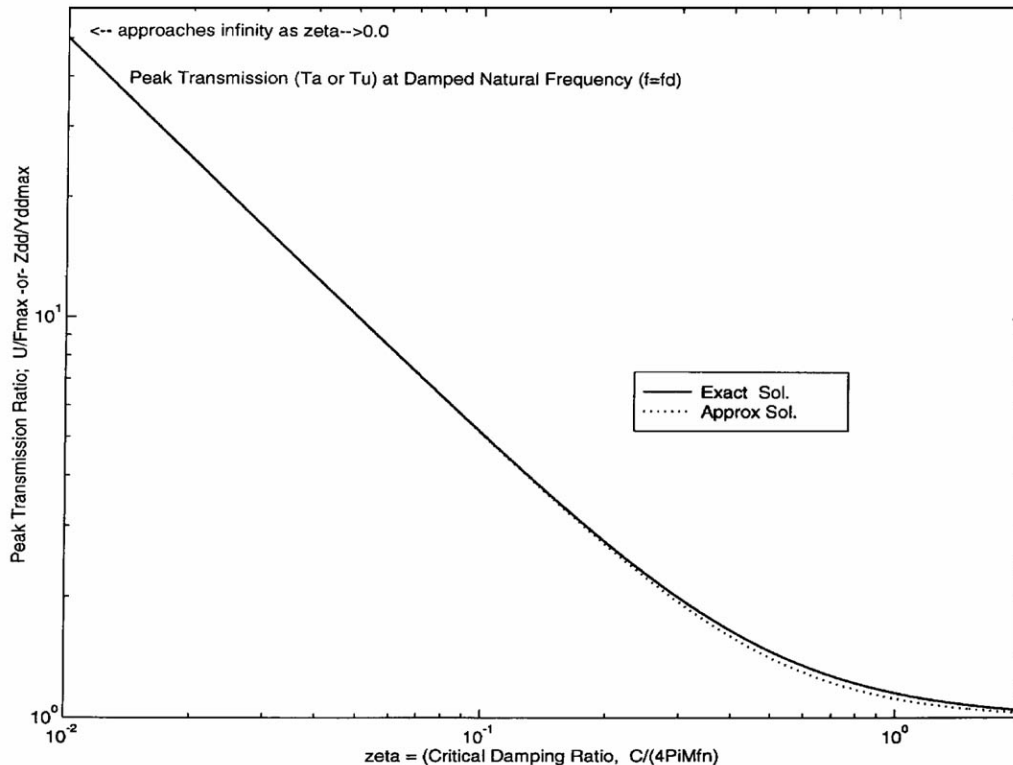
$$\beta^2 = 4\zeta^2 \left( -1 + \sqrt{1 + 8\zeta^2} \right)$$

Substituting this expression into Eq. (17.11) results in a relationship between  $T_A$  and  $\zeta$  as plotted in Fig. 17.5. For design purposes, the exact curve in inverse form can be approximated by the simpler expression

$$\zeta_{\text{res}} \approx 0.5 \sqrt{1/(T_A^2 - 1)} \quad (17.15)$$

In Fig. 17.5 both the exact curve and the Eq. (17.15) approximation are plotted side by side. For example, if the manufacturer states that an isolator has a transmission ratio  $T_A = 1.5$  at resonance, then substituting this value into Eq. (17.15) gives a corresponding critical damping ratio of  $\zeta = 0.45$ . Finally, substituting  $\zeta = 0.45$  into  $C = 4\pi M\zeta f_n$  provides the numerical value for the isolator damping constant in question.

**Figure 17.5** Transmission at resonance (harmonic force or base acceleration).



## Problem 2: Shock Isolation of Fragile Equipment

Consider a 50 lb structural configuration, initially at rest [i.e.,  $x(t=0) = 0$ ,  $\dot{x}(t=0) = 0$ ], that is subject to a half-sine pulse-type BC acceleration input of the form

$$\begin{aligned}\ddot{y} &= A_0 \sin(\omega t), & 0 \leq t \leq T_P; \\ \ddot{y} &= 0.0, & T_P < t \leq \infty; \quad T_P = \pi/\omega\end{aligned}\quad (17.16)$$

and specifically the peak value of the base acceleration input is  $A_0 = 16$  g's and the pulse duration is  $T_P = 11$  ms. In this problem, the direct force  $F(t)$  is taken as zero.

The design problem is to synthesize a shock isolator that will limit the acceleration transmitted to the SC to 4 g's; therefore, a target transmission ratio of  $T_A = 4/16 = 0.25$  is sought. It is also desired that the maximum displacement be limited to  $x = \pm 1.0$  in. At this point it is noted upon comparing the base motion of the previous application [[Eq. (17.10)]] to the base motion of the current problem [[Eq. (17.16)]] that there is a similarity between the first and second problem, except for the fact that the current shock problem has a short pulse length where only one-half of a sine wave ( $N_c = 0.5$ ) is applied. The principle governing shock isolation is different from the corresponding vibration isolation of the previous problem, in that the shock isolation process is characterized as a storage device for a sharply increasing acceleration waveform, and the design concept is to attempt to instantaneously absorb the energy and then release it at the natural frequency of the device, but at a lower-level mass deceleration. One approach is, for guessed  $K$  and  $C$  values, to simply substitute these values and the input waveform of Eq. (17.16) into Eqs. (17.7), (17.8), and (17.1) to get the response motion, observe the response, and reiterate the process with new  $K$  and  $C$  values until a desired response is obtained. We cannot use the steady state method of the previous problem because the steady state assumption of the input waveform would not be valid. An alternative design process is to convert the base input motion into a nearly equivalent initial, suddenly applied velocity problem,  $\dot{x}(t=0) = V_0$ , with no explicit driver on the right-hand side of Eq. (17.5).

The extreme case of a rapidly applied loading is a special mathematical function called a *delta function*,  $\delta(t)$ , whose value is  $\infty$  at  $t = 0$  and zero for  $t > 0$ . The right-hand-side loading can then be represented as

$$-M\ddot{y}(t) \approx -M\tilde{A}\delta(t) \quad (17.17)$$

where

$$\tilde{A} = \int_{\lambda=0}^{\lambda=T_P} \ddot{y}(\lambda) d\lambda \quad (17.18)$$

Upon substituting Eq. (17.17) into Laplace-transformed Eq. (17.7), with initial displacement  $x_0 = 0$  and  $F(s) = 0$ , and noting the Laplace transform of a delta function is 1.0, it can be seen that the  $\dot{x}_0 M$  term and  $-M\tilde{A}$  have exactly the same form (same denominator); therefore, by interchanging the roles of the driver and initial condition, we can let a problem with a zero initial velocity and pulse-type  $\ddot{y}$  driver having area  $\tilde{A}$  [by Eq. (17.18)] be replaced with an equivalent problem having a zero  $\ddot{y}$  driver but an initial velocity of  $\dot{x}_0 \equiv V_0 = -\tilde{A}$ . The  $V_0$  quantity can therefore be interpreted as a suddenly applied velocity change. The advantage of this approach is that estimates of the maximum response can easily be made and used to back-calculate the isolator

properties needed to achieve the desired isolator performance. Thus, for the equivalent initial velocity representation, we immediately get the solution as a special case of Eq. (17.8) that simply reduces to

$$x(t) = -\tilde{A}e^{-\eta t} \sin(\omega_d t)/\omega_d \quad (17.19)$$

Upon differentiating Eq. (17.19) and solving for the maximum displacement  $x_{\max}$  and maximum mass acceleration  $\ddot{z}_{\max}$ , the following result is obtained:

$$x_{\max} = -\tilde{A}e^{-\alpha}/\omega_n, \quad \text{where } \alpha = (\zeta/\sqrt{1-\zeta^2}) \sin^{-1}(\sqrt{1-\zeta^2}) \quad (17.20)$$

$$\ddot{z}_{\max} = \tilde{A}\omega_n D_i \quad (17.21)$$

$$\text{with } D_i = \exp(-\zeta \hat{t}/\sqrt{1-\zeta^2}) \left( \frac{(1-2\zeta^2)}{\sqrt{1-\zeta^2}} \sin(\hat{t}) + 2\zeta \cos(\hat{t}) \right) \quad 0 \leq \zeta \leq 0.5$$

$$D_i = 2\zeta, \quad 0.5 < \zeta \leq 1.0$$

$$\text{where } \hat{t} = \tan^{-1} \left( \frac{(1-4\zeta^2)\sqrt{1-\zeta^2}}{\zeta(3-4\zeta^2)} \right)$$

As in the previous example, the acceleration transmission ratio is defined as  $T_A = \text{peak}|\ddot{z}/\ddot{y}_{\max}|$ . Substituting Eq. (17.21) into this  $T_A$  expression and solving for the natural frequency gives

$$f_n = \frac{|\ddot{y}_{\max}|T_A}{2\pi\tilde{A}D_i} \quad (\text{general shock}) \quad (17.22)$$

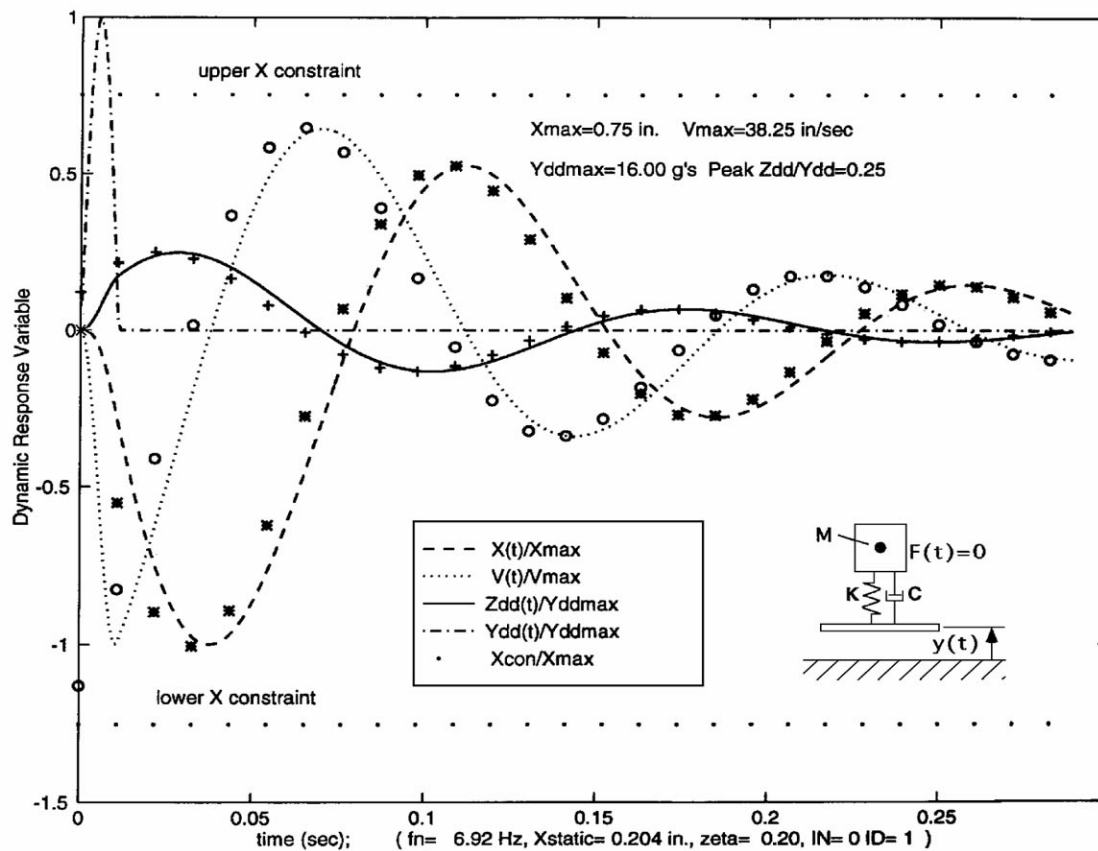
and, for very light damping (say,  $\zeta < 0.1$ ), the approximation  $D_i \approx 1.0$  can be used. The isolator selection follows the same concept as in the previous problem where the natural frequency needed to limit the transmissibility is determined. For the problem at hand, substituting the given half-sine input  $\ddot{y}$  into Eq. (17.18) corresponds to a velocity change of  $\tilde{A} = 2A_0/\omega = 2A_0T_P/\pi$ , where it is also noted that  $|\ddot{y}_{\max}| = A_0$ . Upon using these data, Eq. (17.21) reduces to

$$f_n = \frac{T_A}{4T_P D_i} \quad (\text{for half-sine pulse shock}) \quad (17.23)$$

From this point forward, the selection procedure for obtaining the isolator parameters is conceptually the same, except that the computation of the desired natural frequency is different and is governed by Eq. (17.22) or (17.23) as appropriate. For the problem at hand, substituting the target transmissibility  $T_A = 0.25$ , pulse length  $T_P = 0.011$  s, and designer-selected damping ratio  $\zeta = 0.2$  (using  $\zeta = 0.2$  in the  $D_i$  expression of Eq. (17.21) gives  $D_i = 0.8209$ ) results in a desired natural frequency of  $f_n = 6.92$  Hz. For the 50 lb SC ( $M = 50/386$ ), the  $f_n = 6.92$  and  $\zeta = 0.2$  data translate into spring and damping constants of  $K = (2\pi f_n)^2 M = 244.97$  lb/in. and  $C = 4\pi M \zeta f_n = 2.253$  lb-s/in. Also, using Eq. (17.20), the estimated maximum deflection is  $x_{\max} = 0.7520$  in.; therefore, this isolator design should meet the space constraint imposed on the problem. This  $x_{\max}$  value should also be checked against the isolator manufacturer's maximum allowable spring deflection (sometimes called sway space), which is usually given in the selection catalog. Upon substituting the above  $K$ ,  $M$ , and  $C$  design parameters and actual half-sine base motion  $\ddot{y}(t)$  into Eqs. (17.8), (17.9), and (17.1), the dynamic response for important variables such

as  $x(t)$ ,  $\dot{x}(t)$ ,  $\ddot{y}(t)$ , and  $\ddot{z}(t)/\ddot{y}_{\max}$  (i.e., the SC transient acceleration transmission ratio) is computed. The results are shown in Fig. 17.6, where it is noted that the desired 0.25 transmissibility is achieved and the displacement constraints are not exceeded. It should be noted that the curves labeled in the legend correspond to using the actual half-sine input base motion. These results will not be exactly the same as the equivalent suddenly applied velocity ( $V_0$ ) solution that was used to size the isolators, because the half-sine pulse is not exactly the idealized delta function. For illustrative purposes, the equivalent suddenly applied velocity solution is superimposed on the same plot and denoted with the unconnected symbolic markers "\*" for displacement, "o" for velocity, and "+" for acceleration ratio. As expected, the velocity comparison will be different in the early time [e.g., with the actual  $\ddot{y}(t)$  input the initial condition is  $\dot{x}_0 = 0$ , whereas in the equivalent problem the initial velocity is not zero and represents the entire input to the problem]. As a final comment, the selection of the isolators from the manufacturer's catalog follows along the same lines here and therefore will not be repeated; also check that  $T_P \ll 1/f_n$  when using Eq. (17.22) or (17.23).

**Figure 17.6** Base configuration half-sine acceleration loading.



## 17.2 Structural Configuration Loaded Applications

### Problem 3: Rotating Machinery Force Transmission

In this class of problems, the base configuration is considered fixed; thus  $y(t) = \dot{y}(t) = \ddot{y}(t) = 0$ , and therefore the only system loading comes through the structural configuration loading  $F(t)$ . Perhaps one of the most common such loading problems is the situation where a piece of equipment with some sort of rotating member is spinning in a steady state mode at drive frequency

equipment with some sort of rotating member is spinning in a steady state mode at drive frequency  $\omega$  and is resting on the structural configuration, where the total mass  $M$  of the equipment to be supported by the isolators is  $M = M_s + m_e$  where  $M_s$  is the mass of the structural configuration including the rotating machinery (except for the offset mass  $m_e$ ), and  $m_e$  represents the off-center eccentric mass at radius  $r_e$ . The  $m_e$  term is analogous to wet clothes clinging to the spinning drum during the spin-dry cycle of a common washing machine, where excessive vibrations are set up when the clothes are not uniformly distributed around the drum. The radial acceleration,  $A_r = r_e \omega^2$ , results in a reciprocating force that is represented by

$$F(t) = m_e r_e \omega^2 \cos(\omega t) \quad (17.24)$$

Therefore, by comparing Eq. (17.24) to the first steady state loading example in Eq. (17.10), it is seen that the  $M\omega^2 Y_0 \sin(\omega t)$  driver in the differential equation is just like the current  $m_e r_e \omega^2 \cos(\omega t)$  driver except for a cosine in place of a sine driver function. In fact, the steady state solutions for the transmissibility of an isolator force transmission ratio of  $T_U = \text{peak}|u(t)/F_{\max}|$  has the exact same form as Eq. (17.11); therefore,  $T_U = T_A$  for this class of problem, where the driver force amplitude varies as  $\omega^2$ . It is cautioned that for other harmonically varying forces, ones that have, say, a frequency-independent amplitude, the form of the  $T_U$  would not be the same as Eq. (17.11). It is also noted that although the actual value of the amplitude of  $F(t)$  depends on the size of  $m_e r_e$ , this value cancels out in forming the  $T_U$  force transmission ratio.

As a specific example, consider a structural configuration, whose total weight (including  $m_e$ ) is 800 lb and the equipment is rotating at 540 rpm (i.e.,  $\omega_0 = 2\pi 9$  rad/s), which due to an offset ( $r_e = 10.0$  in.), an eccentric rotating mass ( $m_e = 0.25$  lb-s<sup>2</sup>/in.), transmits unwanted vibrations to the floor. It is required that the force transmission ratio be no more than  $T_U = 0.2$  (sometimes equivalently referred to as 80% isolation). It is also required that the maximum displacement  $x(t)$  not exceed  $\pm 0.5$  in. It is further assumed that the SC is to be supported by four isolators and that they are centered so that the dead weight is distributed equally among them. If this centering assumption is not met, rocking modes will be present and more degrees of freedom will be needed to model the system. The motor cannot go from  $\omega = 0$  up to the operating frequency instantly; therefore, it is assumed that  $\omega$  has a simple linear time ramp:

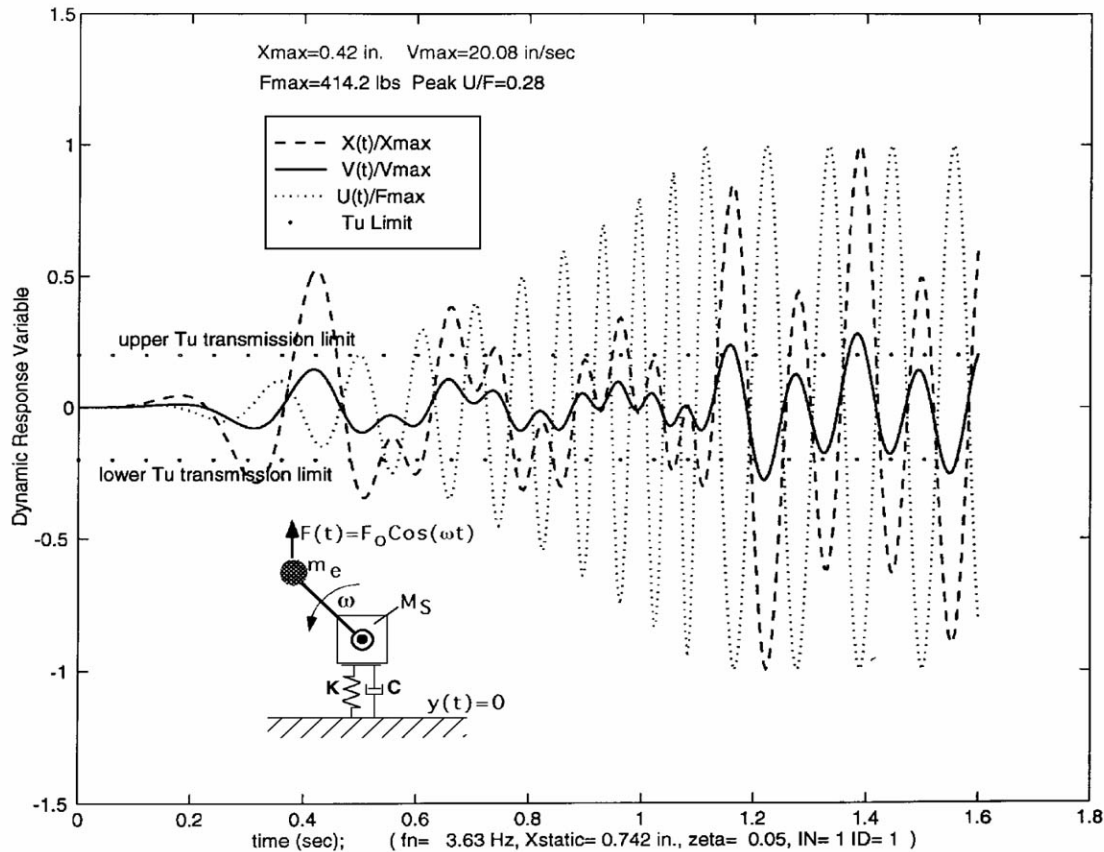
$$\omega = \omega_0 t/t_c, \quad 0 \leq t \leq t_c; \quad \omega = \omega_0, \quad t_c \leq t \quad (17.25)$$

where for the problem at hand  $t_c = 10 \times (2\pi/\omega_0)$ , which corresponds to 10 periods of the steady state frequency. The first step in designing the isolator is to select a critical damping ratio  $\zeta$ . Because the motor has a variable speed and the rotational frequency varies with time during the startup according to Eq. (17.25), the isolator is expected to have a designer-selected amplification of no more than  $T_U = 10$  at resonance, should the motor ever be run at or near the resonant frequency. Thus, substituting  $T_U = T_A = 10$  into Eq. (17.15) results in a damping ratio of  $\zeta = 0.0502$ . The next step is to solve for the natural frequency that will limit the steady state vibration force transmission ratio (operating at speed  $f = \omega_0/2\pi = 9$  Hz) to a value of  $T_U = 0.2$ . This is accomplished by substituting  $\zeta = 0.0502$  and  $T_U = T_A = 0.2$  into Eq. (17.13), which results in  $\beta = 2.4792$ .

There are four isolators; therefore, the mass assigned to each isolator is determined from  $M = (800/386)/4 = 0.5181$  lb-s<sup>2</sup>/in. The natural frequency is computed from  $f_n = f/\beta = 9/2.4792 = 3.63$  Hz. Finally, the spring and damping constants are computed from

$K = (2\pi f_n)^2 M = 269.5 \text{ lb/in.}$  and  $C = 4\pi M\zeta f_n = 1.186 \text{ lb-s/in.}$  (Refer to the discussion on selecting isolators from manufacturer's catalogs in the first suspension isolator design example.) Upon substituting the above  $K$ ,  $M$ ,  $C$  design parameters and driver  $F(t)$  defined by Eqs. (17.24) and (17.25) into Eqs. (17.8), (17.9), and (17.4), the dynamic response for important variables such as  $x(t)$ ,  $\dot{x}(t)$ ,  $F(t)$ , and  $|u(t)/F_{\max}|$  (i.e., the SC force transmission ratio) is computed. It is noted that during the startup phase, the variable angular velocity  $\omega(t)$  results in an additional tangential acceleration component that is neglected in the analysis. The transient portion of the results are shown in Fig. 17.7, where it is noted that the desired 0.20 transmissibility is exceeded ( $T_A = 0.28$ ) during the transient (and also that the displacement constraints are not exceeded). The damping is light in this example; therefore, it takes many more cycles to reach the steady state. The horizontal dotted lines in Fig. 17.7 align with the steady state limit reached for  $u(t)/F_{\max}$  after running the solution out to  $t = 4.0$ .

**Figure 17.7** Transient-into-steady state force transmission applied to foundation.



This demonstration example also illustrates the need to balance rotating machinery. Instead of having to live with the vibrations, it would be better to kill the source of the vibration and rebalance the equipment (e.g., make  $r_e$  smaller in this example), so that even if the vibration is still present, its magnitude will be small enough that it will have the same effect as if isolators were used to reduce the transmissibility.



## Problem 4: Free-Fall Shock

This class of problem is similar to the base-excited impulse problem, where a falling object (structural configuration) strikes the ground (base configuration) and, due to the impact, a suddenly applied velocity is imparted to the falling object. The ground is idealized as rigid, and it is assumed that the object falls from a height  $H$ . The velocity of the object *just prior to hitting the ground* is  $V = \sqrt{2gH}$ . The problem is to design an isolator that limits the maximum force transmissibility  $T_U$  to a prescribed amount (often referred to as the fragility factor, nondimensional  $g$ 's). The solution of this problem follows along the same lines as the initial impulse problem described earlier in this section; therefore, only a rough outline will be presented. The governing differential equations have to be modified slightly to account for the fact that the free-falling body–isolator configuration has no prestretch (static deflection  $\delta_S$ ), and therefore the static deflection force term  $K\delta_S$  does not cancel with the dead weight term in the derivation of Eq. (17.5). Therefore, in solving this problem,  $x(t)$  in Eq. (17.5) should be viewed as the deflection measured *from the unstretched equilibrium position*. Once the base of the isolator has just touched the ground at impact, the isolator starts to compress and the  $x(t)$  solution describes the ensuing motion of the SC. The loading in Eq. (17.5) becomes simply the dead weight (i.e.,  $F(t) = -Mg$  for  $t \geq 0$ , with  $\ddot{y}(t) = 0$ ) and the initial conditions in Eq. (17.6) are  $x_0 = 0$ ,  $\dot{x}_0 = -V = -\sqrt{2gH}$  [the sign convention is that the extension ( $+x$ ) of the isolator is positive]. Thus, with these conditions substituted into Eqs. (17.8), (17.9), and (17.4), the dynamic response for important variables such as  $x(t)$ ,  $\dot{x}(t)$ ,  $F(t)$ , and  $T_U = |u(t)/F_{\max}|$  (i.e., the SC force transmission ratio) can be evaluated, where  $F_{\max}$  in this application is simply the dead weight,  $-Mg$ . Solving for the exact solution for  $u(t)$  as described above, and finding its maximum value in time, one obtains

$$T_U = \left| -1 + \exp \left( -\zeta \hat{t} / \sqrt{1 - \zeta^2} \right) \left[ \frac{(-\Omega(1 - 2\zeta^2) - \zeta)}{\sqrt{1 - \zeta^2}} \sin(\hat{t}) + (1 - 2\Omega\zeta) \cos(\hat{t}) \right] \right|$$

$$0 \leq \zeta \leq \zeta_b \quad (17.26a)$$

or

$$T_U = |-2\zeta\Omega| \quad \zeta_b \leq \zeta \quad (17.26b)$$

with

$$\zeta_b = \frac{1}{4\Omega} + 0.5\sqrt{1 + 1/(2\Omega)^2}, \quad \Omega = \frac{V\omega_n}{g}, \quad (17.26c)$$

$$\hat{t} = \tan^{-1} \left[ \frac{(- (1 - 4\zeta^2) + 2\zeta)\sqrt{1 - \zeta^2}}{(-1 + 3\Omega\zeta + 2\zeta^2 - 4\Omega\zeta^3)} \right]$$

For intermediate damping, Eq. (17.26a) corresponds to the nondimensional time value,  $\hat{t} = \omega_d t_{\max}$ , where the slope  $du/dt = 0.0$ , and Eq. (17.26b) corresponds to the large damping case where the maximum force occurs at the beginning of impact,  $t = 0.0$ . When the damping  $\zeta$  is zero (or very small), then using Eq. (17.26a) with  $\hat{t} \approx \tan^{-1}(\Omega/(-1))$  and solving for  $\Omega$  in terms of  $T_U$ , we can solve for the natural frequency that limits the maximum transmitted force to the desired  $T_U$  value,



where

$$f_n \approx \frac{g}{2\pi V} \sqrt{T_U(T_U - 2)} \quad \text{for light damping, } \zeta \approx 0.0, \text{ and } T_U > 2.0 \quad (17.27)$$

The  $T_U > 2.0$  limitation comes from the fact that *with zero or low damping*, fragility factors lower than 2.0 are not reachable without substantial damping. In the case of heavy damping, the maximum force occurs at the beginning,  $t = 0.0$ , and using Eq. (17.26b), one obtains:

$$f_n = \frac{gT_U}{4\pi V\zeta} \quad \text{for heavy damping, } \zeta_b \leq \zeta \quad (17.28)$$

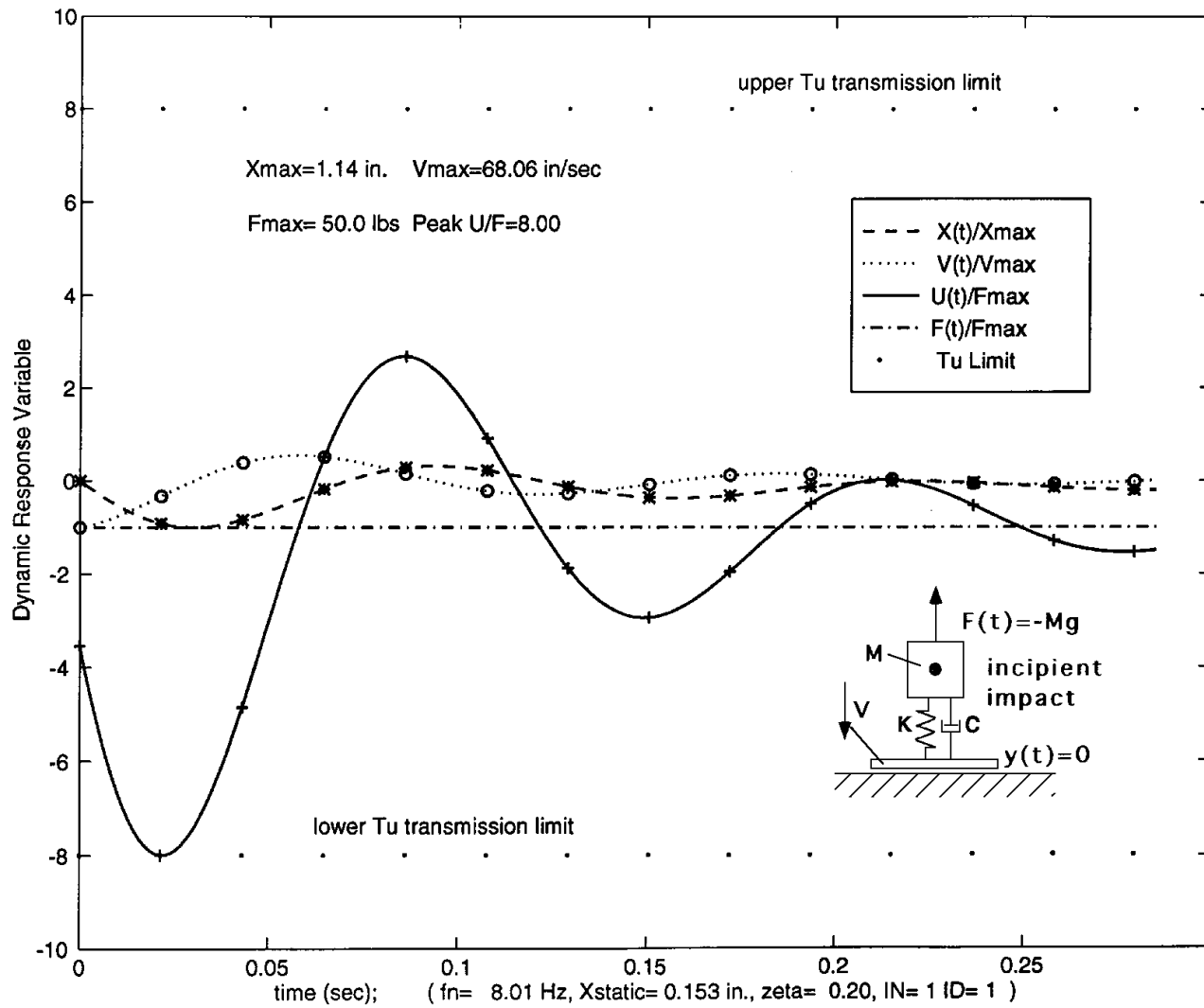
where the breakpoint,  $\zeta_b$ , is determined by setting the numerator of the arctan expression for  $\hat{t}$  equal to zero. Finally the determination of the  $f_n$  value in terms of  $T_U$  for the intermediate damping,  $0 \leq \zeta \leq \zeta_b$ , is the most difficult case because the solution involves obtaining roots to a transcendental equation. It is noted that for large values -  $\geq 10.0$ , the breakpoint value is  $\zeta_b \approx 0.5$ . Use the smallest positive root for  $\tan^{-1}$  in Eq. (17.26c) (e.g., atan2 in MATLAB).

The procedure is numerical and can easily be accomplished by substituting the desired damping ratio  $\zeta$  and fragility factor  $T_U$  into Eq. (17.26a) and iterating - until the left side equals the right side (this root, - = -<sub>rt</sub>, can be found with a simple numerical root finder routine such as the "fzero" routine in MATLAB). Next simply convert -<sub>rt</sub> to  $f_n$  using the second of Eqs. (17.26c), to get

$$f_n = g\Omega_{rt}/(2\pi V) \quad \text{for intermediate damping, } 0 \leq \zeta \leq \zeta_b \quad (17.29)$$

To illustrate the general drop application for intermediate damping, consider a 50 lb weight ( $M = 50/g$ ) that is dropped a height  $H = 6.00$  in. (i.e.,  $V = 68.06$  in./s), where it is required that  $T_U$  be no bigger than 8.0, with a prescribed intermediate damping ratio  $\zeta = 0.2$ . Next insert these data into Eq. (17.26a), solve for the root  $\Omega_{rt} = 8.870$ , and finally compute the isolator natural frequency  $f_n = 8.0066$  Hz with Eq. (17.29). The breakpoint value  $\zeta_b$  must be checked to ensure that the inequality bounds of Eq. (17.29) are not violated; thus, substituting  $\Omega = \Omega_{rt} = 8.870$  into the first of Eqs. (17.26c) gives  $\zeta_b = 0.5290$ , which is above the  $\zeta = 0.2$  value required by the constraint bounds for intermediate damping. These  $f_n$  and  $\zeta$  data translate into  $K = (2\pi f_n)^2 M = 327.82$  lb/in. and  $C = 4\pi M\zeta f_n = 2.607$  lb-s/in. Evaluating the dynamic response as described above with Eqs. (17.8), (17.9), and (17.4) results in the response illustrated in Fig. 17.8 which, as indicated, has peak  $|u(t)/F_{\max}| = 8.0$ . Note that the isolator force turns into tension after 0.062 s, which implies that the structural configuration will jump up off the floor (i.e., "pogo stick effect") at a later time after the energy is absorbed. It is also noted that the  $x(t)$  solution does not settle down to zero, but rather to the static deflection value; this is because of the fact that Fig. 17.1(a) rather than Fig. 17.1(b) was the coordinate reference configuration for this free-fall problem. For comparison purposes, the solution generated with Eqs. (17.8), (17.9), and (17.4) is compared with a solution to the same problem in [Church, 1963] and is indicated by the unconnected symbolic markers "\*" for displacement, "o" for velocity, and "+" for the  $|u(t)/F_{\max}|$  ratio (after adjusting for a different sign convention for positive  $x$ ). As can be seen, the agreement between the two solutions is perfect.

**Figure 17.8** Free-falling structural configuration shock.



## 17.3 Additional Information

---

The sample problems considered here represent simple single-degree-of-freedom (SDF) models of what is often a more complicated multiple-degree-of-freedom (MDF) system. These simple example problems illustrate the main concepts involved in shock and vibration isolation; however, great care must be taken not to apply these idealized SDF-type models in situations where a MDF model is needed to represent the full picture (e.g., the four-wheel vehicle over a roadway needs a MDF model that allows the front and rear wheels to experience different parts of the roadway profile to allow for any rocking modes that may be present). Perhaps one of the most comprehensive references for shock and vibration isolation issues is [Harris, 1988], which not only covers SDF models in detail for all kinds of inputs, but covers MDF models as well. Good design tips are often found directly in the manufacturer's design guides, for example, [Aeroflex, 1994], [Barry, 1993], [Lord, 1994], and [Firestone, 1994]. For some more advanced ideas on optimum shock and vibration isolation concepts, see [Sevin and Pilkey, 1971]. The effects of nonlinearity should also be considered when necessary; for example, the simple Voigt model linear spring-damper isolators considered here eventually turn nonlinear when the deflections get large. In such cases the actual nonlinear isolator force function  $u(x, \dot{x})$ , could be obtained from the manufacturer and Eq. (17.1) resolved as a nonlinear differential equation (which could also easily be solved with a MATLAB script file similar to the one used to generate the sample solutions).

For the analysis of MDF systems consisting of a collection of rigid bodies interconnected by any "Rube Goldberg" collection of linkages, pivot points, and so forth, a handy computer program called Working Model exists [Knowledge Revolution, 1994], enabling the user to geometrically construct the configuration to be analyzed on the computer screen. It is not unlike the object-drawing programs that come with practically all modern-day word processors. The key difference is that once the model is drawn on the screen, following the assignment of initial conditions, spring constants, damping coefficients, and loading via pulldown menus, the user is one "mouse click" away from setting the program to work and getting a graphical display of selected pertinent response variables (displacement, velocity, acceleration) as the solution is in progress. Versions exist for a variety of platforms, including Macintoshes, PCs, and workstations. Depending on the user's on-screen drawing skills, an entire analysis of a rather complex MDF system can be completed in a very short time (e.g., 5–10 minutes).

### Defining Terms

$C$ : isolator damping constant (lb-s/in.)

$K$ : isolator spring constant (lb/in.)

$M$ : net structural configuration mass (lb-s/in.<sup>2</sup>)

$g$ : acceleration of gravity (386 in./s<sup>2</sup>)

$u$ : isolator force not including static deflection (lb)

$\beta = f/f_n$ : ratio of drive frequency to natural frequency (unitless)

$\zeta$ : critical damping ratio =  $C/(4\pi M f_n)$  (unitless)

$\delta_s$ : static deflection =  $Mg/K$ , amount spring deflects under dead weight of mass  $M$  (in.)

$\omega_n$ : natural frequency (rad/s)

$\omega_n = \omega_d \sqrt{1 - \zeta^2}$ : damped natural frequency (rad/s)

$\eta$ : damping decay constant ( $\text{s}^{-1}$ )  
 $T_U$ : force transmission ratio =  $\text{peak}|u(t)/F_{\max}|$  (unitless)  
 $T_A$ : acceleration transmission ratio =  $\text{peak}|\ddot{z}/\ddot{y}_{\max}|$  (unitless)  
 $T_D$ : displacement magnification ratio =  $\text{peak}|x/y_{\max}|$  (unitless)  
 $BC$ : base configuration, refers to the lower isolator attachment point (where base input accelerations  $\ddot{y}$  are applied)  
 $SC$ : structural configuration, refers to the net vibrating mass (upper connection to the isolator)

## References

- Aeroflex. 1994. *Aeroflex Isolators Selection Guide*. Aeroflex International, Inc., Plainview, NY.  
 Barry. 1993. *Barry Controls Bulletin DOEM1*. Barry Controls, Brighton, MA.  
 Church, A. H. 1963. *Mechanical Vibrations*, 2nd ed. John Wiley & Sons, Inc., New York.  
 Firestone. 1994. *Engineering Manual and Design Guide*. Firestone Industrial Products Co., Noblesville, IN.  
 Harris, C. M. 1988. *Shock and Vibration Handbook*, 3rd ed. McGraw-Hill Book Co., New York.  
 Knowledge Revolution. 1994. *Working Model Demonstration Guide and Tutorial*. San Mateo, CA.  
 Lord. 1994. *Lord Industrial Products Catalog*, PC-2201H. Lord Industrial Products, Erie, PA.  
 Math Works. 1992. *The Student Edition of MATLAB*. Prentice Hall, Englewood Cliffs, NJ.  
 Redfern, D. 1994. *The Maple Handbook: Maple V Release 3*. Springer-Verlag, New York.  
 Sevin, E. and Pilkey, W. D. 1971. *Optimum Shock and Vibration*, Monogram SVM-6. Shock and Vibration Information Center, Naval Research Laboratory, Washington, DC.  
 Sigmon, K. 1994. *MATLAB Primer*, 4th ed. CRC Press Inc., Boca Raton, FL.  
 Wolfram, S. 1991. *Mathematica: A System for Doing Mathematics by Computer*, 2nd ed. Addison-Wesley Publishing Co., Redwood City, CA.

Inman, D. J. "Computer Simulation and Nomographs"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Computer Simulation and Nomographs

---

- 18.1 Nomograph Fundamentals
- 18.2 Models for Numerical Simulation
- 18.3 Numerical Integration
- 18.4 Vibration Response by Computer Simulation
- 18.5 Commercial Software for Simulation

## Daniel J. Inman

*Virginia Polytechnic Institute and State University*

A primary concern in performing vibration analysis is just how to represent the response once the model and the various inputs of interest are known. For linear systems with a single degree of freedom, the analytical solution is closed form and can be simply plotted to illustrate the response. Historically, computation has been difficult and response vibration data have been presented in nomographs consisting of log plots of the maximum amplitudes of displacement, velocity, and acceleration versus frequency on a single two-dimensional four-axis plot. Although this approach is useful and incorporated into military and manufacturer specifications, wide availability of high-speed computing and computer codes to simulate detailed responses has produced a trend to display exact responses in the time domain. The following section introduces an example of a commercial computer simulation package and its use in representing the response of vibrating systems. In addition, the basic use of nomographs is presented.

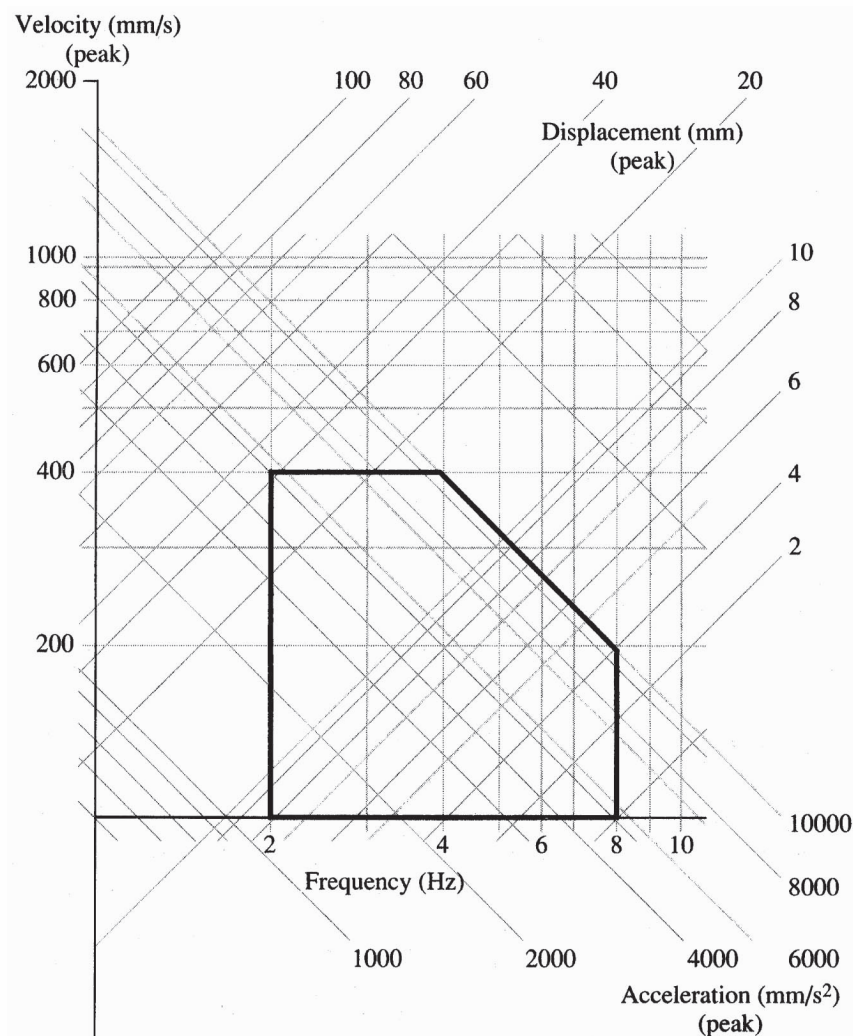
## 18.1 Nomograph Fundamentals

---

**Nomographs** are graphs used to represent the relationship between displacement, velocity, acceleration, and frequency for vibrating systems. These graphs are frequently used to represent vibration limits for given parts, machines, buildings, and components. The basic premise behind a vibration nomograph is that the response of a system is harmonic of the form  $x(t) = A \sin \omega t$ . Here  $A$  is the amplitude of vibration and  $\omega$  is the natural frequency of vibration in rad/s. The velocity is the derivative  $\dot{x}(t) = A\omega \cos \omega t$  and the acceleration is the second derivative  $\ddot{x} = -\omega^2 A \sin \omega t$ . Thus if a vibrating system has displacement amplitude  $A$ , its velocity has amplitude  $\omega A$  and the acceleration amplitude is  $\omega^2 A$ . For a given harmonic motion these three amplitudes can be plotted versus frequency, commonly using a log scale as illustrated in Fig. 18.1. In this log scale plot the log of the frequency (denoted  $f$  in Hz, where  $f = \omega/2\pi$ ) is plotted along the horizontal and the corresponding velocity amplitude is plotted along the vertical, also on a log scale. The lines slanting to the right, of slope +1, correspond to the log of the displacement

amplitude versus frequency, whereas those slanting to the left, with slope  $-1$ , correspond to the log of the acceleration amplitude versus frequency. Thus a given point on the nomograph corresponds to the amplitude of displacement, velocity, and acceleration at a specific frequency. Nomographs can be used to specify regions of acceptable vibration. Often it is not enough to specify just displacement; restrictions may also exist on velocity and acceleration amplitudes. By sketching a closed shape on a nomograph, ranges of acceptable levels of maximum displacement, velocity, and acceleration over a frequency range of interest can be easily specified. An example is illustrated by the bold lines in Fig. 18.1. The bold lines in the figure are used to illustrate vibration between 2 and 8 Hz with displacement amplitude limited by 30 mm, velocity amplitude limited by 400 mm/s and acceleration limited to amplitude by  $10^4$  mm/s<sup>2</sup>.

**Figure 18.1** An example of a vibration nomograph for specifying acceptable limits of sinusoidal vibration. (Source: Inman, D. J. 1994. *Engineering Vibration*. Prentice Hall, Englewood Cliffs, NJ. With permission.)



Rather than maximum amplitude, root mean square values of displacement, velocity, and

acceleration can be plotted as nomographs. As mentioned, such plots are often used in formal documents, vendor specifications, and military specifications. The International Organization for Standardization presents vibration standards for severity, which are often represented in nomograph form.

As useful as nomographs are and as frequently as they appear in vibration literature and in codes and standards, modern computational abilities allow detailed representations of vibration data for much more complicated systems. In particular, it has become very routine to simulate time responses directly.

## 18.2 Models for Numerical Simulation

---

The most common model of a vibrating system is the multiple degree of freedom (MDOF) model, which can be expressed as a vector differential equation with matrix coefficients of the form

$$M\ddot{\mathbf{x}}(t) + C\dot{\mathbf{x}}(t) + K\mathbf{x}(t) = \mathbf{f}(t) \quad \mathbf{x}(0) = \mathbf{x}_0, \dot{\mathbf{x}}(0) = \dot{\mathbf{x}}_0 \quad (18.1)$$

where  $\mathbf{x}(t)$  is an  $n \times 1$  vector of displacement coordinates, its derivative  $\dot{\mathbf{x}}(t)$  is an  $n \times 1$  vector of velocities, and its second derivative  $\ddot{\mathbf{x}}(t)$  is an  $n \times 1$  vector of accelerations. The coefficients  $M$ ,  $C$ , and  $K$  are  $n \times n$  matrices of mass, damping, and stiffness elements, respectively. These coefficient matrices are often symmetric and at least positive semidefinite for most common devices and structures. Equation (18.1) follows from simple modeling using Newton's laws, energy methods, or dynamic finite elements. The constant vectors  $\mathbf{x}_0$  and  $\dot{\mathbf{x}}_0$  represent the required initial conditions. The simulation problem consists of calculating  $\mathbf{x}(t)$ , satisfying Eq. (18.1) as time evolves, and producing a time record of each element of  $\mathbf{x}(t)$ , or of each degree of freedom, as opposed to a single coordinate as used in nomographs.

Currently many very well written numerical integration codes are available commercially for less cost than is involved in writing the code and, more importantly, with less error. Codes used to solve Eq. (18.1) are written based on the definition of a derivative and almost all require the equations of motion to appear in first-order form, that is, with only one instead of two time derivatives. Equation (18.1) can be easily placed into the form of a first-order vector differential equation by some simple matrix manipulations.

If the inverse of the mass matrix  $M$  exists, then the second-order vibration model of Eq. (18.1) can be written as an equivalent first-order equation using new coordinates defined by the  $2n \times 1$  vector:

$$\mathbf{z}(t) = \begin{bmatrix} \mathbf{x}(t) \\ \dot{\mathbf{x}}(t) \end{bmatrix}$$

Let  $\mathbf{z}_1 = \mathbf{x}(t)$  and  $\mathbf{z}_2 = \dot{\mathbf{x}}(t)$ ; then Eq. (18.1) can be written as



$$\begin{aligned}
\dot{\mathbf{z}}_1 &= \mathbf{z}_2 \\
\dot{\mathbf{z}}_2 &= -M^{-1}K\mathbf{z}_1 - M^{-1}C\mathbf{z}_2 + M^{-1}\mathbf{f}(t) \\
\mathbf{z}(0) &= \begin{bmatrix} \mathbf{x}_0 \\ \dot{\mathbf{x}}_0 \end{bmatrix}
\end{aligned} \tag{18.2}$$

which combine to form

$$\dot{\mathbf{z}} = A\mathbf{z} + \mathbf{F}(t), \quad \mathbf{z}(0) = \mathbf{z}_0 \tag{18.3}$$

Here  $\mathbf{z}$  is called a *state vector* and the state matrix  $A$  is defined by

$$A = \begin{bmatrix} \mathbf{0} & I \\ -M^{-1}K & -M^{-1}C \end{bmatrix} \tag{18.4}$$

where  $I$  is the  $n \times n$  identity matrix and  $\mathbf{0}$  is an  $n \times n$  matrix of zeros. The forcing function  $\mathbf{F}(t)$  is "mass" scaled to be the  $2n \times 1$  vector

$$\mathbf{F}(t) = \begin{bmatrix} \mathbf{0} \\ M^{-1}\mathbf{f}(t) \end{bmatrix} \tag{18.5}$$

where  $\mathbf{0}$  denotes an  $n \times 1$  vector of zeros. Note that in solving vibration problems using this state space coordinate system, the first  $n$  components of the solution vector  $\mathbf{z}(t)$  correspond to the individual displacements of the  $n$  degrees of freedom.

## 18.3 Numerical Integration

---

The numerical solution or simulation of the system described by Eq. (18.1) is easiest to discuss by first examining the scalar homogeneous (unforced) case given by

$$\dot{x}(t) = ax(t) \quad x(0) = x_0$$

where  $a$  is a simple constant. The derivative  $\dot{x}(t)$  is written from its definition as

$$\frac{x(t_1 + \Delta t) - x(t_1)}{\Delta t} = ax(t_1) \tag{18.6}$$

where  $\Delta t$  is a finite interval of time. Rewriting this expression yields

$$x(t_1 + \Delta t) = x(t_1) + ax(t_1)\Delta t \tag{18.7}$$

or using a simpler notation

$$x_{i+1} = x_i + ax_i \Delta t \quad (18.8)$$

where  $x_i$  denotes  $x(t_i)$ . This formula gives a value of the response  $x_{i+1}$  at the "next" time interval, given the equation's coefficient  $a$ , the time increment  $\Delta t$ , and the previous value of the response  $x_i$ . Thus, starting with the initial value  $x_0$ , the solution is computed at each time step incrementally until the entire record over the interval of interest is calculated. This simple numerical solution is called the **Euler formula** or tangent line method and only involves addition and multiplication. Of course, the smaller  $\Delta t$  is, the more accurate the approximation becomes (recall the derivative is defined as the limit  $\Delta t \rightarrow 0$ ). Unfortunately, reducing the step size  $\Delta t$  increases the computational time. Numerical errors (rounding and truncation) also prevent the simulations from being perfect, and users should always check their results accordingly.

The rule used to perform the simulation is often called an *algorithm*. One way to improve the accuracy of numerical simulation is to use more sophisticated algorithms. In the late 1800s C. Runge and M. W. Kutta developed some clever formulas to improve the simple tangent or Euler methods. Essentially, these methods examine  $x(t + \Delta t)$  as a Taylor series expanded in powers of  $\Delta t$ . The **Runge-Kutta** methods insert extra values between  $x_i$  and  $x_{i+1}$  to provide estimates of the higher-order derivative in the Taylor expansion and thus improve the accuracy of the simulation. There are several Runge-Kutta methods, thus only an example is given here.

One of the most widely used Runge-Kutta methods solves the scalar equation  $\dot{x} = f(x, t)$  with initial condition  $x(0) = x_0$  where  $f(x, t)$  can be linear or nonlinear as well as time varying. This includes the case  $\dot{x}(t) = ax(t) + g(t)$  where  $a$  is constant and  $g(t)$  is an externally applied force. With  $x_i$  and  $\Delta t$  defined as before, the formulas for the response are

$$x_{i+1} = x_i + \frac{\Delta t}{6}(h_{i1} + 2h_{i2} + 2h_{i3} + h_{i4}) \quad (18.9)$$

where

$$\begin{aligned} h_{i1} &= f(x_i, t_i) & h_{i2} &= f\left(x_i + \frac{\Delta t}{2}h_{i1}, \quad t_i + \frac{\Delta t}{2}\right), \\ h_{i3} &= f\left(x_i + \frac{\Delta t}{2}h_{i2}, \quad t_i + \frac{\Delta t}{2}\right) & h_{i4} &= f(x_i + \Delta t h_{i3}, \quad t_i + \Delta t) \end{aligned}$$

This is referred to as a four-stage formula and represents a substantial improvement over the Euler method.

Additional improvement can be gained by adjusting the time step  $\Delta t$  at each interval based on how rapidly the solution  $x(t)$  is changing. If the solution is not changing very rapidly, a large value of  $\Delta t_i$ , the  $i$ th increment of time is used. On the other hand if  $x(t)$  is changing rapidly, a small  $\Delta t_i$  is chosen. In fact, the  $\Delta t_i$  can be chosen automatically as part of the algorithm.

## 18.4 Vibration Response by Computer Simulation

---

All of these methods (as well as many others not mentioned) can be applied to the simulation of the response of vibrating systems. Essentially the Runge-Kutta and Euler formulas can be applied directly to Eq. (18.3) by simply enforcing a vector notation. For instance, the Euler formula applied to Eq. (18.3) becomes

$$\mathbf{z}(t_{i+1}) = \mathbf{z}(t_i) + \Delta t \mathbf{A} \mathbf{z}(t_i) + \mathbf{F}(t_i) \quad (18.10)$$

using  $\mathbf{z}(0)$  as the initial value. The result will be a list of numerical values for  $\mathbf{z}(t_i)$  versus the successive times  $t_i$ . Equation (18.10) can be programmed on a programmable calculator or computer system. However, many commercially available codes provide more than adequate numerical integration schemes for solving ordinary differential equations and systems of ordinary differential equations as described by Eq. (18.10). Such codes are easy to use and allow studies of the effects of initial conditions and parameter changes while providing detailed solutions to complex problems.

Next a simple example is introduced to illustrate the formulation of a vibration problem into state space form in preparation for numerical simulation. Consider then the equations of motion of a damped two-degree-of-freedom system

$$\begin{bmatrix} 9 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \ddot{x}_1(t) \\ \ddot{x}_2(t) \end{bmatrix} + \begin{bmatrix} 27 & -0.3 \\ -0.3 & 0.3 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} + \begin{bmatrix} 27 & -3 \\ -3 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \end{bmatrix} \sin 3t \quad (18.11)$$

subject to the initial conditions

$$\mathbf{x}(0) = \begin{bmatrix} 0.1 \\ 0 \end{bmatrix}, \quad \dot{\mathbf{x}}(0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Here all units are SI so that  $\mathbf{x}(t)$  is in meters, etc. However, any consistent set of units can be used. This is a simple example with only two degrees of freedom so chosen to fit the given space limitations. The procedure is, however, not dependent on such low order. The matrix  $M^{-1}$  in this case is simply

$$M^{-1} = \begin{bmatrix} \frac{1}{9} & 0 \\ 0 & 1 \end{bmatrix}$$

so that

$$M^{-1} K = \begin{bmatrix} 3 & -0.333 \\ -3 & 3 \end{bmatrix} \quad \text{and} \quad M^{-1} C = \begin{bmatrix} 0.3 & -0.033 \\ -0.3 & 0.3 \end{bmatrix}$$

and the state matrix becomes

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -3 & 0.333 & 0.3 & 0.033 \\ 3 & -3 & 0.3 & -0.3 \end{bmatrix}$$

where the state vector and forcing vector are

$$\mathbf{z} = \begin{bmatrix} x_1 \\ x_2 \\ \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} \quad \text{and} \quad \mathbf{F}(t) = \begin{bmatrix} 0 \\ 0 \\ \sin 3t \\ 0 \end{bmatrix}$$

respectively. Next, Eq. (18.10) can be applied with these values. In the past it was required of the vibration engineer to program Eq. (18.10) or some other versions of it. However, commercial software allows the matrix  $A$  to be computed as well as fairly sophisticated simulation.

## 18.5 Commercial Software for Simulation

A variety of simple-to-use, efficient and relatively inexpensive interactive software packages are available for simulating the response. Such programs reduce by a factor of 10 the amount of computer code that actually has to be written by the user. Some examples of available software containing numerical integration packages are MATLAB® and Mathcad. Many finite element packages also contain numerical integration routines. Here we illustrate the use of MATLAB to simulate the result of the simple example above and to print the results. The MATLAB code is listed in Table 18.1 and the output is plotted in Fig. 18.2. The algorithm used in Table 18.1 is a modification of the formulas given in Eq. (18.9), known as a Runge-Kutta-Fehlberg integration method. This method uses a fourth- and fifth-order pair of formulas and an automatic step size.

**Table 18.1** MATLAB Code.

---

MATLAB code for computing and plotting the displacement versus time response of the two-degree-of-freedom system in the text. The % symbol denotes comments. Part A indicates how to input the given system and part B illustrates how to integrate and plot the response.

Part A

```
function zdot = system (t, z)
%This m-file defines the mechanical properties of the system being studied.
%The input is the current time and the previous state vector (initial conditions).
%The output is obtained by solving the state equation: zdot = A * z + f
%First, the mass, damping and stiffness matrices are defined.
M = [9, 0; 0, 1];
C = [2, 7, -0.3; -0.3, 0.3];
K = [27, -3; -3, 3];
f = [0; 0; sin(3 * t); 0];
%The vector of external forces is defined next.
%The state matrix is assembled.
```

```
A = [zeros(2, 2) eye(2, 2); -inv(M) * K - inv(M) * C];
```

Part B

```
%This is the main file used in the simulation.
```

```
%First the initial state is defined.
```

```
z0 = [0.1; 0; 0; 0];
```

```
%Then, the solution is obtained through numerical integration using the ode command calling the "system" input  
file proposed in part A.
```

```
ti = 0; %Initial time of the simulation
```

```
tf = 50; %Final time of the simulation
```

```
[time,solution] = ode45('system', ti, tf, z0); %Perform integration
```

```
%The displacement of each degree-of-freedom is plotted.
```

```
subplot (2,1,1);
```

```
plot (time,solution(:,1));
```

```
xlabel('t[s]');
```

```
ylabel('z1[m]');
```

```
title('First DOF');
```

```
subplot (2,1,2);
```

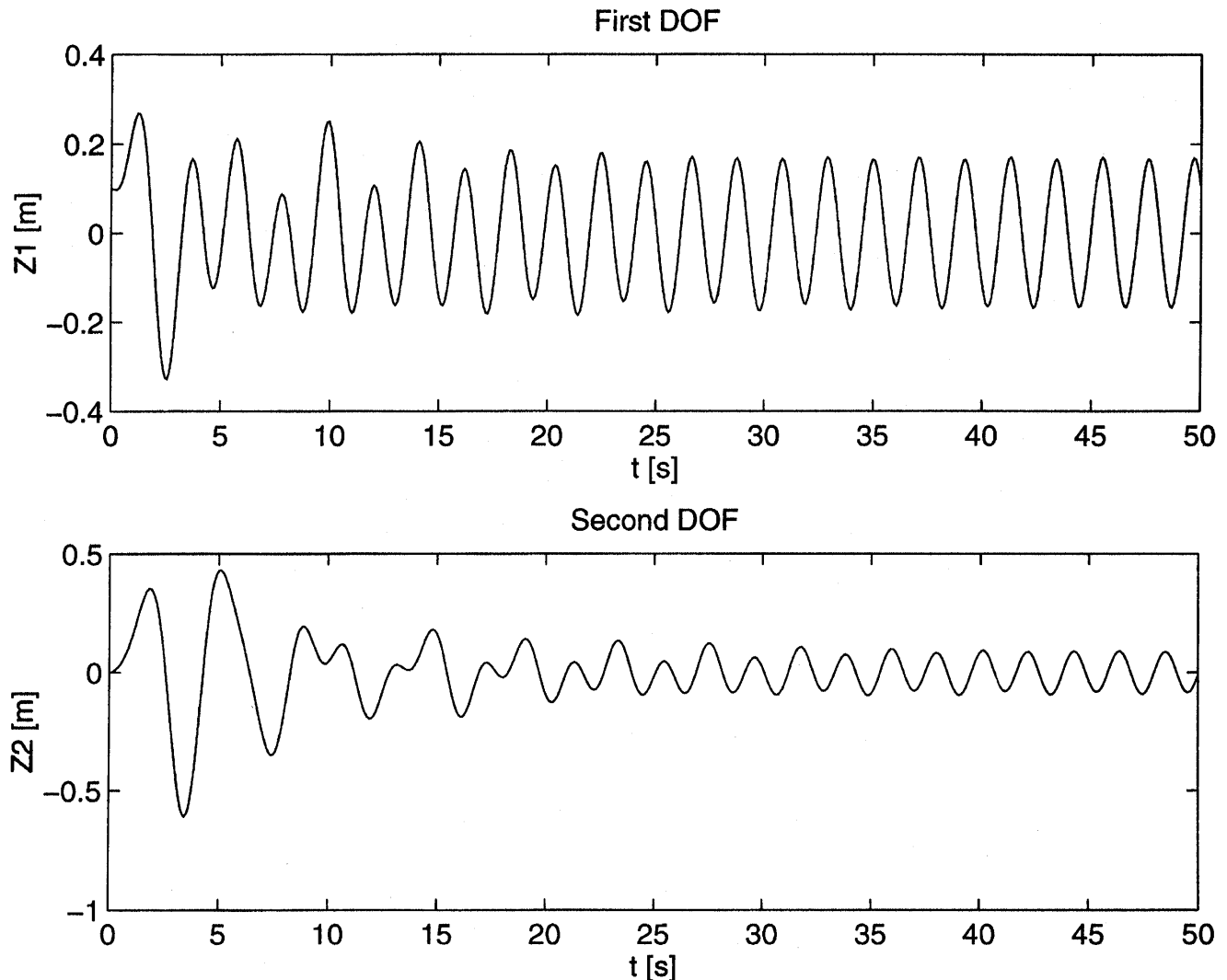
```
plot(time,solution(:,2));
```

```
xlabel('t[s]');
```

```
ylabel('z2[m]');
```

```
title('Second DOF');
```

**Figure 18.2** The output of the MATLAB Code of Table 18.1 illustrating the simulation of the displacement versus time response of the system given by Eq. (18.11).



The use of software such as MATLAB is becoming extremely commonplace. The slide rule has given way to the calculator, and the calculator to the personal computer. Combined with commercial software, the simulation of large and complex vibration problems can be performed without resorting to writing code in lower-level languages. This time savings allows the vibration engineer more time to devote to design and analysis. It is however important to note that simulation through numerical integration is still an approximation and as such is subject to error—both **formula errors** and **round-off errors**. These should be well understood by the user.

## Defining Terms

**Euler method:** A simple numerical solution to a first order ordinary differential equation based on approximating the derivative by a slope.

**Formula error:** Error in the computed response due to the difference between the exact solution and the approximate formula.

**Nomograph:** A graph of displacement, velocity, and acceleration versus frequency for a single-degree-of-freedom system.

**Round-off error:** Error in the computed response due to numerical round-off and truncation in computer arithmetic.

**Runge-Kutta method:** A numerical solution to a first-order ordinary differential equation based on approximating the derivative by several estimates of the first few terms of a Taylor series expansion of the solution.

**Simulation:** Numerical integration to solve an (ordinary) differential equation using time steps to produce the time history of the response (of a vibrating system).

## References

- Boyce, W. E. and DePrima, P. C. 1986. *Elementary Differential Equations and Boundary Value Problems*, 4th ed. John Wiley & Sons, New York.
- Forsythe, G. E., Malcolm, M. A., and Moler, C. B. 1977. *Computer Methods for Mathematical Computation*. Prentice Hall, Englewood Cliffs, NJ.
- Inman, D. J. 1994. *Engineering Vibration*. Prentice Hall, Englewood Cliffs, NJ.
- Macinante, J. A. 1984. *Seismic Mountings for Vibration Isolation*. John Wiley & Sons, New York.
- Moler, C. B. 1980. *MATLAB Users' Guide Technical Report CS81-1*. Department of Computer Sciences, University of New Mexico, Albuquerque.

## Further Information

Further information can be found by consulting the references. More information on MATLAB can be obtained from

The MATHWORKS  
24 Prime Park Way  
Natick, MA 01760-1500

Information on Mathcad can be obtained from

MathSoft, Inc.  
101 Main  
Cambridge, MA 02142

Baird, T. W. "Test Equipment and Measuring Instruments"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Test Equipment and Measuring Instruments

---

## 19.1 Vibration and Shock Test Machines

Vibration Test Machines • Shaker Performance Considerations • Shock Test Machines • Shock Test Machine Performance Considerations

## 19.2 Transducers and Signal Conditioners

Transducers • Transducer Performance Considerations • Signal Conditioners

## 19.3 Digital Instrumentation and Computer Control

### Terrence W. Baird

*Hewlett-Packard Company*

Dynamic test and measurement finds application in many engineering and scientific disciplines. Although each has evolved uniquely in its types of equipment and methods employed, there exists a degree of commonality in the operating principles and performance criteria of the various apparatuses used. Rather than attempting comprehensive treatment of each application or equipment type available, focus will be directed toward this commonality through a discussion of the components of a generalized test system.

The generalized model from which the discussion will be developed is as follows. A device is subjected to a specified dynamic environment produced by a test machine. The test machine input and device response motions are measured by **transducers**, whose signals are conditioned and subsequently analyzed for purposes of data reduction and test machine control. Descriptions will be limited to some commonly used test equipment and instrumentation and their key performance criteria.

Although the model described will be used as the basis for this introduction to test equipment, it is the author's intent that the technical content and principles of equipment performance be relevant to any application. The reader is encouraged to research the references and resources cited at the end of the chapter to more fully explore today's dynamic test technology on an application-specific basis.



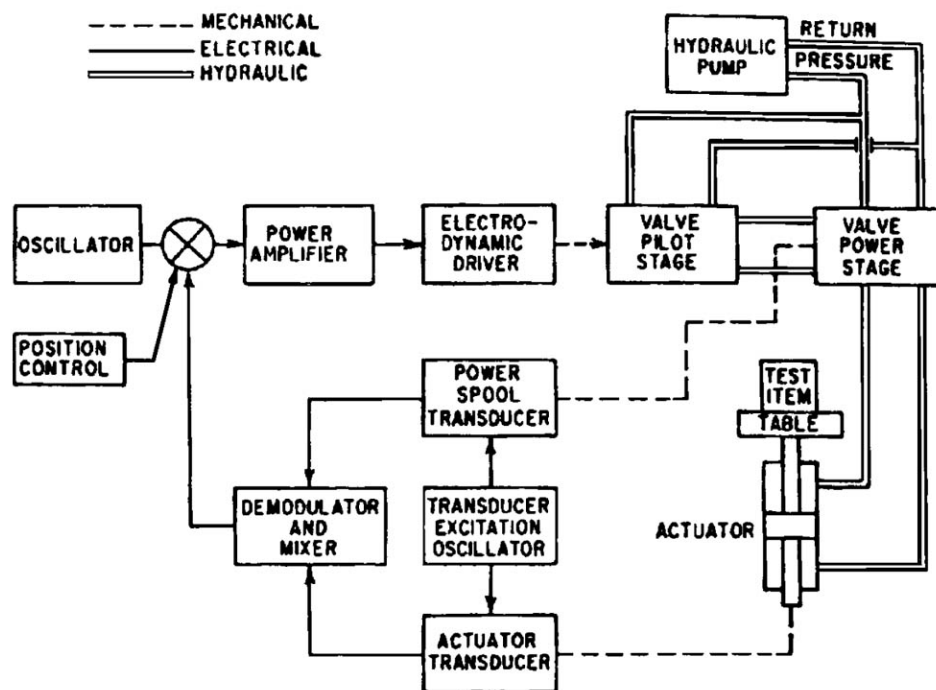
## 19.1 Vibration and Shock Test Machines

### Vibration Test Machines

Vibration test machines, also referred to as *shakers*, are available in several distinct designs. Two common shaker designs are electrohydraulic and electrodynamic, whose names are based on the method of force generation.

Electrohydraulic shakers generate force through electrically controlled hydraulics where power is converted from the high-pressure flow of hydraulic fluid to the vibratory motion of the shaker's table. Figure 19.1 [Unholtz, 1988] illustrates a block diagram of a typical electrohydraulic shaker system and Fig. 19.2 shows an actual system. Availability of large force generation, long displacement stroke and low-frequency performance are advantages of an electrohydraulic shaker.

**Figure 19.1** Block diagram of an electrohydraulic shaker. (Source: Harris, C. and Crede, C. 1988. *Shock and Vibration Handbook*, 3rd ed. McGraw-Hill, New York. Reproduced with permission.)

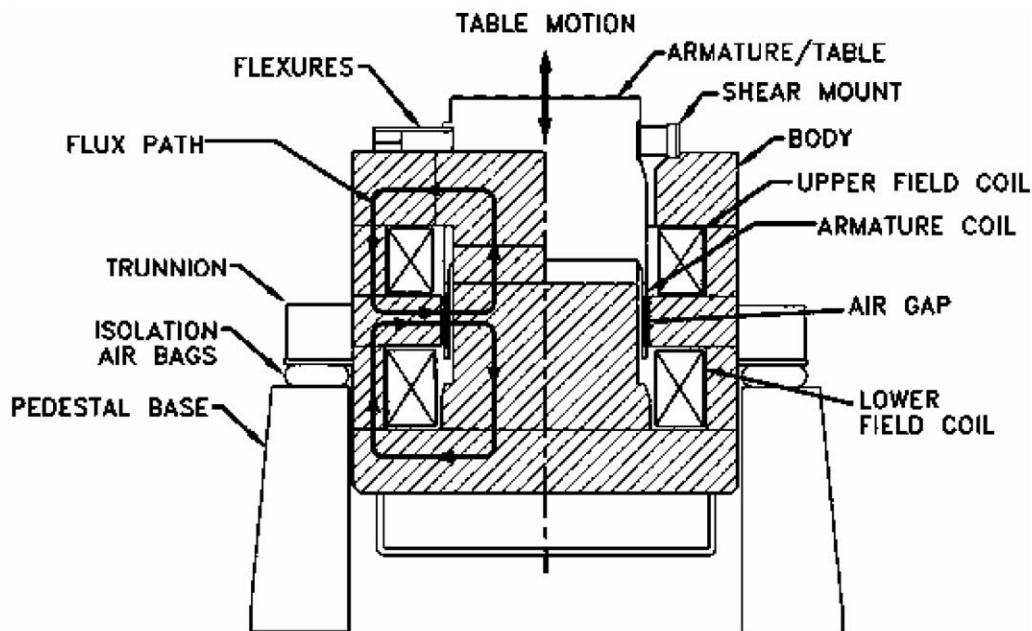


Electrodynamic (also called *electromagnetic*) shakers generate force by means of a coil carrying a current flow placed in a magnetic field, which passes through the coil, causing motion of the coil, moving element, and table assembly (called the *armature*). The operating principle is not unlike that of a loudspeaker system. Some electrodynamic shakers use a "double-ended" design, which incorporates both an upper and lower field coil, resulting in reduced stray magnetic fields above the table and increased operating efficiency. A typical double-ended shaker design is shown in Fig. 19.3. Higher frequency performance is a major advantage of an electrodynamic shaker.

**Figure 19.2** Electrohydraulic shaker system including hydraulic supply, shaker, and digital controller. (Courtesy Lansmont Corp.)



**Figure 19.3** Schematic of double-ended electrodynamic shaker design. (Courtesy Unholtz-Dickie Corp.)



## Shaker Performance Considerations

The following criteria should be evaluated relative to the application regardless of shaker type.

1. *Force rating.* The maximum force available is typically specified as a continuous rating for sine vibration through a usable frequency range. Estimated acceleration performance can be determined from:

$$A = F/W \quad (19.1)$$

where

$A$  = maximum acceleration,  $g$

$F$  = force rating in pounds-force, lbf

$W$  = total load, lb, including armature, table, and test specimen weight

2. *Frequency range.* Frequency versus amplitude performance is generally specified in a series of performance curves presented for various test loads. Representative ranges for typical general-purpose shakers are 1–500 Hz for electrohydraulic and 10–3000 Hz for electrodynamic, depending on test parameters.
3. *Waveform quality/harmonic distortion.* This will vary by design but should be specified.
4. *Magnetic fields.* This may be a concern for some applications, in which case it should be specified for electrodynamic shakers. It is not a concern with electrohydraulic shakers.
5. *Table or head expander **frequency response**.* The practical frequency range will depend, in part, on the table attached to the shaker. Basic design, mass, damping, and frequency response characteristics should be specified.
6. *Test orientation.* Consideration should be given to design type if independent vertical and horizontal test capability is desired. For instance, it is relatively common for an electrodynamic shaker to be supported by a base with a trunnion shaft whereby the entire body can be rotated about its center providing for either vertical or horizontal vibration.

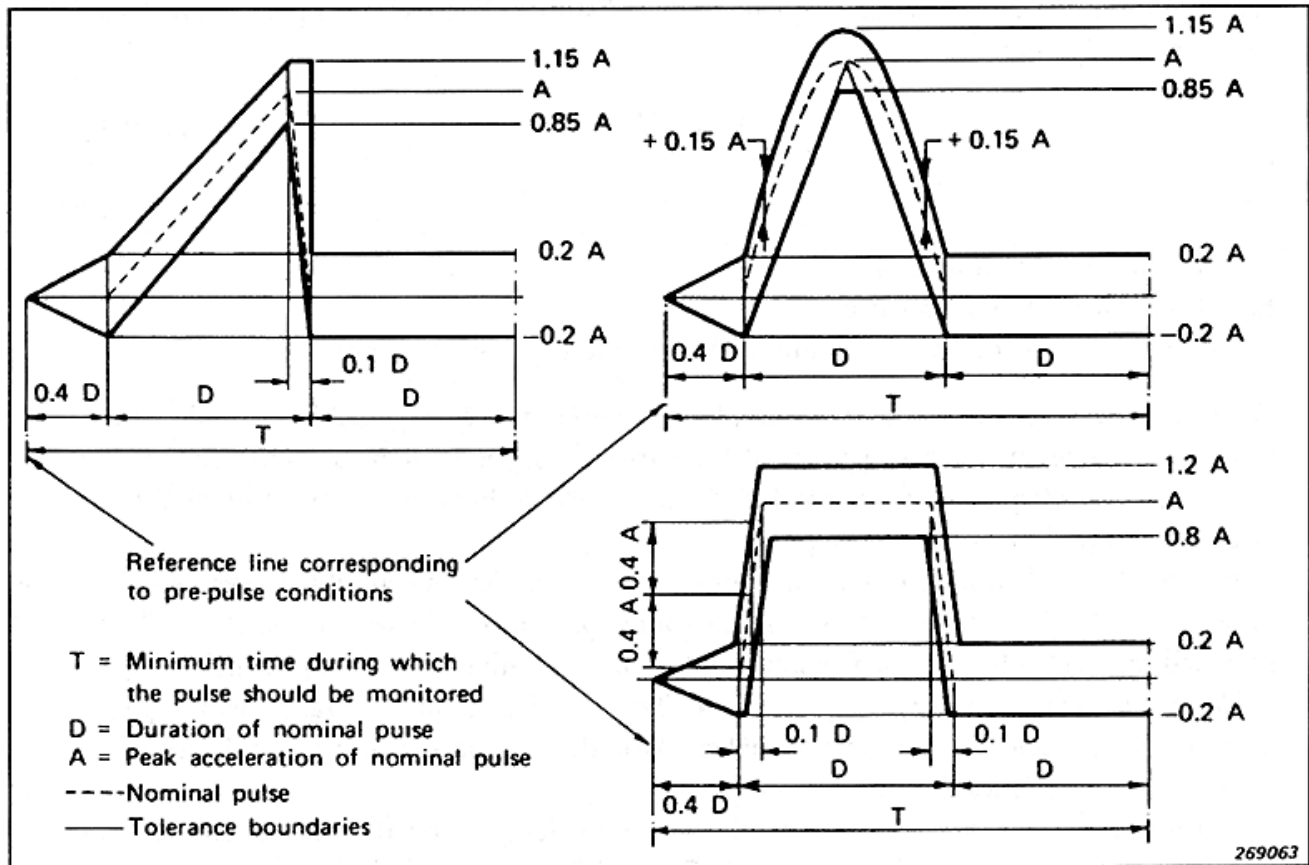
## Shock Test Machines

Three specification types are used in defining a shock test:

1. Specification of the shock test machine (also called *shock machine*) including mounting and operating procedures. Shock machines unique to a specification will not be discussed.
2. Specification of shock motion described by simple shock pulse waveforms and parameters of peak acceleration, duration, and **velocity change**. An example is shown in [Fig. 19.4](#).
3. Specification of the **shock response spectrum (SRS)** that the test produces.

Two types of machines used for shock generation are the free-fall shock machine and the previously described shakers.

**Figure 19.4** Preferred shock pulse waveforms of IEC 68-2-27. (Source: Broch, J. T. 1984. Mechanical Vibration and Shock Measurement, 2nd ed. Bruel & Kjaer, Naerum, Denmark. Reproduced with permission.)



A typical free-fall shock machine (sometimes called a *drop table* or *drop tester*) is shown in Fig. 19.5 and is used for generating simple or classical pulse waveforms such as those in Fig. 19.4. Operation is straightforward. The shock table, whose orientation and free-fall path is controlled by guide rods, is raised to a desired height and allowed to fall and impact upon a pulse programmer. Brakes are employed to prevent multiple impacts after rebound. Shock pulse velocity change is controlled by drop height. Pulse waveform, resulting peak acceleration, and duration are determined by programmer type. Significant velocity and peak acceleration capabilities are advantages of this type of shock machine.

**Figure 19.5** Shock machine and associated instrumentation. (Courtesy MTS Systems Corp.)





Shakers with appropriate digital controllers are not only capable of producing classical waveforms, but can also be used for applications such as SRS programming or capturing a real-world shock pulse and using it as a control waveform for subsequent shock tests. Shakers have limitations for shock test in terms of available displacement, velocity, and peak acceleration. However, for low-level shocks of light- to medium-weight specimens, test flexibility, control, and repeatability are excellent if the test parameters are within the performance limits of the shaker.

## Shock Test Machine Performance Considerations

Performance criteria that should be evaluated against the intended application include:

1. *Maximum peak acceleration.* Some shakers can achieve 100–200 *g*. A general-purpose free-fall shock machine is usually limited to less than 1000 *g*, whereas a high-performance shock machine may be capable of 20 000 *g* for lightweight specimens.
2. *Pulse duration*, maximum and minimum.
3. *Velocity change.* Shakers will normally be limited to less than 100 inches/second (ips). A general-purpose free-fall shock machine will achieve 300 ips and a high-performance shock machine may be capable of 1000 ips. For very low velocity changes (0–50 ips), control and repeatability can be difficult with a free-fall shock machine.
4. *Waveform flexibility* and programmer design.
5. *Table size and weight capacity.*
6. *Table performance*, frequency response, damping, and waveform quality.
7. Potential need for *SRS control* and *waveform synthesis*.

## 19.2 Transducers and Signal Conditioners

---

### Transducers

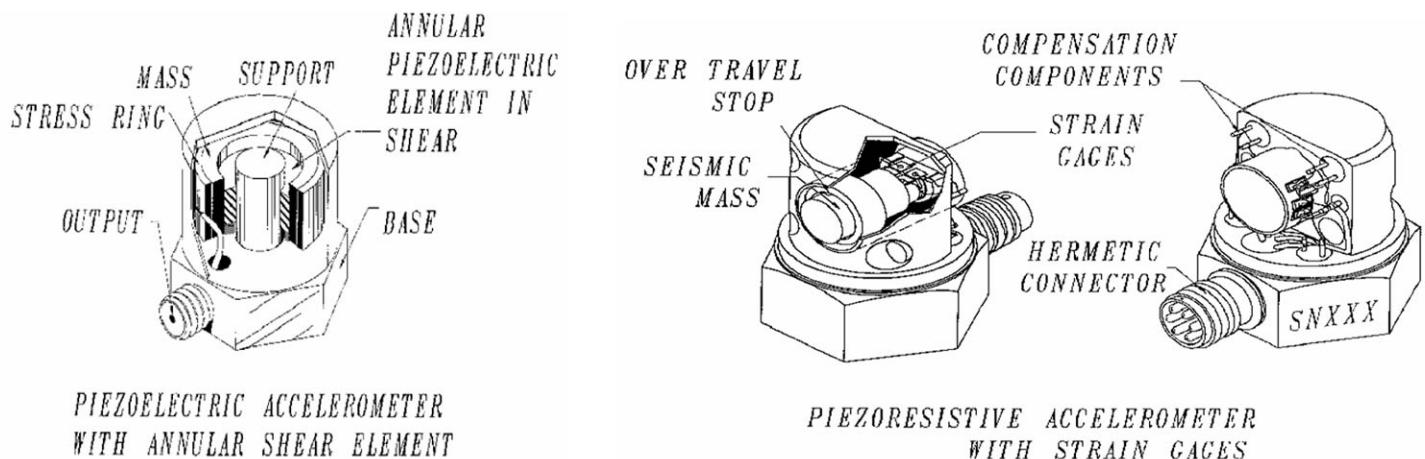
The ANSI/ISA definition of transducer is "a device which provides a usable output in response to a specified measurand" [[ANSI/ISA S37.1](#)]. The measurand is the physical quantity to be measured and the output is an electrical quantity whose amplitude is proportional to that of the measurand. The measurand is the primary descriptor of a transducer type and for dynamic testing would likely be displacement, velocity, or

acceleration. The most common measurand is acceleration, so this discussion will be limited to acceleration transducers, more commonly referred to as *accelerometers*. Two common categories of accelerometers are piezoelectric and piezoresistive, which differ fundamentally in their electrical transduction principles.

*Piezoelectric* (PE) accelerometers incorporate sensing elements, typically quartz or ceramic crystals, that have the property of producing a charge when mechanically stressed. Specifically, when subjected to an acceleration, the PE accelerometer produces a charge proportional to the applied acceleration and is said to be *self-generating* in that the electrical output is produced without the need for auxiliary power excitation to the accelerometer. Typically, this charge then needs to be converted to a voltage in an external signal conditioner for subsequent analysis or readout. An exception to this rule is the now commonly available PE accelerometer with built-in integrated-circuit signal conditioning.

*Piezoresistive* (PR) accelerometers incorporate a semiconductor material such as a solid state silicon resistor, which serves as a strain-sensing or strain gage element. Arranged in pairs and typically connected electrically in a Wheatstone-bridge circuit, these PR elements exhibit a change in electrical resistance proportional to an applied acceleration. The PR accelerometer is referred to as a *passive* type accelerometer in that it does require an external power source to operate. The PR accelerometer's primary advantage is its ability to measure down to DC or steady-state acceleration, making it particularly suitable to long-duration pulses and other low-frequency applications. PE and PR accelerometer designs are shown in [Fig. 19.6](#).

**Figure 19.6** Accelerometer designs. (Courtesy Endevco Corp.)



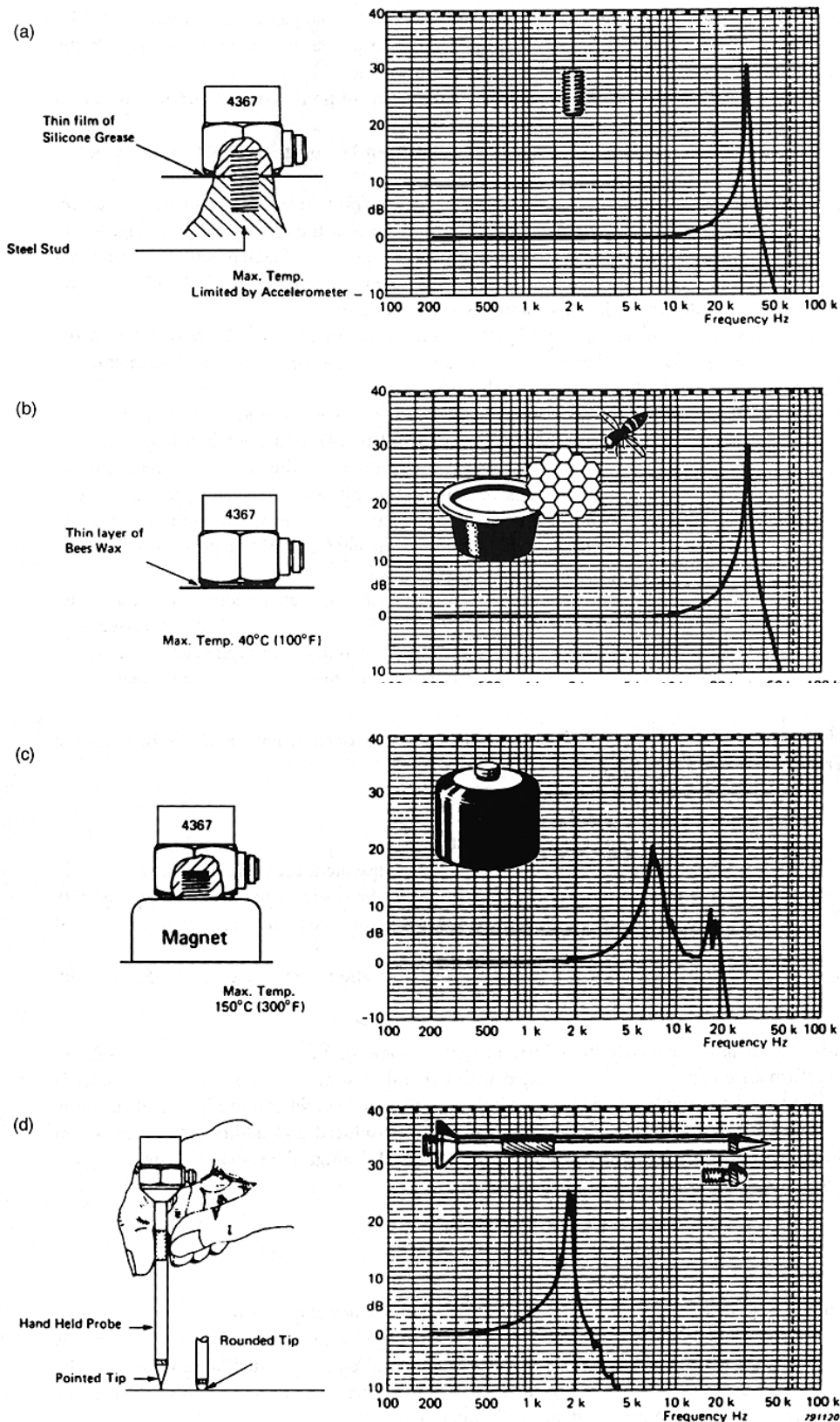
## Transducer Performance Considerations

The following performance criteria should be evaluated prior to an accelerometer's use. The first three are most critical and for a given design are interrelated, representing the likely trade-offs or compromises in accelerometer selection.

1. *Sensitivity*. Defined as the ratio of change in output to change in acceleration, sensitivity is expressed as coulomb/g or volt/g, depending on accelerometer type. The higher the sensitivity is, the greater the system signal-to-noise ratio will be.
2. *Frequency response*. Both low- and high-frequency response may be important to an application.
3. *Mass and size*. Accelerometer weights can range from 1 to 60 g. Typically, minimizing size and mass is desirable.
4. *Mass loading effect*. Mounting an accelerometer with finite mass onto a structure changes the mechanics of the structure at that point. If the mass of the accelerometer is a significant percentage of the effective mass of the structure at the point of attachment, the structure's frequency response will be altered, resulting in a poor measurement. A simple rule of thumb or exercise to determine if mass loading is a problem is to:
  - a. Measure a frequency response function of the structure using the desired accelerometer.
  - b. Mount a second accelerometer of the same mass at the same point of attachment (i.e., mass is now doubled) and repeat the measurement.
  - c. Compare the two measurements for amplitude changes and frequency shifts. If differences are significant, then mass loading is a problem. (Although not discussed in the text, some measurement situations do exist in which mass loading effects, extreme surface temperatures, rotating structures, or other test conditions preclude the practical use of a *contact* transducer, such as an accelerometer. In these instances, a *noncontact* transducer can be employed—such as a *laser Doppler vibrometer* for motion detection—where the *electro-optical* transduction principle is employed.)
5. *Amplitude range and linearity*. Sensitivity is constant within stated tolerances over a certain amplitude range, beyond which sensitivity is nonlinear. This is usually expressed as a percentage deviation from nominal sensitivity as a function of the applied acceleration.
6. *Transverse sensitivity*. For a single-axis accelerometer, there is still a small sensitivity to transverse accelerations, which is usually expressed as a percentage of main axis sensitivity.
7. *Temperature sensitivity*. Percent deviation from the nominal sensitivity is expressed as a function of temperature.
8. *Mounting considerations*. Although not a characteristic of accelerometer design, the way in which an accelerometer is mounted to the structure for measurement has a significant influence on its effective frequency response. [Figure 19.7](#) shows various methods used to mount accelerometers and their effect on frequency response.



**Figure 19.7** Typical accelerometer mounting techniques and relative frequency response characteristics. (Source: Broch, J. T. 1984. *Mechanical Vibration and Shock Measurement*, 2nd ed. Bruel & Kjaer, Naerum, Denmark. Reproduced with permission of Bruel & Kjaer.)  
(continues)



## Signal Conditioners

It is typical for a signal conditioner to be located in the measurement system between the transducer and the final readout or recording instruments. Signal conditioners range from simple to sophisticated and can employ internal electronics for significant signal modification or calculation of related physical quantities. Rather than describe the operating principles of signal conditioners, the following simply lists some of the key functions and available features.

1. Supply excitation voltage to a passive-circuit transducer (e.g., PR accelerometer)
2. Charge conversion to voltage from a PE accelerometer
3. "Dial-in" accelerometer sensitivity normalization
4. Gain or attenuation control to provide optimum signal-to-noise ratio in the readout instrument
5. Low-pass or high-pass filters
6. Grounding options
7. Internal electronics to perform functions such as single or double integration to obtain velocity or displacement data from acceleration signals

## 19.3 Digital Instrumentation and Computer Control

---

It is assumed the reader is familiar with the application and operating principles of time-based instruments such as the oscilloscope. Time-based instruments will be not be discussed in this text.

Today's laboratories are commonly equipped with fast Fourier transform (FFT) analysis and digital microprocessor control capabilities both for data analysis and reduction as well as test machine signal control. It is beyond the scope of this chapter to provide a comprehensive treatment of signal analysis and computer control techniques or of the many software, firmware, and hardware platforms in which these capabilities are available. The reader can find such treatments in several of the references cited at the end of this chapter. Rather, it is the author's intent to simply recognize some of the functionality and features available to the user in applying these techniques and instruments.

*Dynamic signal or FFT analyzers* may be used independently of test machine control

for purposes of data collection and analysis. Available functionality includes:

1. Time domain or frequency domain analysis
2. Shock analysis
  - Waveform capture
  - Digital filtering
  - Math capabilities such as integration, differentiation, and multiaxis vector resolution
  - SRS computation
3. Vibration analysis
  - Power/auto spectrum analysis
  - Frequency response and transfer functions
  - Band power, harmonic power, and harmonic distortion measurements
  - Waterfall analysis
4. Programming capabilities to allow user-defined functionality

*Digital control for vibration testing* is generally tailored specifically to closed-loop shaker control and would normally contain fewer general analysis capabilities as compared to an independent dedicated FFT analyzer. Functions typically include:

1. Sine or random vibration control
2. Swept-sine on random
3. Narrow band random on random
4. Multiple control and response channels
5. Test control and abort limits
6. Capability of using field vibration data as shaker input signal
7. Basic vibration data analysis (e.g., FFT, frequency response, etc.)

*Digital control for shock-testing* refers specifically to the case of controlling a shaker for shock test generation. Functions typically include:

1. Classical waveform control (e.g., half sine)
2. Direction of shock and multiple shock control
3. Transient capture of real-world shock pulses and subsequent use as shaker control
4. SRS and waveform synthesis, a function allowing the operator to specify a required SRS to the controller, which in turn synthesizes a control waveform

resulting in the desired SRS

5. Basic waveform analyses such as single/double integration, vector resolution, FFT, and SRS computation

This section has provided a brief introduction to some of the test equipment, capabilities, and performance considerations associated with dynamic test and measurement. The reader is reminded that the examples cited represent only a small sample of what is available and currently in use. More thorough presentation of the concepts introduced and additional discussion on application-specific methods and equipment can be found through the resources provided at the end of the section.

## Defining Terms

**Frequency response:** As a transducer characteristic, frequency response is the change of transducer sensitivity as a function of frequency. Normally, an operating frequency range is specified over which the sensitivity does not vary more than a stated percentage from the rated sensitivity. More generally, for a mechanical system, frequency response is a ratio of output response to input excitation as a function of frequency.

**g:** The acceleration produced by the force of gravity. Acceleration amplitudes are commonly described as multiples of  $g$ , where

$$1\ g = 980.665\ \text{cm/s}^2 = 386.087\ \text{in./s}^2 = 32.1739\ \text{ft/s}^2.$$

**Shock response spectrum (SRS):** Also called *shock spectrum*, the SRS is a curve that indicates a theoretical maximum response as a function of pulse duration and responding system natural frequency. The shock spectrum of a waveform is an indication of the shock's damage potential in the frequency domain.

**Transducer:** A device that provides a usable output in response to a specified measurand [[ANSI/ISA S37.1-1975](#)].

**Velocity change:** The acceleration-time integral or the area under an acceleration-time shock pulse waveform. It is a function of the energy of the shock pulse and can be related to other physical quantities such as equivalent free-fall drop height.

## References

- ANSI/ISA S37.1-1975. *Electrical Transducer Nomenclature and Terminology*. American National Standard, Instrument Society of America, Research Triangle Park, NC.
- Broch, J. T. 1984. *Mechanical Vibration and Shock Measurement*, 2nd ed. Bruel & Kjaer, Naerum, Denmark.

- IEC 68-2-27. 1987. *Basic Environmental Testing Procedures Part 2: Tests-Test Ea: Shock*. International Electrotechnical Commission, Geneva, Switzerland.
- Unholtz, K. 1988. Vibration testing machines. In *Shock and Vibration Handbook*, 3rd ed., -16. McGraw-Hill, New York.

## Further Information

### Texts

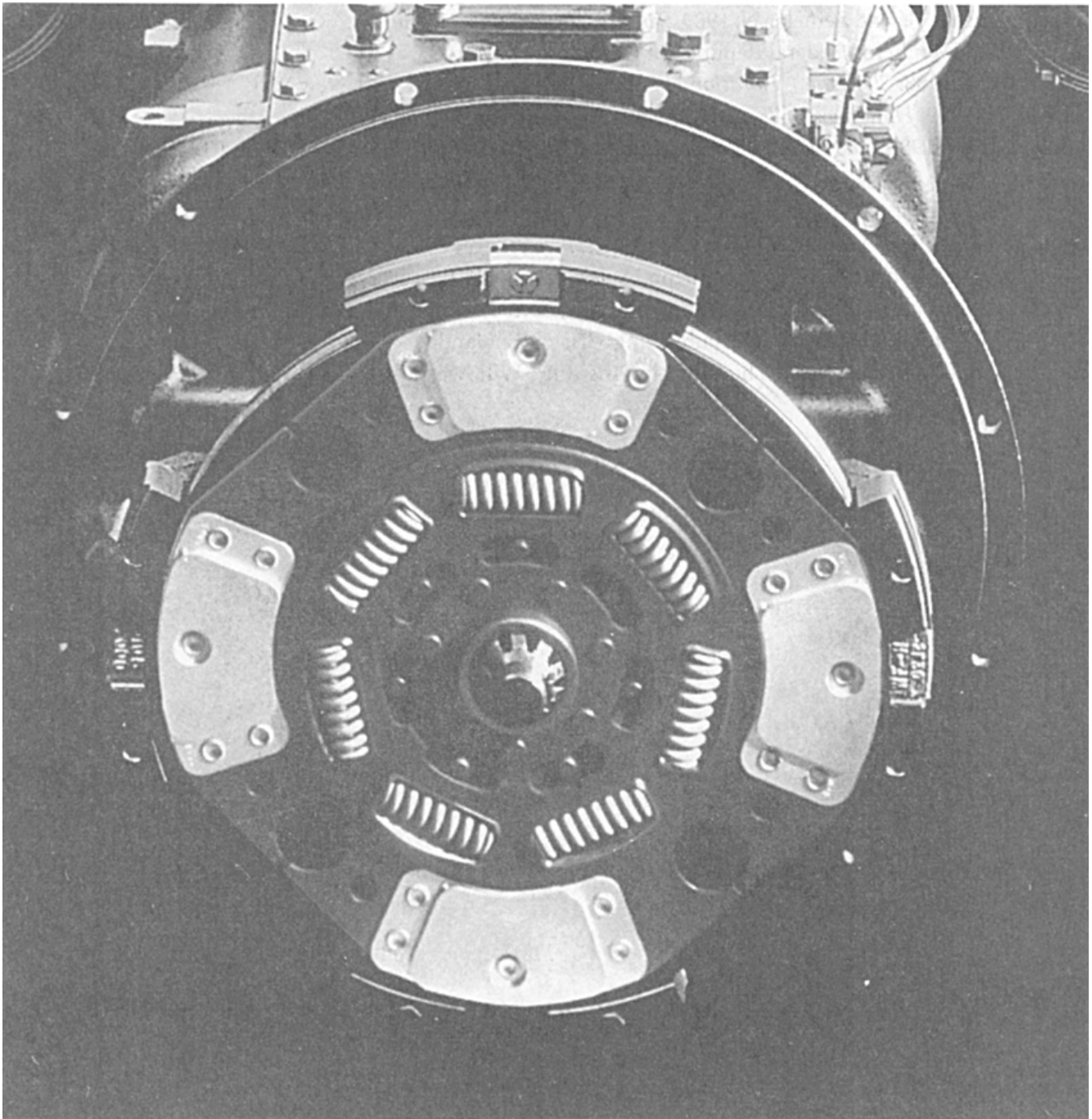
- Bendat, J. S. and Piersol, A. G. 1971. *Random Data: Analysis and Measurement Procedures*. Wiley-Interscience, New York.
- Gopel, W., Hesse, J., and Zemel, J. N. 1989. *Sensors: A Comprehensive Survey*. VCH Verlagsgesellschaft MBH, Weinheim, Federal Republic of Germany.
- Harris, C. M. 1988. *Shock and Vibration Handbook*, 3rd ed. McGraw-Hill, New York.
- Norton, H. 1989. *Handbook of Transducers*. Prentice Hall, Englewood Cliffs, NJ.
- Ramirez, R. W. 1985. *The FFT, Fundamentals and Concepts*. Prentice Hall, Englewood Cliffs, NJ.

### Journals

- Sound and Vibration*, Acoustical Publications, Inc., Bay Village, OH.
- Journal of the IES*, Institute of Environmental Sciences, Mount Prospect, IL.
- Journal of Sound and Vibration*, Academic Press Ltd., London, England.
- Noise and Vibration Worldwide*, IOP Publishing, Bristol, England.
- The Shock and Vibration Digest*, The Vibration Institute, Willowbrook, IL.

Ravani, B. "Kinematics and Mechanisms"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000





**THE LONG TRAVEL DAMPER (LTD) CLUTCH**—The introduction of the Long Travel Damper (LTD) clutch by Rockwell has addressed driver concerns of engine and drivetrain torsional vibration. The 15.5", diaphragm-spring, two-plate, pull-type clutch absorbs and dampens vibrations and torque loads passed through from the engine flywheel, providing a smoother ride for drivers and increased drivetrain component life. The LTD is available in three different capacities for use in low, medium and high horsepower ranges and features a fifth rivet to help alleviate clutch drag. (Photo courtesy of Rockwell Automotive.)

# IV

## Kinematics and Mechanisms

---

**Bahram Ravani**

*University of California, Davis*

- 20 **Linkages and Cams** *J. M. McCarthy and G. L. Long*  
Linkages • Spatial Linkages • Displacement Analysis • Cam Design • Classification of Cams and Followers • Displacement Diagrams
- 21 **Tribology: Friction, Wear, and Lubrication** *B. Bhushan*  
History of Tribology and Its Significance to Industry • Origins and Significance of Micro/nanotribology • Friction • Wear • Lubrication • Micro/nanotribology
- 22 **Machine Elements** *G. R. Pennock*  
Threaded Fasteners • Clutches and Brakes
- 23 **Crankshaft Journal Bearings** *P. K. Subramanyan*  
Role of the Journal Bearings in the Internal Combustion Engine • Construction of Modern Journal Bearings • The Function of the Different Material Layers in Crankshaft Journal Bearings • The Bearing Materials • Basics of Hydrodynamic Journal Bearing Theory • The Bearing Assembly • The Design Aspects of Journal Bearings • Derivations of the Reynolds and Harrison Equations for Oil Film Pressure
- 24 **Fluid Sealing in Machines, Mechanical Devices, and Apparatus** *A. O. Lebeck*  
Fundamentals of Sealing • Static Seals • Dynamic Seals • Gasket Practice • O-Ring Practice • Mechanical Face Seal Practice

THIS SECTION COMBINES KINEMATICS AND MECHANISMS and certain aspects of mechanical design to provide an introductory coverage of certain aspects of the theory of machines and mechanisms. This is the branch of engineering that deals with design and analysis of moving devices (or mechanisms) and machinery and their components. Kinematic analysis is usually the first step in the design and evaluation of mechanisms and machinery, and involves studying the relative motion of various components of a device or evaluating the geometry of the force system acting on a mechanism or its components. Further analysis and evaluation may involve calculation of the magnitude and sense of the forces and the stresses produced in each part of a mechanism or machine as a result of such forces. The overall subject of the theory of machines and mechanisms is broad and would be difficult to cover in this section. Instead, the authors in this section provide an introduction to some topics in this area to give readers an appreciation of the broad nature of this subject as well as to provide a readily available reference on the topics covered.

The first chapter is an introductory coverage of linkages and cams. These are mechanisms found in a variety of applications, from door hinges to robot manipulators and the valve mechanisms used in present-day motor vehicles. The scope of the presentation is displacement analysis dealing with understanding the relative motion between the input and output in such mechanisms. The second chapter goes beyond kinematic analysis and deals with the effects of the interactions between two surfaces in relative motion. This subject is referred to as tribology, and it is an important topic in



mechanical design, the theory of machines, and other fields. Tribology is an old field but still has many applications in areas where mechanical movement is achieved by relative motion between two surfaces. Present applications of tribology range from understanding the traction properties of tires used in automobiles to understanding the interfacial phenomena in magnetic storage systems and devices. The third chapter in this section deals with mechanical devices used for stopping relative motion between the contacting surfaces of machine elements or for coupling two moving mechanical components. These include mechanical fasteners, brakes, and clutches. Many mechanical devices and machines require the use of bolts and nuts (which are fasteners) for their construction. Brakes are usually used to stop the relative motion between two moving surfaces, and clutches reduce any mismatch in the speed of two mechanical elements. These components are used in a variety of applications; probably their best-known application is their use in the motor vehicle.

The fourth chapter deals with another mechanical element in the automotive industry, namely, the journal bearing used in the crankshaft of the automotive engine (which is usually an internal combustion engine). The last chapter in this section deals with mechanical seals used to protect against leakage of fluids from mechanical devices and machines. When two mechanical components are brought into contact or relative motion as part of a machine, the gap between the contacting surfaces must be sealed if fluid is used for lubrication or other purposes in the machine. This chapter provides an introduction to the mechanical seals used to protect against leakage of fluids.

In summary, the authors in this section have provided easy-to-read introductions to selected topics in the field of theory of machines and mechanisms that can be used as a basis for further studies or as a readily available reference on the subject.

McCarthy, J. M., Long, G. L. "Linkages and Cams"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

- 20.1 Linkages
- 20.2 Spatial Linkages
- 20.3 Displacement Analysis
- 20.4 Cam Design
- 20.5 Classification of Cams and Followers
- 20.6 Displacement Diagrams

**J. Michael McCarthy**  
*University of California, Irvine*

**Gregory L. Long**  
*University of California, Irvine*

Mechanical movement of various machine components can be coordinated using linkages and cams. These devices are assembled from hinges, ball joints, sliders, and contacting surfaces and transform an input movement such as a rotation into an output movement that may be quite complex.

## 20.1 Linkages

---

Rigid links joined together by hinges parallel to each other are constrained to move in parallel planes and the system is called a **planar linkage**. A generic value for the **degree of freedom**, or mobility, of the system is given by the formula  $F = 3(n - 1) - 2j$ , where  $n$  is the number of links and  $j$  is the number of hinges.

Two links and one hinge form the simplest *open chain linkage*. Open chains appear as the structure of robot manipulators. In particular, a three-degree-of-freedom planar robot is formed by four bodies joined in a series by three hinges, as in [Fig. 20.1\(b\)](#).

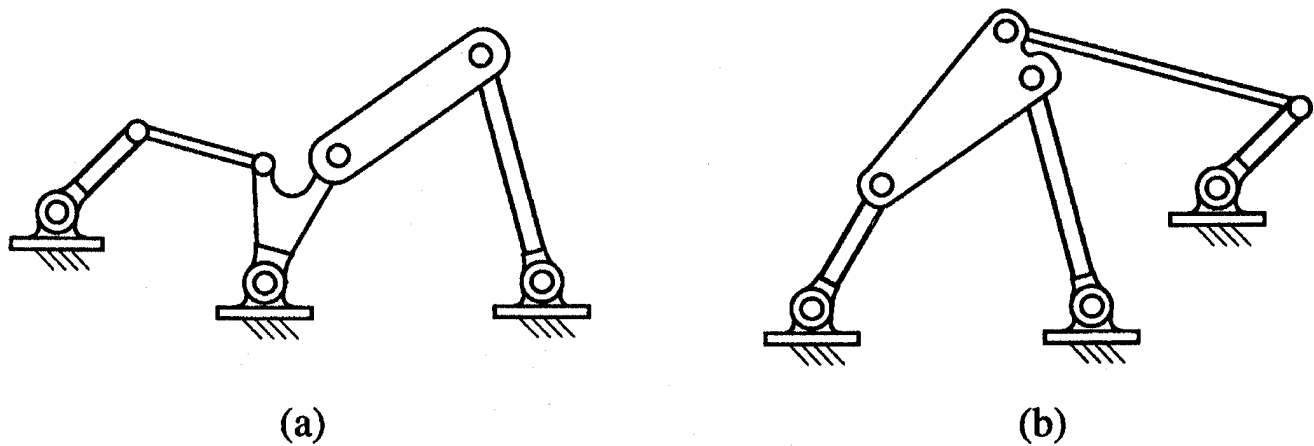
If the series of links close to form a loop, the linkage is a simple *closed chain*. The simplest case is a quadrilateral ( $n = 4$ ,  $j = 4$ ) with one degree of freedom (See [Figs. 20.1\(a\)](#) and [20.3](#)); notice that a triangle has mobility zero. A single loop with five links has two degrees of freedom and one with six links has three degrees of freedom. This latter linkage also appears when two planar robots hold the same object.

A useful class of linkages is obtained by attaching a two-link chain to a four-link quadrilateral in various ways to obtain a one-degree-of-freedom linkage with two loops. The two basic forms of this linkage are known as the Stephenson and Watt six-bar linkages, shown in [Fig. 20.2](#).

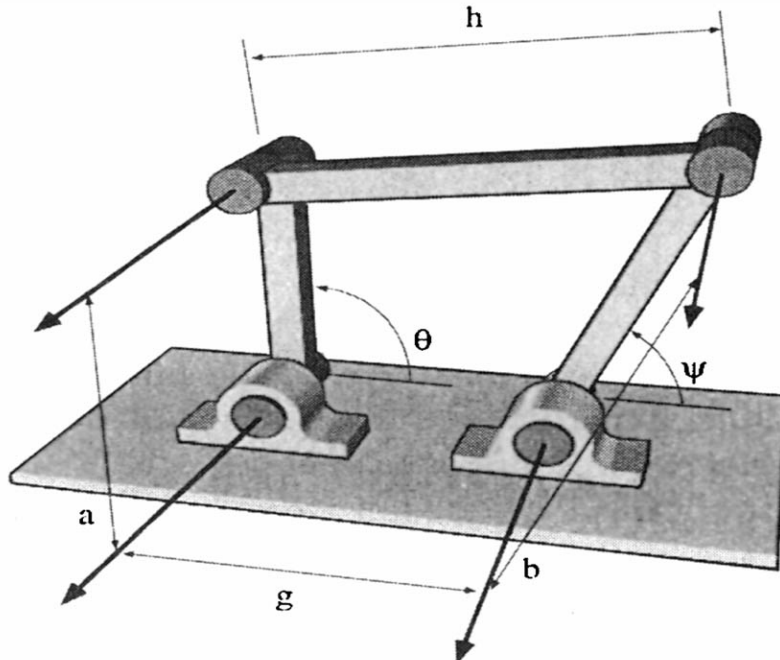
**Figure 20.1** (a) Planar four-bar linkage; and (b) planar robot.



**Figure 20.2** (a) A Watt six-bar linkage; and (b) a Stephenson six-bar linkage.



**Figure 20.3** Dimensions used to analyze a planar 4R linkage.



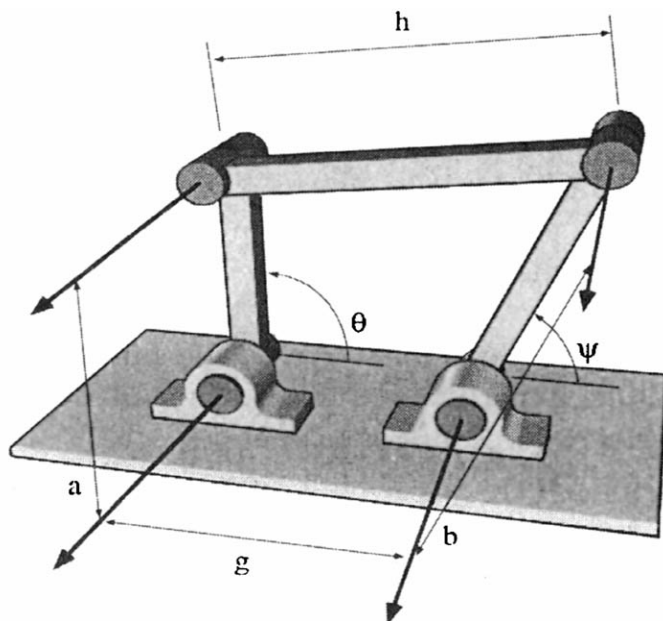
In each of these linkages a sliding joint, which constrains a link to a straight line rather than a circle, can replace a hinge to obtain a different movement. For example, a slider-crank linkage is a four-bar closed chain formed by three hinges and a sliding joint.

## 20.2 Spatial Linkages

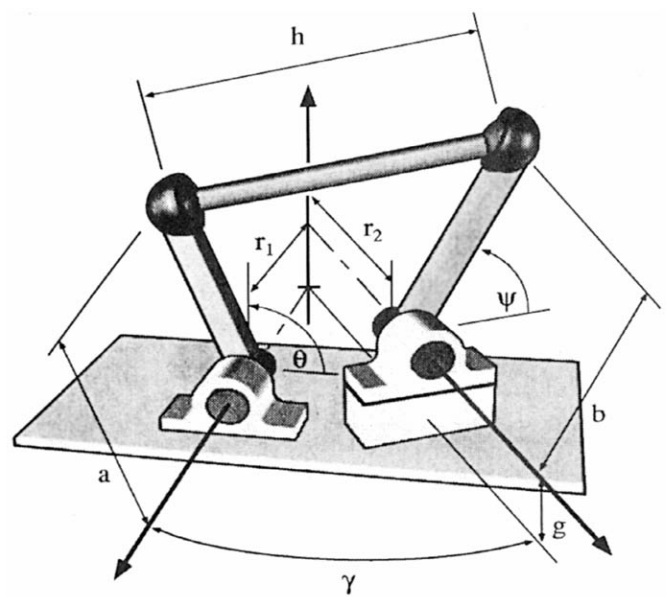
The axes of the hinges connecting a set of links need not be parallel. In this case the system is no longer constrained to move in parallel planes and forms a **spatial linkage**. The robot manipulator with six hinged joints (denoted R for **revolute joint**) is an example of a spatial 6R open chain.

Spatial linkages are often constructed using joints that constrain a link to a sphere about a point, such as a ball-in-socket joint, or a gimbal mounting formed by three hinges with concurrent axes—each termed a **spherical joint** (denoted S). The simplest spatial closed chain is the RSSR linkage, which is often used in place of a planar four-bar linkage to allow for misalignment of the cranks (Fig. 20.4).

**Figure 20.4** A spatial RSSR linkage.



**Figure 20.5** A spherical 4R linkage.



Another useful class of spatial mechanisms is produced by four hinges with concurrent axes that form a spherical quadrilateral known as a **spherical linkage**. These linkages provide a controlled reorientation movement of a body in space (Fig. 20.5).

## 20.3 Displacement Analysis

The closed loop of the planar 4R linkage (Fig. 20.3) introduces a constraint between the crank angles  $\theta$  and  $\psi$  given by the equation

$$A \cos \psi + B \sin \psi = C \quad (20.1)$$

where

$$\begin{aligned} A &= 2gb - 2ab \cos \theta \\ B &= -2ab \sin \theta \\ C &= h^2 - g^2 - b^2 - a^2 + 2ga \cos \theta \end{aligned}$$

This equation can be solved to give an explicit formula for the angle  $\psi$  of the output crank in terms of the input crank rotation  $\theta$ :

$$\psi(\theta) = \tan^{-1} \left( \frac{B}{A} \right) \pm \cos^{-1} \left( \frac{C}{\sqrt{A^2 + B^2}} \right) \quad (20.2)$$

The constraint equations for the spatial RSSR and spherical 4R linkages have the same form as that of the planar 4R linkage, but with coefficients as follows. For spatial RSSR linkage (Fig. 20.4):

$$\begin{aligned} A &= -2ab \cos \gamma \cos \theta - 2br_1 \sin \gamma \\ B &= 2bg - 2ab \sin \theta \\ C &= h^2 - g^2 - b^2 - a^2 - r_1^2 - r_2^2 + 2r_1 r_2 \cos \gamma \\ &\quad + 2ar_2 \sin \gamma \cos \theta + 2ga \sin \theta \end{aligned}$$

For spherical 4R linkage (Fig. 20.5):

$$\begin{aligned} A &= \sin \alpha \sin \beta \cos \gamma \cos \theta - \cos \alpha \sin \beta \sin \gamma \\ B &= \sin \alpha \sin \beta \sin \theta \\ C &= \cos \eta - \sin \alpha \cos \beta \sin \gamma \cos \theta \\ &\quad - \cos \alpha \cos \beta \cos \gamma \end{aligned}$$

The formula for the output angle  $\psi$  in terms of  $\theta$  for both cases is identical to that already given for the planar 4R linkage.

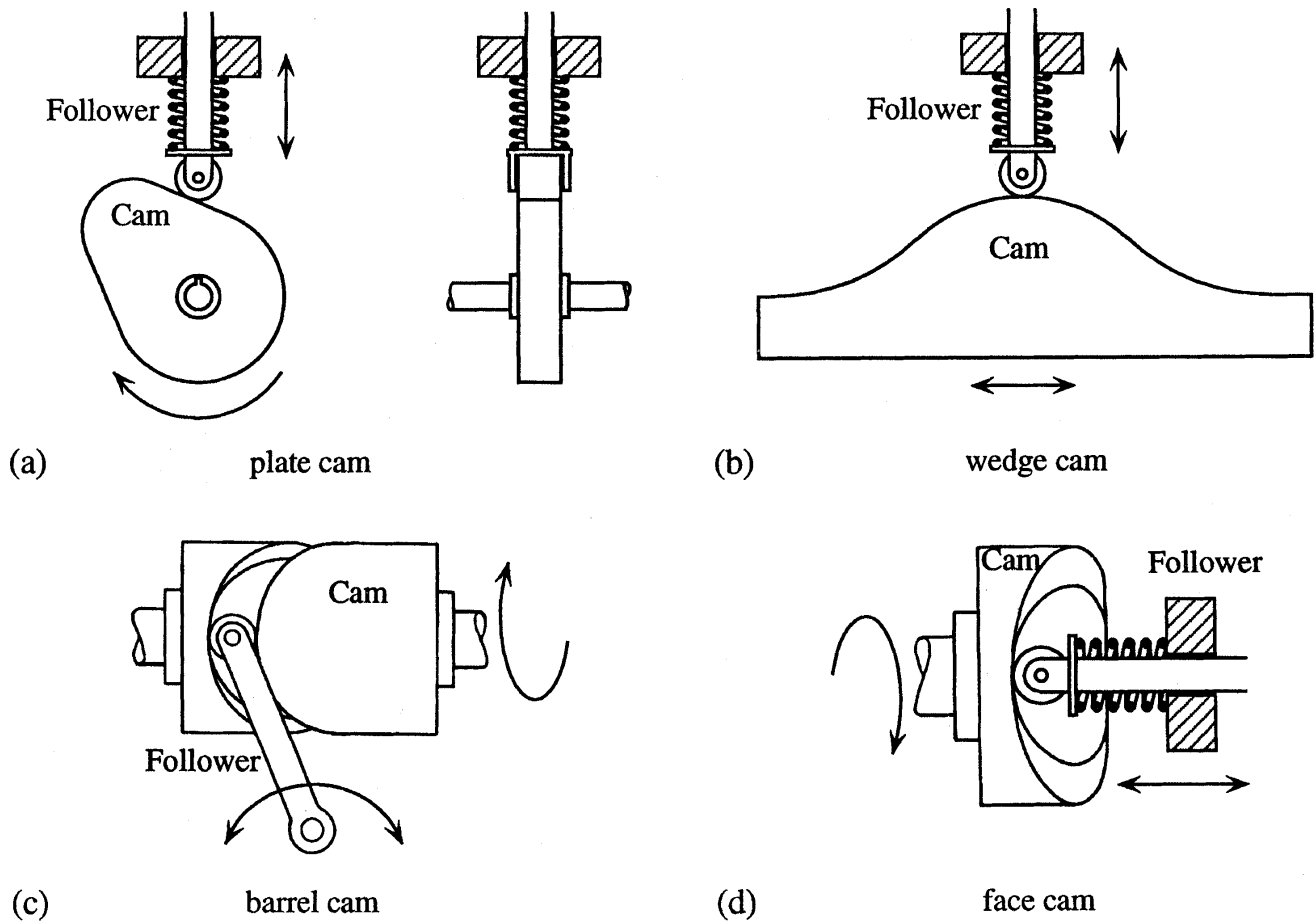
## 20.4 Cam Design

A *cam pair* (or *cam-follower*) consists of two primary elements called the *cam* and *follower*. The cam's motion, which is usually rotary, is transformed into either follower translation, oscillation, or combination, through direct mechanical contact. Cam pairs are found in numerous manufacturing and commercial applications requiring motion, path, and/or function generation. Cam pair mechanisms are usually simple, inexpensive, compact, and robust for the most demanding design applications. Moreover, a **cam profile** can be designed to generate virtually any desired follower motion, by either graphical or analytical methods.

## 20.5 Classification of Cams and Followers

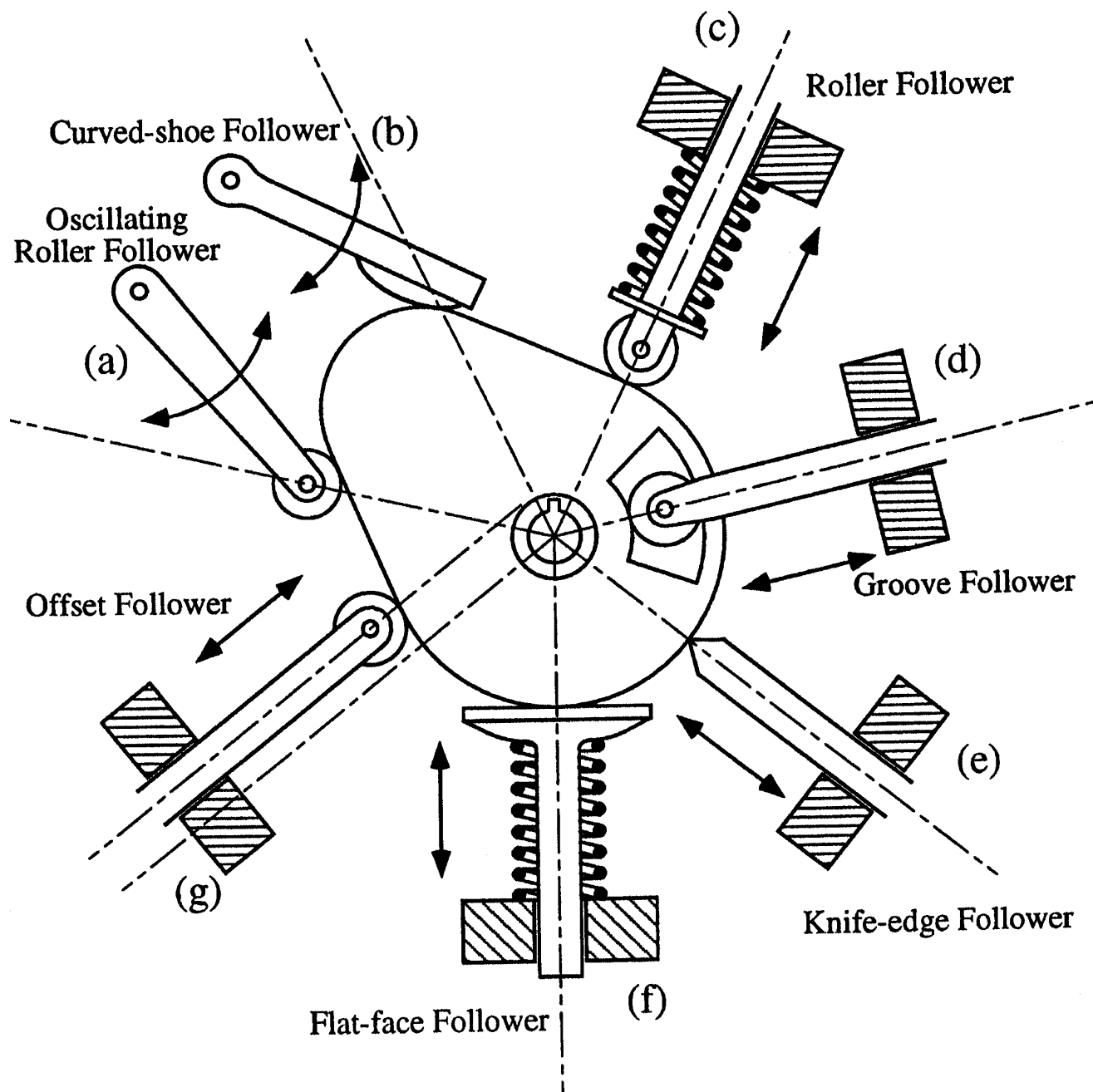
The versatility of cam pairs is evidenced by the variety of shapes, forms, and motions for both cam and follower. Cams are usually classified according to their basic shape as illustrated in Fig. 20.6: (a) plate cam, (b) wedge cam, (c) cylindric or barrel cam, and (d) end or face cam.

**Figure 20.6** Basic types of cams.



Followers are also classified according to their basic shape with optional modifiers describing their motion characteristics. For example, a follower can oscillate [Figs. 20.7(a–b)] or translate [20.7(c–g)]. As required by many applications, follower motion may be offset from the cam shaft's center as illustrated in Fig. 20.7(g). For all cam pairs, however, the follower must maintain constant contact with cam surface. Constant contact can be achieved by gravity, springs, or other mechanical constraints such as grooves.

**Figure 20.7** Basic types of followers.



## 20.6 Displacement Diagrams

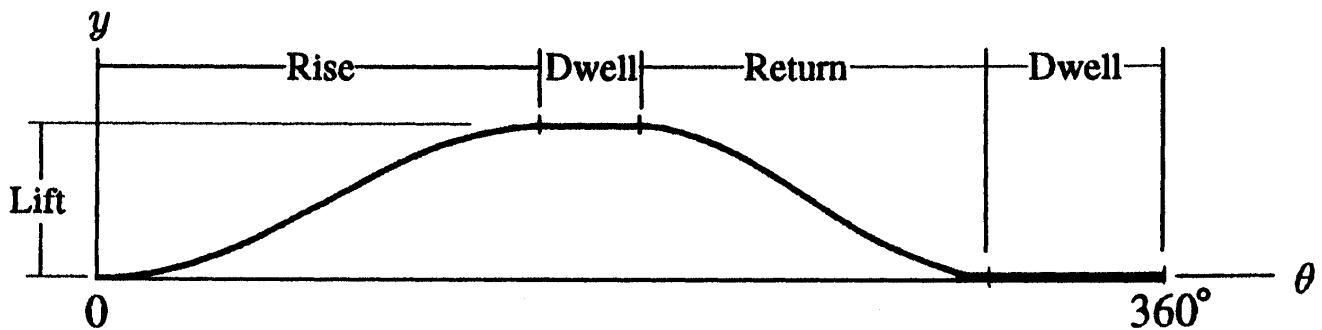
The cam's primary function is to create a well-defined follower displacement. If the cam's displacement is designated by  $\theta$  and follower displacement by  $y$ , a given cam is designed such that a displacement function

$$y = f(\theta) \quad (20.3)$$



is satisfied. A graph of  $y$  versus  $\theta$  is called the *follower displacement diagram* (Fig. 20.8). On a displacement diagram, the abscissa represents one revolution of cam motion ( $\theta$ ) and the ordinate represents the corresponding follower displacement ( $y$ ). Portions of the displacement diagram, when follower motion is away from the cam's center, are called *rise*. The maximum rise is called *lift*. Periods of follower rest are referred to as *dwells*, and *returns* occur when follower motion is toward the cam's center.

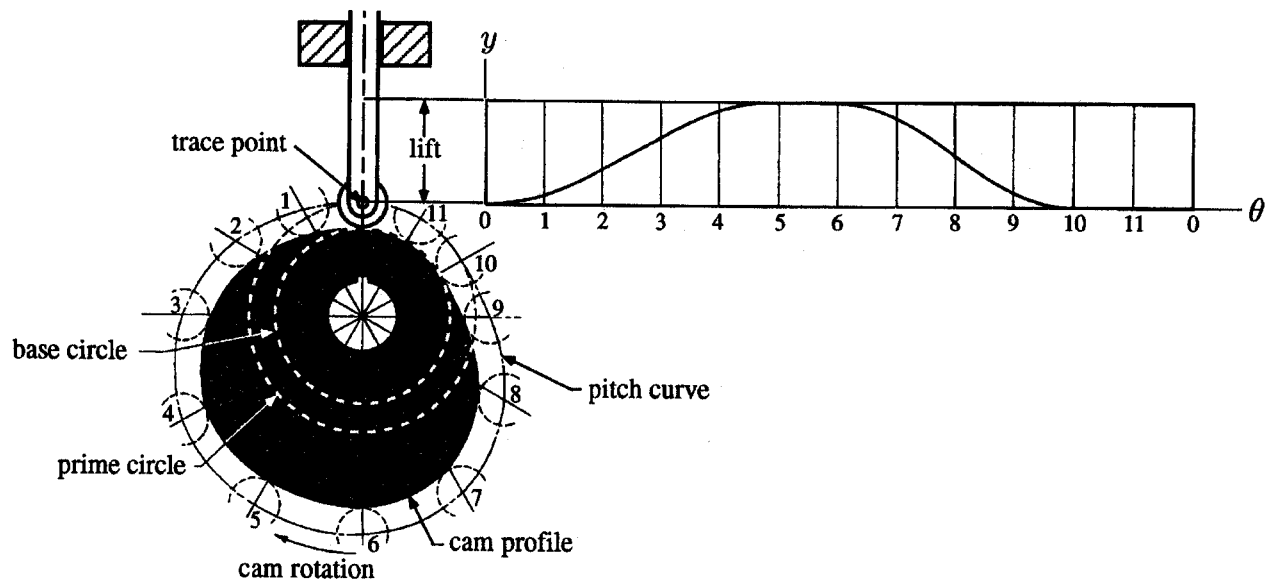
**Figure 20.8** Displacement diagram.



The cam profile is generated from the follower displacement diagram via graphical or analytical methods that use parabolic, simple harmonic, cycloidal, and/or polynomial profiles. For many applications, the follower's velocity, acceleration, and higher time derivatives are necessary for proper cam design.

Cam profile generation is best illustrated using graphical methods where the cam profile can be constructed from the follower displacement diagram using the principle of kinematic inversion. As shown in Fig. 20.9, the prime circle is divided into a number of equal angular segments and assigned station numbers. The follower displacement diagram is then divided along the abscissa into corresponding segments. Using dividers, the distances are then transferred from the displacement diagram directly onto the cam layout to locate the corresponding trace point position. A smooth curve through these points is the pitch curve. For the case of a roller follower, the roller is drawn in its proper position at each station and the cam profile is then constructed as a smooth curve tangent to all roller positions. Analytical methods can be employed to facilitate computer-aided design of cam profiles.

**Figure 20.9** Cam layout.



**Figure 20.9** Cam layout.

## Defining Terms

### Linkage Terminology

Standard terminology for linkages includes the following:

**Degree of freedom:** The number of parameters, available as input, that prescribe the configuration of a given linkage, also known as its *mobility*.

**Planar linkage:** A collection of links constrained to move in parallel planes.

**Revolute joint:** A hinged connection between two links that constrains their relative movement to the plane perpendicular to the hinge axis.

**Spatial linkage:** A linkage with at least one link that moves out of a plane.

**Spherical joint:** A connection between two links that constrains their relative movement to a sphere about a point at the center of the joint.

**Spherical linkage:** A collection of links constrained to move on concentric spheres.

### Cam Terminology

The standard cam terminology is illustrated in [Fig. 20.10](#) and defined as follows:

**Base circle:** The smallest circle, centered on the cam axis, that touches the cam profile (radius  $R_b$ ).

**Cam profile:** The cam's working surface.

**Pitch circle:** The circle through the pitch point, centered on the cam axis (radius  $R_p$ ).

**Pitch curve:** The path of the trace point.

**Pitch point:** The point on the pitch curve where pressure angle is maximum.

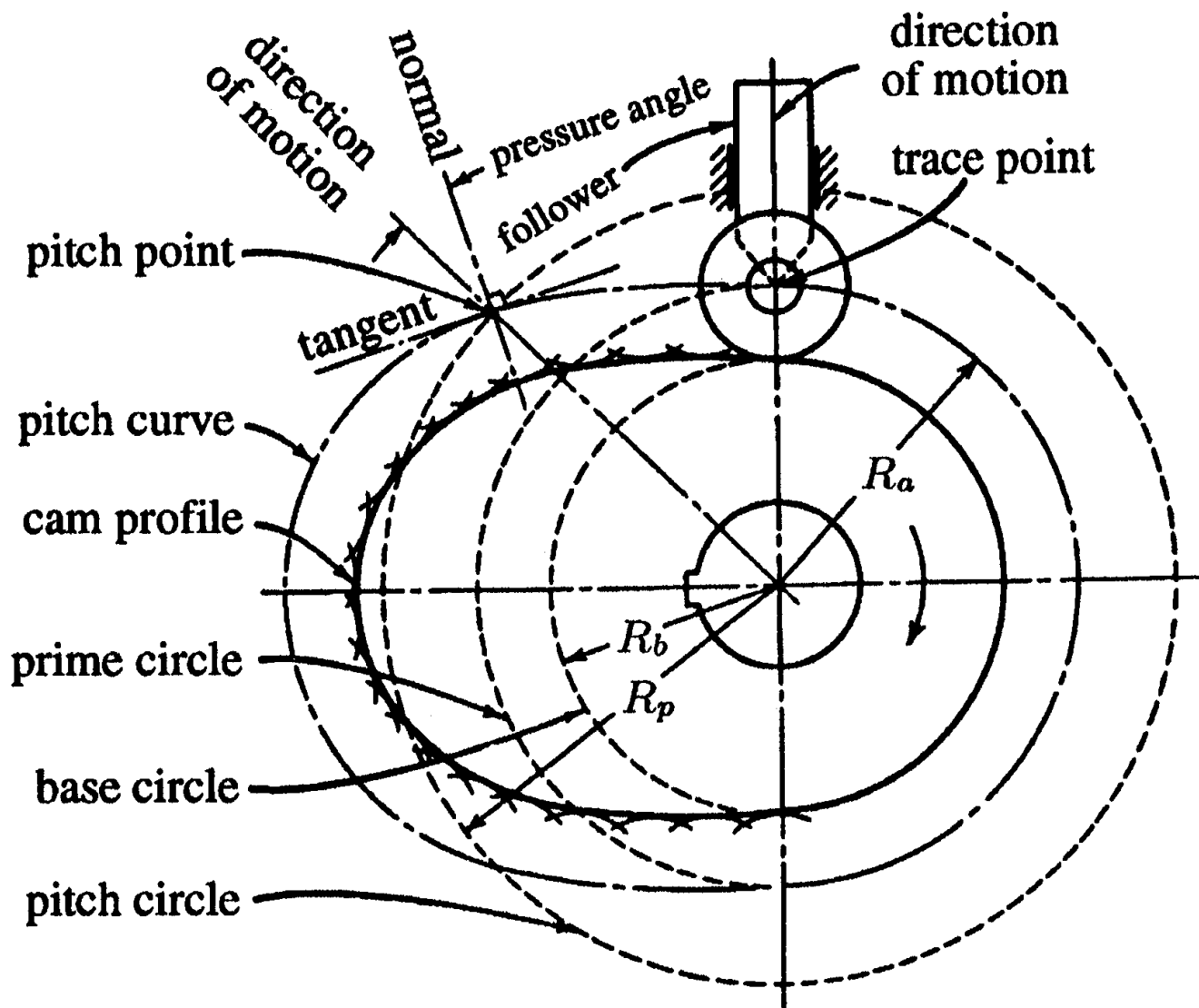
**Pressure angle:** The angle between the normal to the pitch curve and the instantaneous direction

of trace point motion.

**Prime circle:** The smallest circle, centered on the cam axis, that touches the pitch curve (radius  $R_a$ ).

**Trace point:** The contact point of a knife-edge follower, the center of a roller follower, or a reference point on a flat-faced follower.

**Figure 20.10** Cam terminology.



## References

- Chironis, N. P. 1965. *Mechanisms, Linkages, and Mechanical Controls*. McGraw-Hill, New York.
- Erdman, A. G. and Sandor, G. N. 1984. *Mechanism Design: Analysis and Synthesis*, vol. 1. Prentice Hall, Englewood Cliffs, NJ.

- Paul, B. 1979. *Kinematics and Dynamics of Planar Machinery*. Prentice Hall, Englewood Cliffs, NJ.
- Shigley, J. E. and Uicker, J. J. 1980. *Theory of Machines and Mechanisms*. McGraw-Hill, New York.
- Suh, C. H. and Radcliffe, C. W. 1978. *Kinematics and Mechanism Design*. John Wiley & Sons, New York.

## **Further Information**

An interesting array of linkages that generate specific movements can be found in *Mechanisms and Mechanical Devices Sourcebook* by Nicholas P. Chironis.

Design methodologies for planar and spatial linkages to guide a body in a desired way are found in *Mechanism Design: Analysis and Synthesis* by George Sandor and Arthur Erdman and in *Kinematics and Mechanism Design* by Chung Ha Suh and Charles W. Radcliffe.

*Theory of Machines and Mechanisms* by Joseph E. Shigley and John J. Uicker is particularly helpful in design of cam profiles for various applications.

Proceedings of the ASME Design Engineering Technical Conferences are published annually by the American Society of Mechanical Engineers. These proceedings document the latest developments in mechanism and machine theory.

The quarterly *ASME Journal of Mechanical Design* reports on advances in the design and analysis of linkage and cam systems. For a subscription contact American Society of Mechanical Engineers, 345 E. 47th St., New York, NY 10017.

Bhushan, B. "Tribology: Friction, Wear, and Lubrication"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Tribology: Friction, Wear, and Lubrication

---

## 21.1 History of Tribology and Its Significance to Industry

## 21.2 Origins and Significance of Micro/nanotribology

## 21.3 Friction

Definition of Friction • Theories of Friction • Measurements of Friction

## 21.4 Wear

Adhesive Wear • Abrasive Wear • Fatigue Wear • Impact Wear • Corrosive Wear • Electrical Arc–Induced Wear • Fretting and Fretting Corrosion

## 21.5 Lubrication

Solid Lubrication • Fluid Film Lubrication

## 21.6 Micro/nanotribology

### **Bharat Bhushan**

*Ohio State University*

In this chapter we first present the history of macrotribology and micro/nanotribology and their significance. We then describe mechanisms of friction, wear, and lubrication, followed by micro/nanotribology.

## 21.1 History of Tribology and Its Significance to Industry

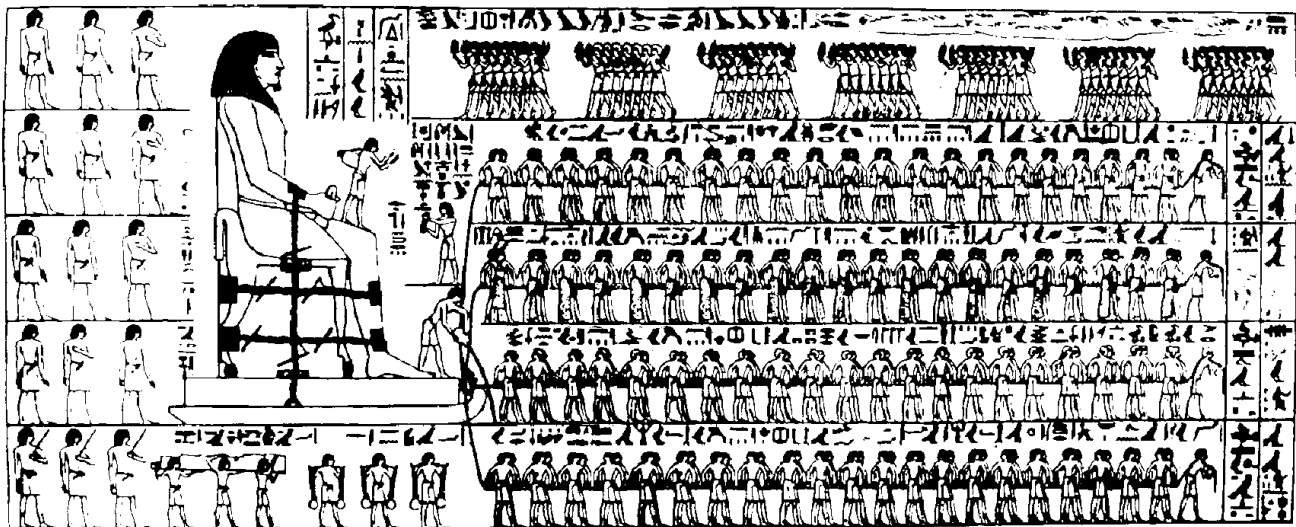
---

**Tribology** is the science and technology of two interacting surfaces in relative motion and of related subjects and practices. The popular equivalent is friction, wear, and lubrication. The word *tribology*, coined in 1966, is derived from the Greek word *tribos* meaning "rubbing," so the literal translation would be the science of rubbing [Jost, 1966]. It is only the name tribology that is relatively new, because interest in the constituent parts of tribology is older than recorded history [Dowson, 1979]. It is known that drills made during the Paleolithic period for drilling holes or producing fire were fitted with bearings made from antlers or bones, and potters' wheels or stones for grinding cereals clearly had a requirement for some form of bearings [Davidson, 1957]. A ball thrust bearing dated about 40 A.D. was found in Lake Nemi near Rome.

Records show the use of wheels from 3500 B.C., which illustrates our ancestors' concern with reducing friction in translationary motion. The transportation of large stone building blocks and monuments required the know-how of frictional devices and lubricants, such as water-lubricated

sleds. Figure 21.1 illustrates the use of a sledge to transport a heavy statue by Egyptians circa 1880 B.C. [Layard, 1853]. In this transportation, 172 slaves are being used to drag a large statue weighing about 600 kN along a wooden track. One man, standing on the sledge supporting the statue, is seen pouring a liquid into the path of motion; perhaps he was one of the earliest lubrication engineers. [Dowson (1979) has estimated that each man exerted a pull of about 800 N. On this basis the total effort, which must at least equal the friction force, becomes  $172 \times 800$  N. Thus, the coefficient of friction is about 0.23.] A tomb in Egypt that was dated several thousand years B.C. provides the evidence of use of lubricants. A chariot in this tomb still contained some of the original animal-fat lubricant in its wheel bearings.

**Figure 21.1** Egyptians using lubricant to aid movement of Colossus, El-Bersheh, c. 1880 B.C.



During and after the glory of the Roman empire, military engineers rose to prominence by devising both war machinery and methods of fortification, using tribological principles. It was the Renaissance engineer and artist Leonardo da Vinci (1452–1519), celebrated in his days for his genius in military construction as well as for his painting and sculpture, who first postulated a scientific approach to friction. Leonardo introduced for the first time the concept of coefficient of friction as the ratio of the friction force to normal load. In 1699 Amontons found that the friction force is directly proportional to the normal load and is independent of the apparent area of contact. These observations were verified by Coulomb in 1781, who made a clear distinction between static friction and kinetic friction.

Many other developments occurred during the 1500s, particularly in the use of improved bearing materials. In 1684 Robert Hooke suggested the combination of steel shafts and bell-metal bushes as preferable to wood shod with iron for wheel bearings. Further developments were associated with the growth of industrialization in the latter part of the eighteenth century. Early developments in the petroleum industry started in Scotland, Canada, and the U.S. in the 1850s [Parish, 1935; Dowson, 1979].

Though essential laws of viscous flow had earlier been postulated by Newton, scientific

understanding of lubricated bearing operations did not occur until the end of the nineteenth century. Indeed, the beginning of our understanding of the principle of hydrodynamic lubrication was made possible by the experimental studies of Tower [1884] and the theoretical interpretations of Reynolds [1886] and related work by Petroff [1883]. Since then developments in hydrodynamic bearing theory and practice have been extremely rapid in meeting the demand for reliable bearings in new machinery.

Wear is a much younger subject than friction and bearing development, and it was initiated on a largely empirical basis.

Since the beginning of the 20th century, from enormous industrial growth leading to demand for better tribology, our knowledge in all areas of tribology has expanded tremendously [Holm, 1946; Bowden and Tabor, 1950, 1964; Bhushan, 1990, 1992; Bhushan and Gupta, 1991].

Tribology is crucial to modern machinery, which uses sliding and rolling surfaces. Examples of productive wear are writing with a pencil, machining, and polishing. Examples of productive friction are brakes, clutches, driving wheels on trains and automobiles, bolts, and nuts. Examples of unproductive friction and wear are internal combustion and aircraft engines, gears, cams, bearings, and seals. According to some estimates, losses resulting from ignorance of tribology amount in the U.S. to about 6% of its gross national product or about 200 billion dollars per year, and approximately one-third of the world's energy resources in present use appear as friction in one form or another. Thus, the importance of friction reduction and wear control cannot be overemphasized for economic reasons and long-term reliability. According to Jost [1966, 1976], the United Kingdom could save approximately 500 million pounds per annum and the U.S. could save in excess of 16 billion dollars per annum by better tribological practices. The savings are both substantial and significant and could be obtained without the deployment of large capital investment.

The purpose of research in tribology is understandably the minimization and elimination of losses resulting from friction and wear at all levels of technology where the rubbing of surfaces are involved. Research in tribology leads to greater plant efficiency, better performance, fewer breakdowns, and significant savings.

## 21.2 Origins and Significance of Micro/nanotribology

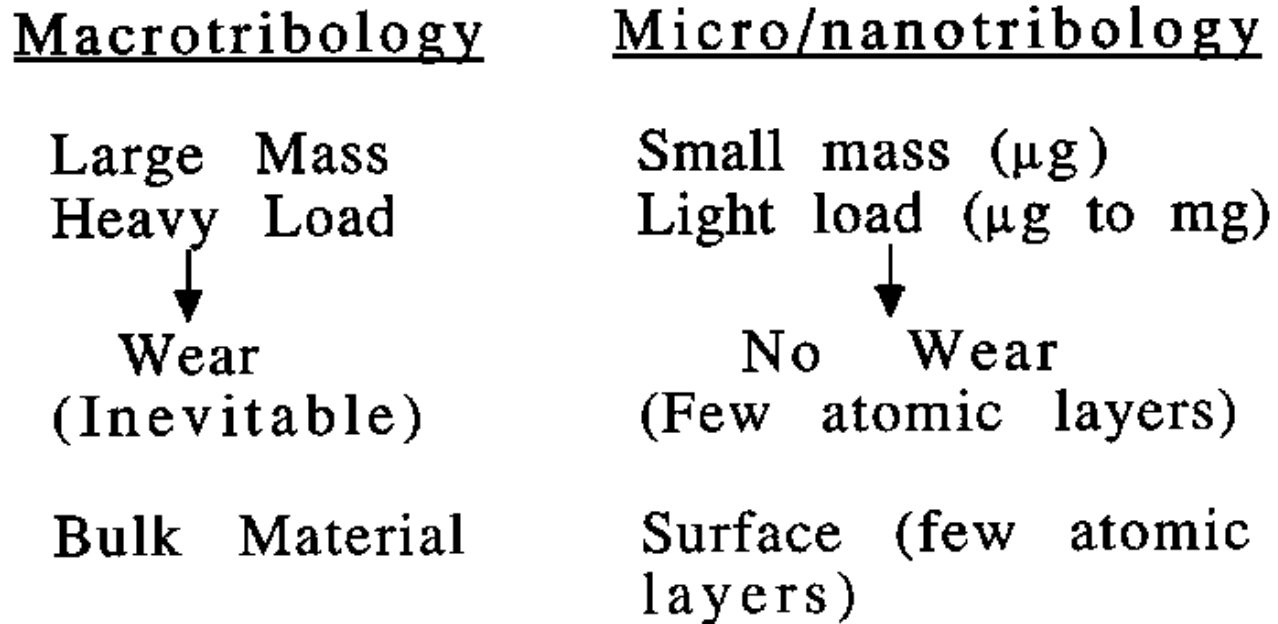
---

The advent of new techniques to measure surface topography, adhesion, friction, wear, lubricant film thickness, and mechanical properties all on micro- to nanometer scale; to image lubricant molecules; and to conduct atomic-scale simulations with the availability of supercomputers has led to development of a new field referred to as *microtribology*, *nanotribology*, *molecular tribology*, or *atomic-scale tribology*. This field deals with experimental and theoretical investigations of processes ranging from atomic and molecular scales to micro scales, occurring during adhesion, friction, wear, and thin-film lubrication at sliding surfaces. The differences between the conventional or macrotribology and micro/nanotribology are contrasted in Fig. 21.2. In macrotribology, tests are conducted on components with relatively large mass under heavily loaded conditions. In these tests, wear is inevitable and the bulk properties of mating components dominate the tribological performance. In **micro/nanotribology**, measurements are made on components, at least one of the mating components with relatively small mass under lightly loaded



conditions. In this situation negligible wear occurs and the surface properties dominate the tribological performance.

**Figure 21.2** Comparison between macrotribology and micro/nanotribology.



The micro/nanotribological studies are needed to develop fundamental understanding of interfacial phenomena on a small scale and to study interfacial phenomena in micro- and nanostructures used in magnetic storage systems, microelectromechanical systems (MEMS) and other industrial applications [Bhushan, 1990, 1992]. The components used in micro- and nanostructures are very light (on the order of few micrograms) and operate under very light loads (on the order of few micrograms to few milligrams). As a result, friction and wear (on a nanoscale) of lightly loaded micro/nanocomponents are highly dependent on the surface interactions (few atomic layers). These structures are generally lubricated with molecularly thin films. Micro- and nanotribological techniques are ideal to study the friction and wear processes of micro- and nanostructures. Although micro/nanotribological studies are critical to study micro- and nanostructures, these studies are also valuable in fundamental understanding of interfacial phenomena in macrostructures to provide a bridge between science and engineering. Friction and wear on micro- and nanoscales have been found to be generally small compared to that at macroscales. Therefore, micro/nanotribological studies may identify the regime for ultra-low friction and near zero wear.

To give a historical perspective of the field [Bhushan, 1995], the *scanning tunneling microscope* (STM) developed by Dr. Gerd Binnig and his colleagues in 1981 at the IBM Zurich Research Laboratory, Forschungslabor, is the first instrument capable of directly obtaining three-dimensional (3-D) images of solid surfaces with atomic resolution [Binnig *et al.*, 1982]. G. Binnig and H. Rohrer received a Nobel Prize in Physics in 1986 for their discovery. STMs can

only be used to study surfaces that are electrically conductive to some degree. Based on their design of STM Binnig *et al.* developed, in 1985, an *atomic force microscope* (AFM) to measure ultrasmall forces (less than  $1\ \mu\text{N}$ ) present between the AFM tip surface and the sample surface [1986]. AFMs can be used for measurement of *all engineering surfaces*, which may be either electrically conducting or insulating. AFM has become a popular surface profiler for topographic measurements on micro- to nanoscale. Mate *et al.* [1987] were the first to modify an AFM in order to measure both normal and friction forces and this instrument is generally called *friction force microscope* (FFM) or *lateral force microscope* (LFM). Since then, Bhushan and other researchers have used FFM for atomic-scale and microscale friction and boundary lubrication studies [Bhushan and Ruan, 1994; Bhushan *et al.*, 1994; Ruan and Bhushan, 1994; Bhushan, 1995; Bhushan *et al.*, 1995]. By using a standard or a sharp diamond tip mounted on a stiff cantilever beam, Bhushan and other researchers have used AFM for scratching, wear, and measurements of elastic/plastic mechanical properties (such as indentation hardness and modulus of elasticity) [Bhushan *et al.*, 1994; Bhushan and Koinkar, 1994a,b; Bhushan, 1995; Bhushan *et al.*, 1995].

Surface force apparatuses (SFAs), first developed in 1969 [Tabor and Winterton, 1969], are other instruments used to study both static and dynamic properties of the molecularly thin liquid films sandwiched between two molecularly smooth surfaces [Israelachvili and Adams, 1978; Klein, 1980; Tonck *et al.*, 1988; Georges *et al.*, 1993,1994]. These instruments have been used to measure the dynamic shear response of liquid films [Bhushan, 1995]. Recently, new friction attachments were developed that allow for two surfaces to be sheared past each other at varying sliding speeds or oscillating frequencies while simultaneously measuring both the friction forces and normal forces between them [Peachey *et al.*, 1991; Bhushan, 1995]. The distance between two surfaces can also be independently controlled to within  $\pm 0.1\ \text{nm}$  and the force sensitivity is about  $10\ \text{nN}$ . The SFAs are used to study rheology of molecularly thin liquid films; however, the liquid under study has to be confined between molecularly smooth optically transparent surfaces with radii of curvature on the order of  $1\ \text{mm}$  (leading to poorer lateral resolution as compared to AFMs). SFAs developed by Tonck *et al.* [1988] and Georges *et al.* [1993, 1994] use an opaque and smooth ball with large radius ( $\approx 3\ \text{mm}$ ) against an opaque and smooth flat surface. Only AFMs/FFMs can be used to study *engineering surfaces* in the *dry and wet conditions* with *atomic resolution*.

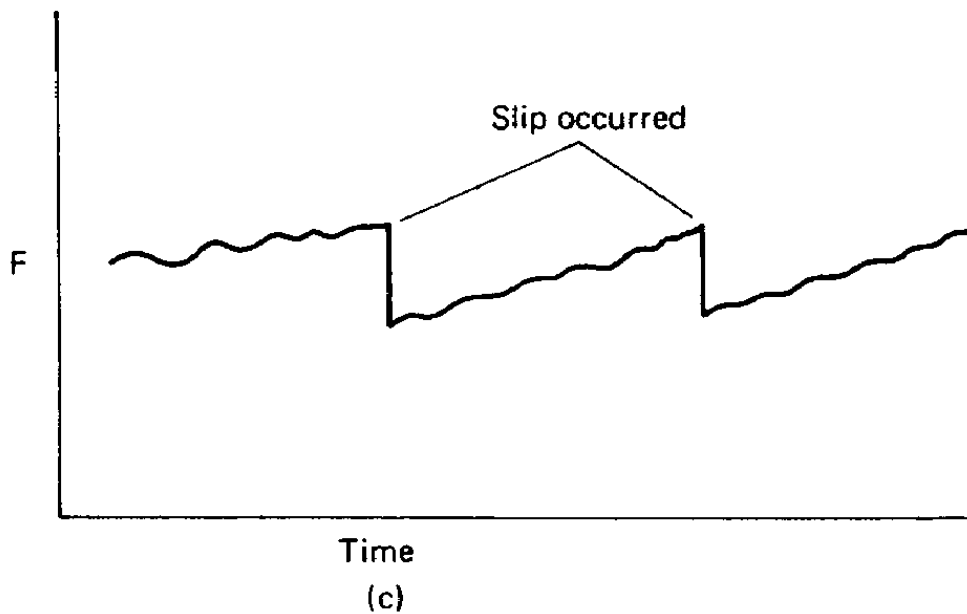
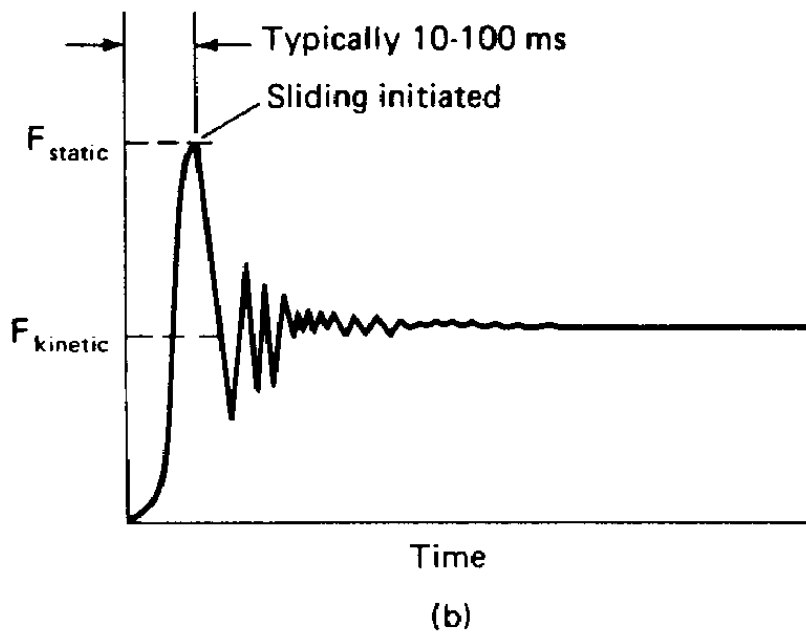
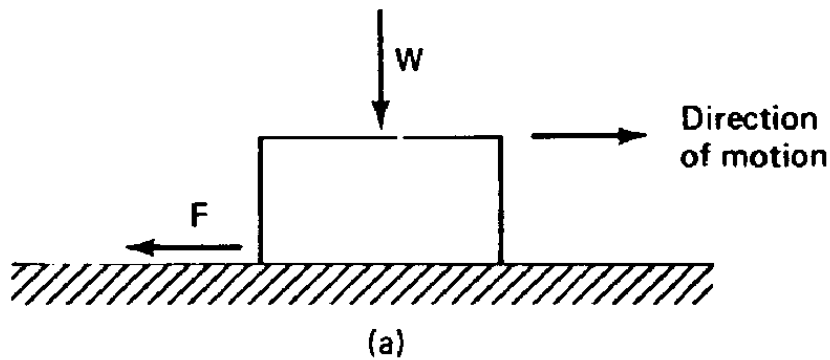
## 21.3 Friction

---

### Definition of Friction

**Friction** is the resistance to motion that is experienced whenever one solid body slides over another. The resistive force, which is parallel to the direction of motion, is called the friction force, Fig. 21.3(a). If the solid bodies are loaded together and a tangential force ( $F$ ) is applied, then the value of the tangential force that is required to initiate sliding is the static friction force. It may take a few milliseconds before sliding is initiated at the interface ( $F_{\text{static}}$ ). The tangential force required to maintain sliding is the kinetic (or dynamic) friction force ( $F_{\text{kinetic}}$ ). The kinetic friction force is either lower than or equal to the static friction force, Fig. 21.3(b).

**Figure 21.3** (a) Schematic illustration of a body sliding on a horizontal surface.  $W$  is the normal load and  $F$  is the friction force. (b) Friction force versus time or displacement.  $F_{\text{static}}$  is the force required to initiate sliding and  $F_{\text{kinetic}}$  is the force required to sustain sliding. (c) Kinetic friction force versus time or displacement showing irregular stick-slip.



It has been found experimentally that there are two basic laws of intrinsic (or conventional) friction that are generally obeyed over a wide range of applications. The first law states that the friction is independent of the apparent area of contact between the contacting bodies, and the second law states that the friction force  $F$  is proportional to the normal load  $W$  between the bodies. These laws are often referred to as *Amontons laws*, after the French engineer Amontons, who presented them in 1699 [Dowson, 1979].

The second law of friction enables us to define a *coefficient of friction*. The law states that the friction force  $F$  is proportional to the normal load  $W$ . That is,

$$F = \mu W \quad (21.1)$$

where  $\mu$  is a constant known as the *coefficient of friction*. It should be emphasized that  $\mu$  is a constant only for a given pair of sliding materials under a given set of operating conditions (temperature, humidity, normal pressure, and sliding velocity). Many materials show sliding speed and normal load dependence on the coefficients of static and kinetic friction in dry and lubricated contact.

It is a matter of common experience that the sliding of one body over another under a steady pulling force proceeds sometimes at constant or nearly constant velocity, and on other occasions at velocities that fluctuate widely. If the friction force (or sliding velocity) does not remain constant as a function of distance or time and produces a form of oscillation, it is generally called a *stick-slip phenomena*, Fig. 21.3(c). During the stick phase, the friction force builds up to a certain value and then slip occurs at the interface. Usually, a sawtooth pattern in the friction force–time curve [Fig. 21.3(c)] is observed during the stick-slip process. Stick-slip generally arises whenever the coefficient of static friction is markedly greater than the coefficient of kinetic friction or whenever the rate of change of coefficient of kinetic friction as a function of velocity at the sliding velocity employed is negative. The stick-slip events can occur either repetitively or in a random manner.

The stick-slip process generally results in squealing and chattering of sliding systems. In most sliding systems the fluctuations of sliding velocity resulting from the stick-slip process and associated squeal and chatter are considered undesirable, and measures are normally taken to eliminate, or at any rate to reduce, the amplitude of the fluctuations.

## Theories of Friction

All engineering surfaces are rough on a microscale. When two nominally flat surfaces are placed in contact under load, the contact takes place at the tips of the asperities and the load is supported by the deformation of contacting asperities, and the discrete contact spots (junctions) are formed, Fig. 21.4. The sum of the areas of all the contact spots constitutes the real (true) area of the contact ( $A_r$ ) and for most materials at normal loads, this will be only a small fraction of the apparent (nominal) area of contact ( $A_a$ ). The proximity of the asperities results in adhesive contacts caused by either physical or chemical interaction. When these two surfaces move relative to each other, a lateral force is required to overcome adhesion. This force is referred to as *adhesional friction*

force. From classical theory of adhesion, this friction force ( $F_A$ ) is defined as follows [Bowden and Tabor, 1950]. For a dry contact,

$$F_A = A_r \tau_a \quad (21.2a)$$

and for a lubricated contact,

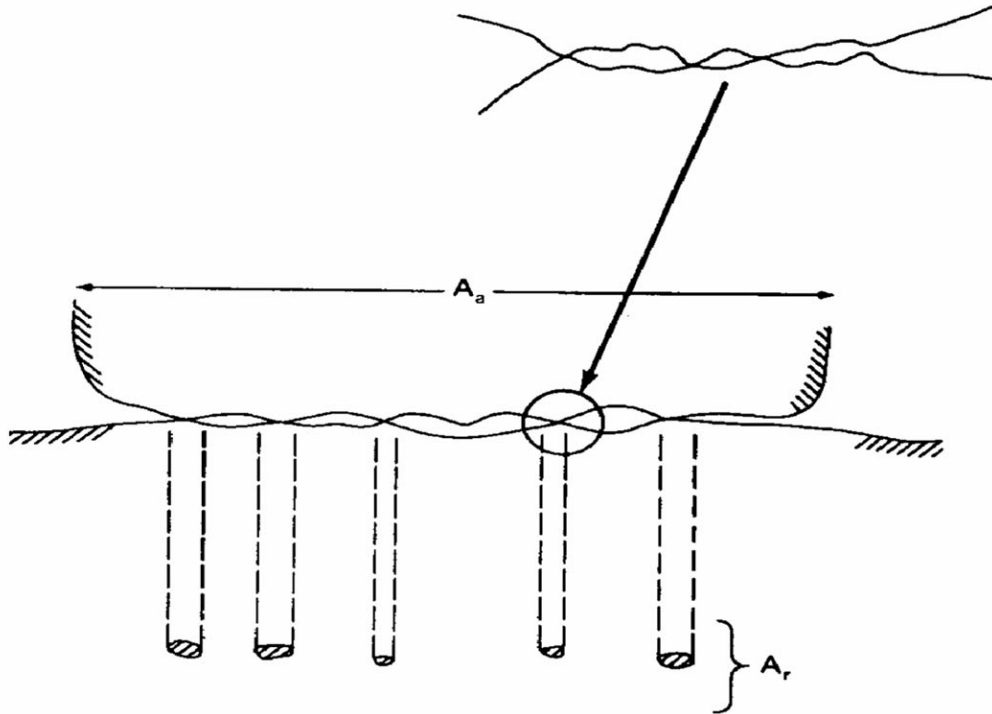
$$F_A = A_r [\alpha \tau_a + (1 - \alpha) \tau_l] \quad (21.2b)$$

and

$$\tau_l = \eta_l V / h \quad (21.2c)$$

where  $\tau_a$  and  $\tau_l$  are the shear strengths of the dry contact and of the lubricant film, respectively;  $\alpha$  is the fraction of unlubricated area;  $\eta_l$  is the dynamic viscosity of the lubricant;  $V$  is the relative sliding velocity; and  $h$  is the lubricant film thickness.

**Figure 21.4** Schematic representation of an interface, showing the apparent ( $A_a$ ) and real ( $A_r$ ) areas of contact. Typical size of an asperity contact is from submicron to a few microns. Inset shows the details of a contact on a submicron scale.



The contacts can be either elastic or plastic, depending primarily on the surface topography and the mechanical properties of the mating surfaces. The expressions for real area of contact for elastic ( $e$ ) and plastic ( $p$ ) contacts are as follows [Greenwood and Williamson, 1966; Bhushan, 1984, 1990]. For  $\psi < 0.6$ , elastic contacts,

$$A_{re}/W \sim 3.2/E_c (\sigma_p/R_p)^{1/2} \quad (21.3a)$$

For  $\psi > 1$ , plastic contacts,

$$A_{rp}/W = 1/H \quad (21.3b)$$

Finally,

$$\psi = (E_c/H) (\sigma_p/R_p)^{1/2} \quad (21.3c)$$

where  $E_c$  is the composite modulus of elasticity,  $H$  is the hardness of the softer material, and  $\sigma_p$  and  $1/R_p$  are the composite standard deviation and composite mean curvature of the summits of the mating surfaces. The real area of contact is reduced by improving the mechanical properties and in some cases by increasing the roughness (in the case of bulk of the deformation being in the elastic contact regime).

The adhesion strength depends upon the mechanical properties and the physical and chemical interaction of the contacting bodies. The adhesion strength is reduced by reducing surface interactions at the interface. For example, presence of contaminants or deliberately applied fluid film (e.g., air, water, or lubricant) would reduce the adhesion strength. Generally, most interfaces in vacuum with intimate solid-solid contact would exhibit very high values for coefficient of friction. Few pp of contaminants (air, water) may be sufficient to reduce  $\mu$  dramatically. Thick films of liquids or gases would further reduce  $\mu$ , as it is much easier to shear into a fluid film than to shear a solid-solid contact.

So far we have discussed theory of adhesional friction. If one of the sliding surfaces is harder than the other, the asperities of the harder surface may penetrate and plough into the softer surface. Ploughing into the softer surface may also occur as a result of impacted wear debris. In addition, interaction of two rather rough surfaces may result into mechanical interlocking on micro or macro scale. During sliding, interlocking would result into ploughing of one of the surfaces. In tangential motion the ploughing resistance is in addition to the adhesional friction. There is yet other mechanism of friction—deformation (or hysteresis) friction—which may be prevalent in materials with elastic hysteresis losses such as in polymers. In boundary lubricated conditions or unlubricated interfaces exposed to humid environments, presence of some liquid may result in formation of menisci or adhesive bridges and the meniscus/viscous effects may become important; in some cases these may even dominate the overall friction force [Bhushan, 1990].

## Measurements of Friction

In a friction measurement apparatus two test specimens are loaded against each other at a desired normal load, one of the specimens is allowed to slide relative to the other at a desired sliding speed, and the tangential force required to initiate or maintain sliding is measured. There are numerous apparatuses used to measure friction force [Benzing *et al.*, 1976; Bhushan and Gupta, 1991]. The simplest method is an inclined-plane technique. In this method the flat test specimen of weight  $W$  is placed on top of another flat specimen whose inclination can be adjusted, as shown in Fig. 21.5. The inclination of the lower specimen is increased from zero to an angle at which the block begins to slide. At this point, downward horizontal force being applied at the interface exceeds the static friction force,  $F_{\text{static}}$ . At the inclination angle  $\theta$ , at which the block just begins to slide,

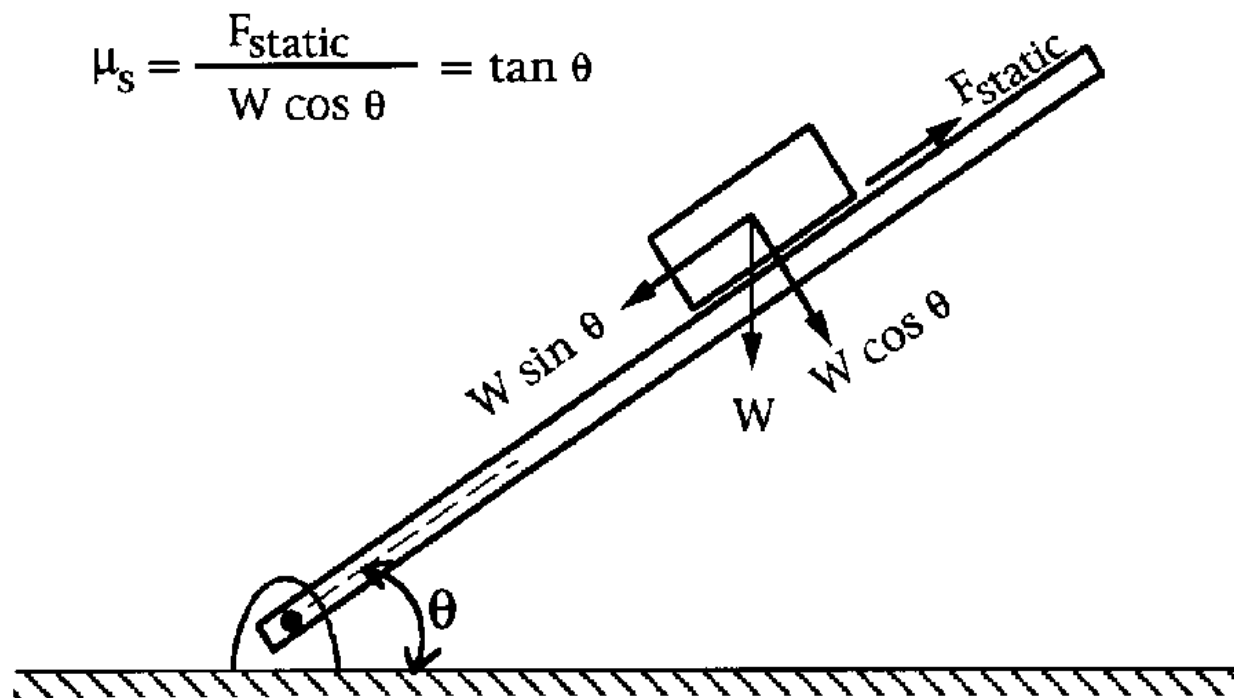
$$F_{\text{static}} = W \sin \theta$$

and the coefficient of static friction  $\mu_s$  is

$$\mu_s = \frac{F_{\text{static}}}{W \cos \theta} = \tan \theta \quad (21.4)$$

The angle  $\theta$  is referred to as *friction angle*. This simple method only measures the coefficient of static friction and does not allow the measurements of the effect of sliding. However, this method demonstrates the effects of friction and provides the simplest method to measure coefficient of static friction.

**Figure 21.5** Inclined-plane technique to measure static friction force.



Typical values of coefficient of friction of various material pairs are presented in [Table 21.1](#) [Avallo and Baumeister, 1987]. It should be noted that values of coefficient of friction depend on the operating conditions—loads, speeds, and the environment—and the values reported in [Table 21.1](#) should therefore be used with caution.

**Table 21.1** Coefficient of Friction  $\mu$  for Various Material Combinations

Materials	$\mu$ , static		$\mu$ , sliding (kinetic)	
	Dry	Greasy	Dry	Greasy
Hard steel on hard steel	0.78	0.11(a)	0.42	0.029(h)
		0.23(b)		0.081(c)
		0.15(c)		0.080(i)
		0.11(d)		0.058(j)

		0.0075( <i>p</i> )		0.084( <i>d</i> )
		0.0052( <i>h</i> )		0.105( <i>k</i> )
				0.096( <i>l</i> )
				0.108( <i>m</i> )
				0.12( <i>a</i> )
Mild steel on mild steel	0.74		0.57	0.09( <i>a</i> )
				0.19( <i>u</i> )
Hard steel on graphite	0.21	0.09( <i>a</i> )		
Hard steel on babbitt (ASTM 1)	0.70	0.23( <i>b</i> )	0.33	0.16( <i>b</i> )
		0.15( <i>c</i> )		0.06( <i>c</i> )
		0.08( <i>d</i> )		0.11( <i>d</i> )
		0.085( <i>e</i> )		
Hard steel on babbitt (ASTM 8)	0.42	0.17( <i>b</i> )	0.35	0.14( <i>b</i> )
		0.11( <i>c</i> )		0.065( <i>c</i> )
		0.09( <i>d</i> )		0.07( <i>d</i> )
		0.08( <i>e</i> )		0.08( <i>h</i> )
Hard steel on babbitt (ASTM 10)		0.25( <i>b</i> )		0.13( <i>b</i> )
		0.12( <i>c</i> )		0.06( <i>c</i> )
		0.10( <i>d</i> )		0.055( <i>d</i> )
		0.11( <i>e</i> )		
Mild steel on cadmium silver				0.097( <i>f</i> )
Mild steel on phosphor bronze			0.34	0.173( <i>f</i> )
Mild steel on copper lead				0.145( <i>f</i> )
Mild steel on cast iron		0.183( <i>c</i> )	0.23	0.133( <i>f</i> )
Mild steel on lead	0.95	0.5( <i>f</i> )	0.95	0.3( <i>f</i> )
Nickel on mild steel			0.64	0.178( <i>x</i> )
Aluminum on mild steel	0.61		0.47	
Magnesium on mild steel			0.42	
Magnesium on magnesium	0.6	0.08( <i>y</i> )		
Teflon on Teflon	0.04			0.04( <i>f</i> )
Teflon on steel	0.04			0.04( <i>f</i> )
Tungsten carbide on tungsten carbide	0.2	0.12( <i>a</i> )		
Tungsten carbide on steel	0.5	0.08( <i>a</i> )		
Tungsten carbide on copper	0.35			
Tungsten carbide on iron	0.8			
Bonded carbide on copper	0.35			
Bonded carbide on iron	0.8			
Cadmium on mild steel			0.46	
Copper on mild steel	0.53		0.36	0.18( <i>a</i> )
Nickel on nickel	1.10		0.53	0.12( <i>w</i> )



Brass on mild steel	0.51		0.44	
Brass on cast iron			0.30	
Zinc on cast iron	0.85		0.21	
Magnesium on cast iron			0.25	
Copper on cast iron	1.05		0.29	
Tin on cast iron			0.32	
Lead on cast iron			0.43	
Aluminum on aluminum	1.05		1.4	
Glass on glass	0.94	0.01( <i>p</i> )	0.40	0.09( <i>a</i> )
		0.005( <i>q</i> )		0.116( <i>v</i> )
Carbon on glass			0.18	
Garnet on mild steel			0.39	
Glass on nickel	0.78		0.56	
Copper on glass	0.68		0.53	
Cast iron on cast iron	1.10		0.15	0.070( <i>d</i> )
				0.064( <i>n</i> )
Bronze on cast iron			0.22	0.077( <i>n</i> )
Oak on oak (parallel to grain)	0.62		0.48	0.164( <i>r</i> )
				0.067( <i>s</i> )
Oak on oak (perpendicular)	0.54		0.32	0.072( <i>s</i> )
Leather on oak (parallel)	0.61		0.52	
Cast iron on oak			0.49	0.075( <i>n</i> )
Leather on cast iron			0.56	0.36( <i>t</i> )
				0.13( <i>n</i> )
Laminated plastic on steel			0.35	0.05( <i>t</i> )
Fluted rubber bearing on steel				0.05( <i>t</i> )

---

*Source:* Adapted from Avallone, E. A. and Baumeister, T., III, 1987. *Marks' Standard Handbook for Mechanical Engineers*, 9th ed. McGraw-Hill, New York.

*Note:* Reference letters indicate the lubricant used:

- a* = oleic acid
- b* = Atlantic spindle oil (light mineral)
- c* = castor oil
- d* = lard oil
- e* = Atlantic spindle oil plus 2% oleic acid
- f* = medium mineral oil
- g* = medium mineral oil plus ½% oleic acid
- h* = stearic acid
- i* = grease (zinc oxide base)
- j* = graphite
- k* = turbine oil plus 1% graphite
- l* = turbine oil plus 1% stearic acid
- m* = turbine oil (medium mineral)
- n* = olive oil
- p* = palmitic acid

$q$  = ricinoleic acid  
 $r$  = dry soap  
 $s$  = lard  
 $t$  = water  
 $u$  = rape oil  
 $v$  = 3-in-1 oil  
 $w$  = octyl alcohol  
 $x$  = triolein  
 $y$  = 1% lauric acid in paraffin oil

## 21.4 Wear

---

**Wear** is the removal of material from one or both of two solid surfaces in a solid-state contact. It occurs when solid surfaces are in a sliding, rolling, or impact motion relative to one another. Wear occurs through surface interactions at asperities, and components may need replacement after a relatively small amount of material has been removed or if the surface is unduly roughened. In well-designed tribological systems, the removal of material is usually a very slow process but it is very steady and continuous. The generation and circulation of wear debris—particularly in machine applications where the clearances are small relative to the wear particle size—may be more of a problem than the actual amount of wear.

Wear includes six principal, quite distinct phenomena that have only one thing in common: the removal of solid material from rubbing surfaces. These are (1) adhesive; (2) abrasive; (3) fatigue; (4) impact by erosion or percussion; (5) corrosive; and (6) electrical arc-induced wear [Archard, 1980; Bhushan *et al.*, 1985a,b; Bhushan, 1990]. Other commonly encountered wear types are fretting and fretting corrosion. These are not distinct mechanisms, but rather combinations of the adhesive, corrosive, and abrasive forms of wear. According to some estimates, two-thirds of all wear encountered in industrial situations occurs because of adhesive- and abrasive-wear mechanisms.

Of the aforementioned wear mechanisms, one or more may be operating in one particular machinery. In many cases wear is initiated by one mechanism and results in other wear mechanisms, thereby complicating failure analysis.

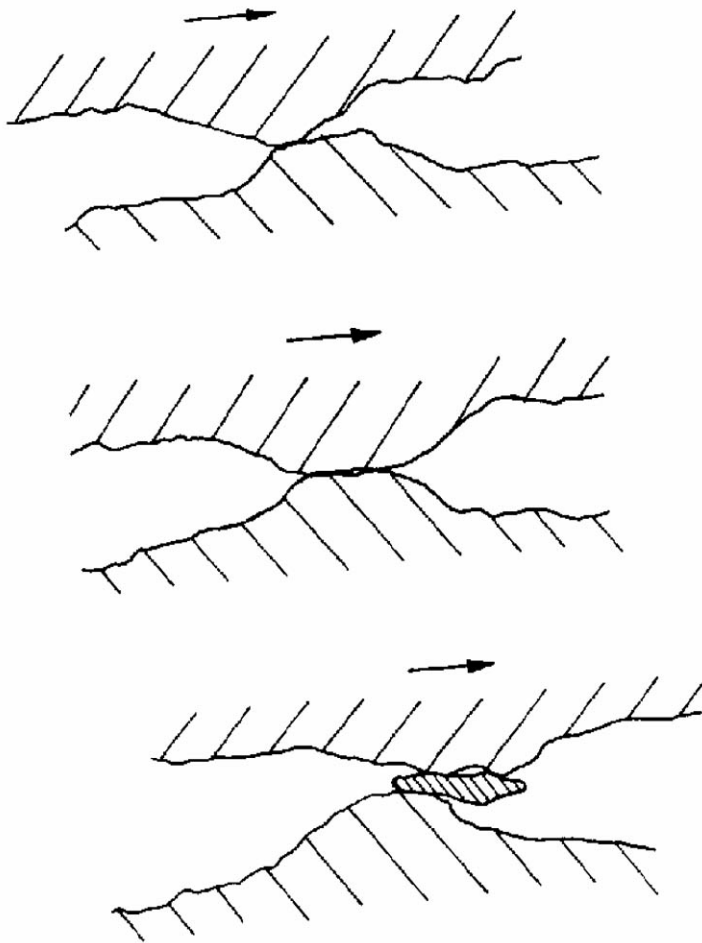
### Adhesive Wear

Adhesive wear occurs when two nominally flat solid bodies are in rubbing contact, whether lubricated or not. Adhesion (or bonding) occurs at the asperity contacts on the interface, and fragments are pulled off one surface to adhere to the other surface. Subsequently, these fragments may come off the surface on which they are formed and either be transferred back to the original surface or form loose wear particles. Severe types of adhesive wear are often called *galling*, *scuffing*, *scoring*, or *smearing*, although these terms are sometimes used loosely to describe other types of wear.

Although the adhesive-wear theory can explain transferred wear particles, it does not explain how loose wear particles are formed. We now describe the actual process of formation of wear

particles. Asperity contacts are sheared by sliding and a small fragment of *either surface* becomes attached to the other surface. As sliding continues, the fragment constitutes a new asperity that becomes attached once more to the original surface. This transfer element is repeatedly passed from one surface to the other and grows quickly to a large size, absorbing many of the transfer elements so as to form a flakelike particle from materials of both rubbing elements. Rapid growth of this transfer particle finally accounts for its removal as a wear particle, as shown in [Fig. 21.6](#). The occurrence of wear of the harder of the two rubbing surfaces is difficult to understand in terms of the adhesion theory. It is believed that the material transferred by adhesion to the harder surface may finally get detached by a fatigue process.

**Figure 21.6** Schematic showing generation of wear particle as a result of adhesive wear mechanism.



As a result of experiments carried out with various unlubricated materials—the vast majority being metallic—it is possible to write the laws of adhesive wear, commonly referred to as Archard's law, as follows [[Archard, 1953](#)]. For plastic contacts,

$$V = kWx/H \quad (21.5)$$

where  $V$  is the volume worn away,  $W$  is the normal load,  $x$  is the sliding distance,  $H$  is the hardness of the surface being worn away, and  $k$  is a nondimensional wear coefficient dependent on the materials in contact and their exact degree of cleanliness. The term  $k$  is usually interpreted as the probability that a wear particle is formed at a given asperity encounter.

Equation (21.5) suggests that the probability of a wear-particle formation increases with an increase in the real area of contact,  $A_r$  ( $A_r = W/H$  for plastic contacts), and the sliding distance. For elastic contacts occurring in materials with a low modulus of elasticity and a very low surface roughness Eq. (21.5) can be rewritten for elastic contacts (Bhushan's law of adhesive wear) as [Bhushan, 1990]

$$V = k'Wx/E_c(\sigma_p/R_p)^{1/2} \quad (21.6)$$

where  $k'$  is a nondimensional wear coefficient. According to this equation, elastic modulus and surface roughness govern the volume of wear. We note that in an elastic contact—though the normal stresses remain compressive throughout the entire contact—strong adhesion of some contacts can lead to generation of wear particles. Repeated elastic contacts can also fail by surface/subsurface fatigue. In addition, as the total number of contacts increases, the probability of a few plastic contacts increases, and the plastic contacts are specially detrimental from the wear standpoint.

Based on studies by Rabinowicz [1980], typical values of wear coefficients for metal on metal and nonmetal on metal combinations that are unlubricated (clean) and in various lubricated conditions are presented in Table 21.2. Wear coefficients and coefficients of friction for selected material combinations are presented in Table 21.3 [Archard, 1980].

**Table 21.2** Typical Values of Wear Coefficients for Metal on Metal and Nonmetal on Metal Combinations

Condition	Metal on Metal		Nonmetal on Metal
	Like	Unlike*	
Clean (unlubricated)	$1500 \cdot 10^{-6}$	15 to $500 \cdot 10^{-6}$	$1.5 \cdot 10^{-6}$
Poorly lubricated	300	3 to 100	1.5
Average lubrication	30	0.3 to 10	0.3
Excellent lubrication	1	0.03 to 0.3	0.03

\*The values depend on the metallurgical compatibility (degree of solid solubility when the two metals are melted together). Increasing degree of incompatibility reduces wear, leading to higher value of the wear coefficients.

**Table 21.3** Coefficient of Friction and Wear Coefficients for Various Materials in the Unlubricated Sliding

Materials		Vickers Microhardness (kg/mm <sup>2</sup> )	Coefficient of Friction	Wear Coefficient ( <i>k</i> )
Wearing Surface	Counter Surface			
Mild steel	Mild steel	186	0.62	$7.0 \cdot 10^{-3}$
60/40 leaded brass	Tool steel	95	0.24	$6.0 \cdot 10^{-4}$
Ferritic stainless steel	Tool steel	250	0.53	$1.7 \cdot 10^{-5}$
Stellite	Tool steel	690	0.60	$5.5 \cdot 10^{-5}$
PTFE	Tool steel	5	0.18	$2.4 \cdot 10^{-5}$
Polyethylene	Tool steel	17	0.53	$1.3 \cdot 10^{-7}$
Tungsten carbide	Tungsten carbide	1300	0.35	$1.0 \cdot 10^{-6}$

Source: Archard, J. F. 1980. Wear theory and mechanisms. In *Wear Control Handbook*, ed. M. B. Peterson and W. O. Winer, pp. 35–80. ASME, New York.

Note: Load = 3.9 N; speed = 1.8 m/s. The stated value of the hardness is that of the softer (wearing) material in each example.

## Abrasive Wear

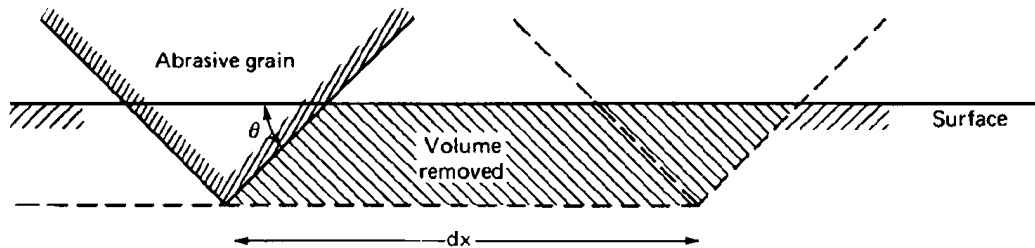
Abrasive wear occurs when a rough, hard surface slides on a softer surface and ploughs a series of grooves in it. The surface can be ploughed (plastically deformed) without removal of material. However, after the surface has been ploughed several times, material removal can occur by a low-cycle fatigue mechanism. Abrasive wear is also sometimes called *ploughing*, *scratching*, *scoring*, *gouging*, or *cutting*, depending on the degree of severity. There are two general situations for this type of wear. In the first case the hard surface is the harder of two rubbing surfaces (two-body abrasion), for example, in mechanical operations such as grinding, cutting, and machining. In the second case the hard surface is a third body, generally a small particle of grit or abrasive, caught between the two other surfaces and sufficiently harder that it is able to abrade either one or both of the mating surfaces (three-body abrasion), for example, in lapping and polishing. In many cases the wear mechanism at the start is adhesive, which generates wear debris that gets trapped at the interface, resulting in a three-body abrasive wear.

To derive a simple quantitative expression for abrasive wear, we assume a conical asperity on the hard surface (Fig. 21.7). Then the volume of wear removed is given as follows [Rabinowicz, 1965]:

$$V = kWx \overline{\tan \theta} / H \quad (21.7)$$

where  $\overline{\tan \theta}$  is a weighted average of the  $\tan \theta$  values of all the individual cones and  $k$  is a factor that includes the geometry of the asperities and the probability that a given asperity cuts (removes) rather than ploughs. Thus, the roughness effect on the volume of wear is very distinct.

**Figure 21.7** Abrasive wear model in which a cone removes material from a surface. (Source: Rabinowicz, E. 1965. *Friction and Wear of Materials*. John Wiley & Sons, New York. With permission.)



## Fatigue Wear

Subsurface and surface fatigue are observed during repeated rolling and sliding, respectively. For pure rolling condition the maximum shear stress responsible for nucleation of cracks occurs some distance below the surface, and its location moves towards the surface with an application of the friction force at the interface. The repeated loading and unloading cycles to which the materials are exposed may induce the formation of subsurface or surface cracks, which eventually, after a critical number of cycles, will result in the breakup of the surface with the formation of large fragments, leaving large pits in the surface. Prior to this critical point, negligible wear takes place, which is in marked contrast to the wear caused by adhesive or abrasive mechanism, where wear causes a gradual deterioration from the start of running. Therefore, the amount of material removed by fatigue wear is not a useful parameter. Much more relevant is the useful life in terms of the number of revolutions or time before fatigue failure occurs. Time to fatigue failure is dependent on the amplitude of the reversed shear stresses, the interface lubrication conditions, and the fatigue properties of the rolling materials.

## Impact Wear

Two broad types of wear phenomena belong in the category of impact wear: erosive and percussive wear. Erosion can occur by jets and streams of solid particles, liquid droplets, and implosion of bubbles formed in the fluid. Percussion occurs from repetitive solid body impacts. Erosive wear by impingement of solid particles is a form of abrasion that is generally treated rather differently because the contact stress arises from the kinetic energy of a particle flowing in an air or liquid stream as it encounters a surface. The particle velocity and impact angle combined with the size of the abrasive give a measure of the kinetic energy of the erosive stream. The volume of wear is proportional to the kinetic energy of the impinging particles, that is, to the square of the velocity.

Wear rate dependence on the impact angle differs between ductile and brittle materials. [Bitter, 1963].

When small drops of liquid strike the surface of a solid at high speeds (as low as 300 m/s), very high pressures are experienced, exceeding the yield strength of most materials. Thus, plastic deformation or fracture can result from a single impact, and repeated impact leads to pitting and erosive wear. Cavitation erosion arises when a solid and fluid are in relative motion and bubbles formed in the fluid become unstable and implode against the surface of the solid. Damage by this process is found in such components as ships' propellers and centrifugal pumps.

Percussion is a repetitive solid body impact, such as experienced by print hammers in high-speed electromechanical applications and high asperities of the surfaces in a gas bearing (e.g., head-medium interface in magnetic storage systems). In most practical machine applications the impact is associated with sliding; that is, the relative approach of the contacting surfaces has both normal and tangential components known as *compound impact* [Engel, 1976].

## Corrosive Wear

Corrosive wear occurs when sliding takes place in a corrosive environment. In the absence of sliding, the products of the corrosion (e.g., oxides) would form a film typically less than a micrometer thick on the surfaces, which would tend to slow down or even arrest the corrosion, but the sliding action wears the film away, so that the corrosive attack can continue. Thus, corrosive wear requires both corrosion and rubbing. Machineries operating in an industrial environment or near the coast generally corrode more rapidly than those operating in a clean environment. Corrosion can occur because of chemical or electrochemical interaction of the interface with the environment. Chemical corrosion occurs in a highly corrosive environment and in high temperature and high humidity environments. Electrochemical corrosion is a chemical reaction accompanied by the passage of an electric current, and for this to occur a potential difference must exist between two regions.

## Electrical Arc- Induced Wear

When a high potential is present over a thin air film in a sliding process, a dielectric breakdown results that leads to arcing. During arcing, a relatively high-power density (on the order of  $1 \text{ kW/mm}^2$ ) occurs over a very short period of time (on the order of  $100 \mu\text{s}$ ). The heat affected zone is usually very shallow (on the order of  $50 \mu\text{m}$ ). Heating is caused by the Joule effect due to the high power density and by ion bombardment from the plasma above the surface. This heating results in considerable melting, corrosion, hardness changes, other phase changes, and even the direct ablation of material. Arcing causes large craters, and any sliding or oscillation after an arc either shears or fractures the lips, leading to abrasion, corrosion, surface fatigue, and fretting. Arcing can thus initiate several modes of wear, resulting in catastrophic failures in electrical machinery [Bhushan and Davis, 1983].

## Fretting and Fretting Corrosion

Fretting occurs where low-amplitude vibratory motion takes place between two metal surfaces loaded together [Anonymous, 1955]. This is a common occurrence because most machinery is subjected to vibration, both in transit and in operation. Examples of vulnerable components are shrink fits, bolted parts, and splines. Basically, fretting is a form of adhesive or abrasive wear where the normal load causes adhesion between asperities and vibrations cause ruptures, resulting in wear debris. Most commonly, fretting is combined with corrosion, in which case the wear mode is known as *fretting corrosion*.

## 21.5 Lubrication

---

Sliding between clean solid surfaces is generally characterized by a high coefficient of friction and severe wear due to the specific properties of the surfaces, such as low hardness, high surface energy, reactivity, and mutual solubility. Clean surfaces readily adsorb traces of foreign substances, such as organic compounds, from the environment. The newly formed surfaces generally have a much lower coefficient of friction and wear than the clean surfaces. The presence of a layer of foreign material at an interface cannot be guaranteed during a sliding process; therefore, lubricants are deliberately applied to produce low friction and wear. The term **lubrication** is applied to two different situations: solid lubrication and fluid (liquid or gaseous) film lubrication.

### Solid Lubrication

A solid lubricant is any material used in bulk or as a powder or a thin, solid film on a surface to provide protection from damage during relative movement to reduce friction and wear. Solid lubricants are used for applications in which any sliding contact occurs, for example, a bearing operative at high loads and low speeds and a hydrodynamically lubricated bearing requiring start/stop operations. The term *solid lubricants* embraces a wide range of materials that provide low friction and wear [Bhushan and Gupta, 1991]. Hard materials are also used for low wear under extreme operating conditions.

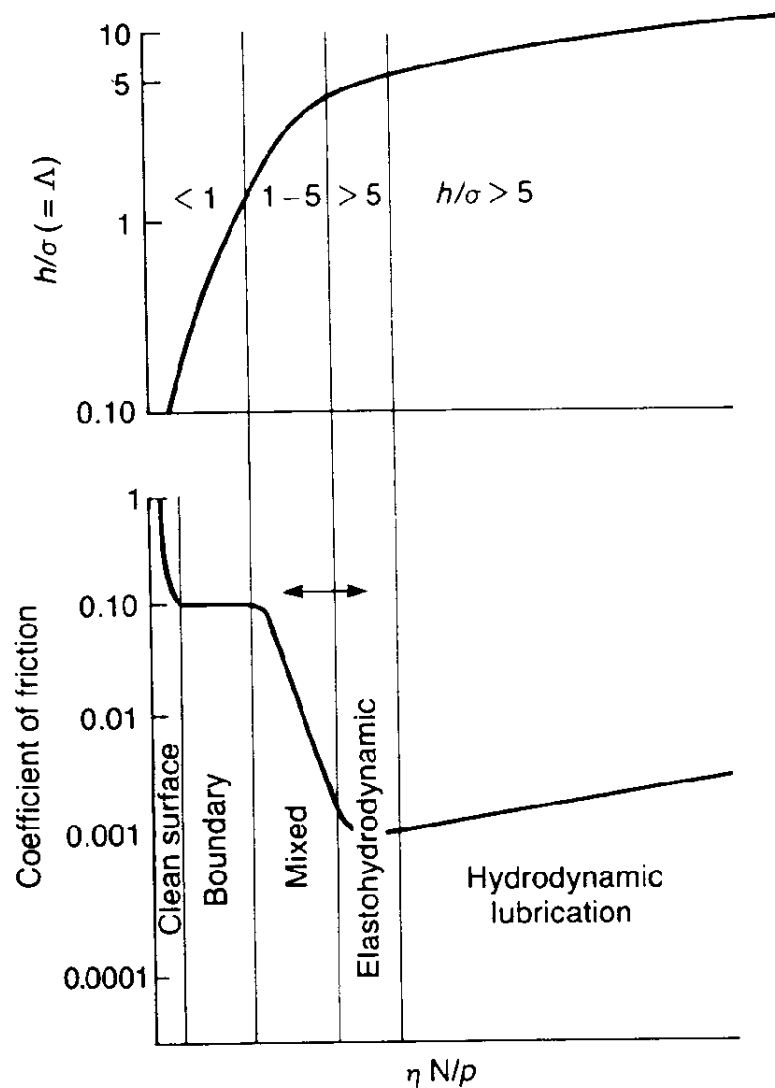
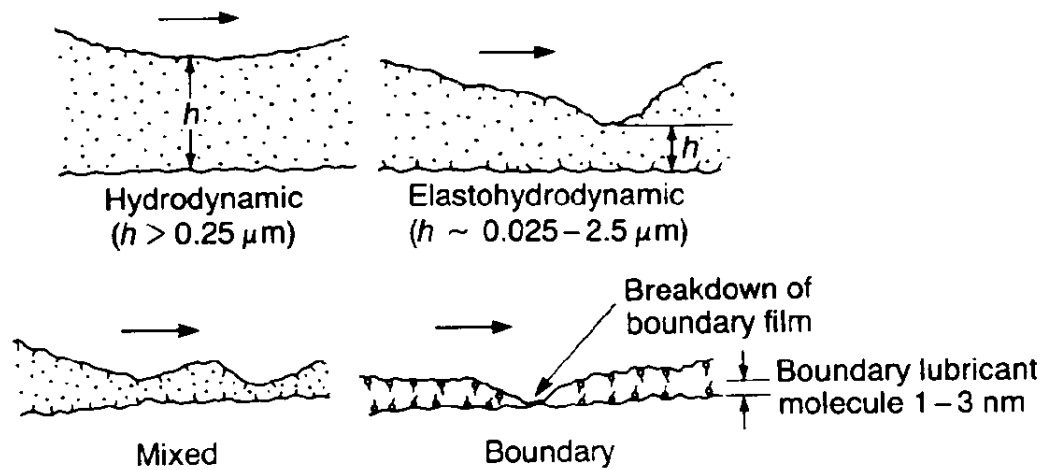
### Fluid Film Lubrication

A regime of lubrication in which a thick fluid film is maintained between two sliding surfaces by an external pumping agency is called *hydrostatic lubrication*.

A summary of the lubrication regimes observed in fluid (liquid or gas) lubrication without an external pumping agency (self-acting) can be found in the familiar Stribeck curve in Fig. 21.8. This plot for a hypothetical fluid-lubricated bearing system presents the coefficient of friction as a function of the product of viscosity ( $\eta$ ) and rotational speed ( $N$ ) divided by the normal pressure ( $p$ ). The curve has a minimum, which immediately suggests that more than one lubrication mechanism is involved. The regimes of lubrication are sometimes identified by a lubricant film parameter  $\Lambda$  equal to  $h/\sigma$ , which is mean film thickness divided by composite standard deviation of surface roughnesses. Descriptions of different regimes of lubrication follow [Booser, 1984; Bhushan, 1990].



**Figure 21.8** Lubricant film parameter ( $\Lambda$ ) and coefficient of friction as a function of  $\eta N/p$  (Stribeck curve) showing different lubrication regimes observed in fluid lubrication without an external pumping agency. Schematics of interfaces operating in different lubrication regimes are also shown.



## Hydrostatic Lubrication

Hydrostatic bearings support load on a thick film of fluid supplied from an external pressure source—a pump—which feeds pressurized fluid to the film. For this reason, these bearings are often called "externally pressurized." Hydrostatic bearings are designed for use with both incompressible and compressible fluids. Since hydrostatic bearings do not require relative motion of the bearing surfaces to build up the load-supporting pressures as necessary in hydrodynamic bearings, hydrostatic bearings are used in applications with little or no relative motion between the surfaces. Hydrostatic bearings may also be required in applications where, for one reason or another, touching or rubbing of the bearing surfaces cannot be permitted at startup and shutdown. In addition, hydrostatic bearings provide high stiffness. Hydrostatic bearings, however, have the disadvantage of requiring high-pressure pumps and equipment for fluid cleaning, which adds to space and cost.

## Hydrodynamic Lubrication

Hydrodynamic (HD) lubrication is sometimes called *fluid-film* or *thick-film lubrication*. As a bearing with convergent shape in the direction of motion starts to spin (slide in the longitudinal direction) from rest, a thin layer of fluid is pulled through because of viscous entrainment and is then compressed between the bearing surfaces, creating a sufficient (hydrodynamic) pressure to support the load without any external pumping agency. This is the principle of hydrodynamic lubrication, a mechanism that is essential to the efficient functioning of the self-acting journal and thrust bearings widely used in modern industry. A high load capacity can be achieved in the bearings that operate at high speeds and low loads in the presence of fluids of high viscosity.

Fluid film can also be generated only by a reciprocating or oscillating motion in the normal direction (*squeeze*), which may be fixed or variable in magnitude (transient or steady state). This load-carrying phenomenon arises from the fact that a viscous fluid cannot be instantaneously squeezed out from the interface with two surfaces that are approaching each other. It takes time for these surfaces to meet, and during that interval—because of the fluid's resistance to extrusion—a pressure is built up and the load is actually supported by the fluid film. When the load is relieved or becomes reversed, the fluid is sucked in and the fluid film often can recover its thickness in time for the next application. The squeeze phenomenon controls the buildup of a water film under the tires of automobiles and airplanes on wet roadways or landing strips (commonly known as *hydroplaning*) that have virtually no relative sliding motion.

HD lubrication is often referred to as the ideal lubricated contact condition because the lubricating films are normally many times thicker (typically 5–500  $\mu\text{m}$ ) than the height of the irregularities on the bearing surface, and solid contacts do not occur. The coefficient of friction in the HD regime can be as small as 0.001 (Fig. 21.8). The friction increases slightly with the sliding speed because of viscous drag. The behavior of the contact is governed by the bulk physical properties of the lubricant, notable viscosity, and the frictional characteristics arise purely from the shearing of the viscous lubricant.

## Elastohydrodynamic Lubrication

Elastohydrodynamic (EHD) lubrication is a subset of HD lubrication in which the elastic deformation of the bounding solids plays a significant role in the HD lubrication process. The film thickness in EHD lubrication is thinner (typically  $0.5\text{--}2.5\ \mu\text{m}$ ) than that in HD lubrication (Fig. 21.8), and the load is still primarily supported by the EHD film. In isolated areas, asperities may actually touch. Therefore, in liquid lubricated systems, boundary lubricants that provide boundary films on the surfaces for protection against any solid-solid contact are used. Bearings with heavily loaded contacts fail primarily by a fatigue mode that may be significantly affected by the lubricant. EHD lubrication is most readily induced in heavily loaded contacts (such as machine elements of low geometrical conformity), where loads act over relatively small contact areas (on the order of one-thousandth of journal bearing), such as the point contacts of ball bearings and the line contacts of roller bearings and gear teeth. EHD phenomena also occur in some low elastic modulus contacts of high geometrical conformity, such as seals and conventional journal and thrust bearings with soft liners.

## Mixed Lubrication

The transition between the hydrodynamic/elastohydrodynamic and boundary lubrication regimes constitutes a gray area known as *mixed lubrication*, in which two lubrication mechanisms may be functioning. There may be more frequent solid contacts, but at least a portion of the bearing surface remains supported by a partial hydrodynamic film (Fig. 21.8). The solid contacts, if between unprotected virgin metal surfaces, could lead to a cycle of adhesion, metal transfer, wear particle formation, and snowballing into seizure. However, in liquid lubricated bearings, the physisorbed or chemisorbed or chemically reacted films (boundary lubrication) prevent adhesion during most asperity encounters. The mixed regime is also sometimes referred to as *quasihydrodynamic*, *partial fluid*, or *thin-film* (typically  $0.5\text{--}2.5\ \mu\text{m}$ ) *lubrication*.

## Boundary Lubrication

As the load increases, speed decreases or the fluid viscosity decreases in the Stribeck curve shown in Fig. 21.8; the coefficient of friction can increase sharply and approach high levels (about 0.2 or much higher). In this region it is customary to speak of boundary lubrication. This condition can also occur in a starved contact. Boundary lubrication is that condition in which the solid surfaces are so close together that surface interaction between monomolecular or multimolecular films of lubricants (liquids or gases) and the solids dominate the contact. (This phenomenon does not apply to solid lubricants.) The concept is represented in Fig. 21.8, which shows a microscopic cross section of films on two surfaces and areas of asperity contact. In the absence of boundary lubricants and gases (no oxide films), friction may become very high ( $>1$ ).

## 21.6 Micro/nanotribology

---

AFM/FFMs are commonly used to study engineering surfaces on micro- to nanoscales. These instruments measure the normal and friction forces between a sharp tip (with a tip radius of  $30\text{--}100\ \text{nm}$ ) and an engineering surface. Measurements can be made at loads as low as less than 1 nN and at scan rates up to about 120 Hz. A sharp AFM/ FFM tip sliding on a surface simulates a single asperity contact. FFMs are used to measure coefficient of friction on micro- to nanoscales

and AFMs are used for studies of surface topography, scratching/wear and boundary lubrication, mechanical property measurements, and nanofabrication/nanomachining [Bhushan and Ruan, 1994; Bhushan *et al.*, 1994; Bhushan and Koinkar, 1994a,b; Ruan and Bhushan, 1994; Bhushan, 1995; Bhushan *et al.*, 1995]. For surface roughness, friction force, nanoscratching and nanowear measurements, a microfabricated square pyramidal  $\text{Si}_3\text{N}_4$  tip with a tip radius of about 30 nm is generally used at loads ranging from 10 to 150 nN. For microscratching, microwear, nanoindentation hardness measurements, and nanofabrication, a three-sided pyramidal single-crystal natural diamond tip with a tip radius of about 100 nm is used at relatively high loads ranging from 10  $\mu\text{N}$  to 150  $\mu\text{N}$ . Friction and wear on micro- and nanoscales are found to be generally smaller compared to that at macroscales. For an example of comparison of coefficients of friction at macro- and microscales see Table 21.4.

**Table 21.4** Surface Roughness and Micro- and Macroscale Coefficients of Friction of Various Samples

Material	RMS Roughness,nm	Microscale Coefficient of Friction versus $\text{Si}_3\text{N}_4$ Tip <sup>1</sup>	Macroscale Coefficient of Friction versus Alumina Ball <sup>2</sup>	
			0.1 N	1 N
Si (111)	0.11	0.03	0.18	0.60
C <sup>+</sup> -implanted Si	0.33	0.02	0.18	0.18

<sup>1</sup> $\text{Si}_3\text{N}_4$  tip (with about 50 nm radius) in the load range of 10–150 nN (1.5–3.8 GPa), a scanning speed of 4  $\mu\text{m/s}$  and scan area of 1  $\mu\text{m} \times 1 \mu\text{m}$ .

<sup>2</sup>Alumina ball with 3-mm radius at normal loads of 0.1 and 1 N (0.23 and 0.50 GPa) and average sliding speed of 0.8 mm/s.

## Defining Terms

**Friction:** The resistance to motion whenever one solid slides over another.

**Lubrication:** Materials applied to the interface to produce low friction and wear in either of two situations—solid lubrication or fluid (liquid or gaseous) film lubrication.

**Micro/nanotribology:** The discipline concerned with experimental and theoretical investigations of processes (ranging from atomic and molecular scales to microscales) occurring during adhesion, friction, wear, and lubrication at sliding surfaces.

**Tribology:** The science and technology of two interacting surfaces in relative motion and of related subjects and practices.

**Wear:** The removal of material from one or both solid surfaces in a sliding, rolling, or impact motion relative to one another.

## References

- Anonymous. 1955. Fretting and fretting corrosion. *Lubrication*. 41:85–96.
- Archard, J. F. 1953. Contact and rubbing of flat surfaces. *J. Appl. Phys.* 24:981–988.
- Archard, J. F. 1980. Wear theory and mechanisms. *Wear Control Handbook*, ed. M. B. Peterson and W. O. Winer, pp. 35–80. ASME, New York.
- Avallone, E. A. and Baumeister, T., III. 1987. *Marks' Standard Handbook for Mechanical Engineers*, 9th ed. McGraw-Hill, New York.
- Benzing, R., Goldblatt, I., Hopkins, V., Jamison, W., Mecklenburg, K., and Peterson, M. 1976. *Friction and Wear Devices*, 2nd ed. ASLE, Park Ridge, IL.
- Bhushan, B. 1984. Analysis of the real area of contact between a polymeric magnetic medium and a rigid surface. *ASME J. Lub. Tech.* 106:26–34.
- Bhushan, B. 1990. *Tribology and Mechanics of Magnetic Storage Devices*. Springer-Verlag, New York.
- Bhushan, B. 1992. *Mechanics and Reliability of Flexible Magnetic Media*. Springer-Verlag, New York.
- Bhushan, B. 1995. *Handbook of Micro/Nanotribology*. CRC Press, Boca Raton, FL.
- Bhushan, B. and Davis, R. E. 1983. Surface analysis study of electrical-arc-induced wear. *Thin Solid Films*. 108:135–156.
- Bhushan, B., Davis, R. E., and Gordon, M. 1985a. Metallurgical re-examination of wear modes. I: Erosive, electrical arcing and fretting. *Thin Solid Films*. 123:93–112.
- Bhushan, B., Davis, R. E., and Kolar, H. R. 1985b. Metallurgical re-examination of wear modes. II: Adhesive and abrasive. *Thin Solid Films*. 123:113–126.
- Bhushan, B. and Gupta, B. K. 1991. *Handbook of Tribology: Materials, Coatings, and Surface Treatments*. McGraw-Hill, New York.
- Bhushan, B., Israelachvili, J. N., and Landman, U. 1995. Nanotribology: Friction, Wear and Lubrication at the Atomic Scale. *Nature*. 374:607–616.
- Bhushan, B. and Koinkar, V. N. 1994a. Tribological studies of silicon for magnetic recording applications. *J. Appl. Phys.* 75:5741–5746.
- Bhushan, B. and Koinkar, V. N. 1994b. Nanoindentation hardness measurements using atomic force microscopy. *Appl. Phys. Lett.* 64:1653–1655.
- Bhushan, B., Koinkar, V. N., and Ruan, J. 1994. Microtribology of magnetic media. *Proc. Inst. Mech. Eng., Part J: J. Eng. Tribol.* 208:17–29.
- Bhushan, B. and Ruan, J. 1994. Atomic-scale friction measurements using friction force microscopy: Part II—Application to magnetic media. *ASME J. Tribology*. 116:389–396.
- Binnig, G., Quate, C. F., and Gerber, C. 1986. Atomic force microscope. *Phys. Rev. Lett.* 56:930–933.
- Binnig, G., Rohrer, H., Gerber, C., and Weibel, E. 1982. Surface studies by scanning tunnelling microscopy. *Phys. Rev. Lett.* 49:57–61.
- Bitter, J. G. A. 1963. A study of erosion phenomena. *Wear*. 6:5–21; 169–190.
- Booser, E. R. 1984. *CRC Handbook of Lubrication*, vol. 2. CRC Press, Boca Raton, FL.
- Bowden, F. P. and Tabor, D. 1950. *The Friction and Lubrication of Solids*, vols. I and II. Clarendon Press, Oxford.
- Davidson, C. S. C. 1957. Bearing since the stone age. *Engineering*. 183:2–5.

- Dowson, D. 1979. *History of Tribology*. Longman, London.
- Engel, P. A. 1976. *Impact Wear of Materials*. Elsevier, Amsterdam.
- Fuller, D. D. 1984. *Theory and Practice of Lubrication for Engineers*, 2nd ed. John Wiley & Sons, New York.
- Georges, J. M., Millot, S., Loubet, J. L., and Tonck, A. 1993. Drainage of thin liquid films between relatively smooth surfaces. *J. Chem. Phys.* 98:7345–7360.
- Georges, J. M., Tonck, A., and Mazuyer, D. 1994. Interfacial friction of wetted monolayers. *Wear*. 175:59–62.
- Greenwood, J. A. and Williamson, J. B. P. 1966. Contact of nominally flat surfaces. *Proc. R. Soc. Lond.* A295:300–319.
- Holm, R. 1946. *Electrical Contact*. Springer-Verlag, New York.
- Israelachvili, J. N. and Adams, G. E. 1978. Measurement of friction between two mica surfaces in aqueous electrolyte solutions in the range 0–100 nm. *Chem. Soc. J., Faraday Trans. I.* 74:975–1001.
- Jost, P. 1966. *Lubrication (Tribology)*<sup>3/4</sup>A Report on the Present Position and Industry's Needs. Department of Education and Science, H.M. Stationary Office, London.
- Jost, P. 1976. Economic impact of tribology. *Proc. Mechanical Failures Prevention Group*. NBS Special Pub. 423, Gaithersburg, MD.
- Klein, J. 1980. Forces between mica surfaces bearing layers of adsorbed polystyrene in Cyclohexane. *Nature*. 288:248–250.
- Layard, A. G. 1853. *Discoveries in the Ruins of Nineveh and Babylon*, I and II. John Murray, Albemarle Street, London.
- Mate, C. M., McClelland, G. M., Erlandsson, R., and Chiang, S. 1987. Atomic-scale friction of a tungsten tip on a graphite surface. *Phys. Rev. Lett.* 59:1942–1945.
- Parish, W. F. 1935. Three thousand years of progress in the development of machinery and lubricants for the hand crafts. *Mill and Factory*. Vols. 16 and 17.
- Peachey, J., Van Alsten, J., and Granick, S. 1991. Design of an apparatus to measure the shear response of ultrathin liquid films. *Rev. Sci. Instrum.* 62:463–473.
- Petroff, N. P. 1883. Friction in machines and the effects of the lubricant. *Eng. J.* (in Russian; St. Petersburg) 71–140, 228–279, 377–436, 535–564.
- Rabinowicz, E. 1965. *Friction and Wear of Materials*. John Wiley & Sons, New York.
- Rabinowicz, E. 1980. Wear coefficients—metals. *Wear Control Handbook*, ed. M. B. Peterson and W. O. Winer, pp. 475–506. ASME, New York.
- Reynolds, O. O. 1886. On the theory of lubrication and its application to Mr. Beauchamp Tower's experiments. *Phil. Trans. R. Soc. (Lond.)* 177:157–234.
- Ruan, J. and Bhushan, B. 1994. Atomic-scale and microscale friction of graphite and diamond using friction force microscopy. *J. Appl. Phys.* 76:5022–5035.
- Tabor, D. and Winterton, R. H. S. 1969. The direct measurement of normal and retarded van der Waals forces. *Proc. R. Soc. Lond.* A312:435–450.
- Tonck, A., Georges, J. M., and Loubet, J. L. 1988. Measurements of intermolecular forces and the rheology of dodecane between alumina surfaces. *J. Colloid Interf. Sci.* 126:1540–1563.

Tower, B. 1884. Report on friction experiments. *Proc. Inst. Mech. Eng.* 632.

## **Further Information**

Major conferences:

ASME/STLE Tribology Conference held every October in the U.S.

Leeds-Lyon Symposium on Tribology held every year at Leeds, U.K., or Lyon, France (alternating locations).

International Symposium on Advances in Information Storage and Processing Systems held annually at ASME International Congress and Exposition in November/December in the U.S.

International Conference on Wear of Materials held every two years; next one to be held in 1995.

Eurotrib held every four years; next one to be held in 1997.

Societies:

Information Storage and Processing Systems Division, The American Society of Mechanical Engineers, New York.

Tribology Division, The American Society of Mechanical Engineers, New York.

Institution of Mechanical Engineers, London, U.K.

Society of Tribologists and Lubrication Engineers, Park Ridge, IL.

Pennock, G. R. "Machine Elements"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



**22.1 Threaded Fasteners****22.2 Clutches and Brakes**

Rim-Type Clutches and Brakes • Axial-Type Clutches and Brakes • Disk Clutches and Brakes • Cone Clutches and Brakes • Positive-Contact Clutches

**Gordon R. Pennock**

*Purdue University*

Section 22.1 presents a discussion of threaded fasteners, namely, the nut and bolt, the machine screw, the cap screw, and the stud. Equations are presented for the spring stiffness of the portion of a bolt, or a cap screw, within the clamped zone, which generally consists of the unthreaded shank portion and the threaded portion. Equations for the resultant bolt load and the resultant load on the members are also included in the discussion. The section concludes with a relation that provides an estimate of the torque that is required to produce a given preload. Section 22.2 presents a discussion of clutches and brakes and the important features of these machine elements. Various types of frictional-contact clutches and brakes are included in the discussion, namely, the radial, axial, disk, and cone types. Information on positive-contact clutches and brakes is also provided. The section includes energy considerations, equations for the temperature-rise, and the characteristics of a friction material.

---

**22.1 Threaded Fasteners**

---

The bolted joint with hardened steel washers is a common solution when a connection is required that can be easily disassembled (without destructive methods) and is strong enough to resist external tensile loads and shear loads. The clamping load, which is obtained by twisting the nut until the bolt is close to the elastic limit, stretches or elongates the bolt. This bolt tension will remain as the clamping force, or preload, providing the nut does not loosen. The preload induces compression in the members, which are clamped together, and exists in the connection after the nut has been properly tightened, even if there is no external load. Care must be taken to ensure that a bolted joint is properly designed and assembled [Blake, 1986]. When tightening the connection, the bolt head should be held stationary and the nut twisted. This procedure will ensure that the bolt shank will not experience the thread-friction torque. During the tightening process, the first thread on the nut tends to carry the entire load. However, yielding occurs with some strengthening due to the cold work that takes place, and the load is eventually distributed over about three nut threads. For this reason, it is recommended that nuts should not be reused; in fact, it can be dangerous if

this practice is adopted [Shigley and Mischke, 1989].

There are several styles of hexagonal nut, namely, (1) the general hexagonal nut, (2) the washer-faced regular nut, (3) the regular nut chamfered on both sides, (4) the jam nut with washer face, and (5) the jam nut chamfered on both sides. Flat nuts only have a chamfered top [Shigley and Mischke, 1986]. The material of the nut must be selected carefully to match that of the bolt. Carbon steel nuts are usually made to conform to ASTM A563 Grade A specifications or to SAE Grade 2. A variety of machine screw head styles also exist; they include (1) fillister head, (2) flat head, (3) round head, (4) oval head, (5) truss head, (6) binding head, and (7) hexagonal head (trimmed and upset). There are also many kinds of locknuts, which have been designed to prevent a nut from loosening in service. Spring and lock washers placed beneath an ordinary nut are also common devices to prevent loosening.

Another tension-loaded connection uses cap screws threaded into one of the members. Cap screws can be used in the same applications as nuts and bolts and also in situations where one of the clamped members is threaded. The common head styles of the cap screw include (1) hexagonal head, (2) fillister head, (3) flat head, and (4) hexagonal socket head. The head of a hexagon-head cap screw is slightly thinner than that of a hexagon-head bolt. An alternative to the cap screw is the stud, which is a rod threaded on both ends. Studs should be screwed into the lower member first, then the top member should be positioned and fastened down with hardened steel washers and nuts. The studs are regarded as permanent and the joint should be disassembled by removing only the nuts and washers. In this way, the threaded part of the lower member is not damaged by reusing the threads.

The grip of a connection is the total thickness of the clamped material [Shigley and Mischke, 1989]. In the bolted joint the grip is the sum of the thicknesses of both the members and the washers. In a stud connection the grip is the thickness of the top member plus that of the washer. The spring stiffness, or spring rate, of an elastic member such as a bolt is the ratio of the force applied to the member and the deflection caused by that force. The spring stiffness of the portion of a bolt, or cap screw, within the clamped zone generally consists of two parts, namely, (1) that of the threaded portion, and (2) that of the unthreaded shank portion. Therefore, the stiffness of a bolt is equivalent to the stiffness of two springs in series:

$$\frac{1}{k_b} = \frac{1}{k_T} + \frac{1}{k_d} \quad \text{or} \quad k_b = \frac{k_T k_d}{k_T + k_d} \quad (22.1)$$

The spring stiffnesses of the threaded and unthreaded portions of the bolt in the clamped zone, respectively, are

$$k_T = \frac{A_t E}{L_T} \quad \text{and} \quad k_d = \frac{A_d E}{L_d} \quad (22.2)$$

where  $A_t$  is the tensile-stress area,  $L_T$  is the length of the threaded portion in the grip,  $A_d$  is the major-diameter area of the fastener,  $L_d$  is the length of the unthreaded portion in the grip, and  $E$  is the modulus of elasticity. Substituting Eq. (22.2) into Eq. (22.1), the estimated effective stiffness of the bolt (or cap screw) in the clamped zone can be expressed as

$$k_b = \frac{A_t A_d E}{A_t L_d + A_d L_T} \quad (22.3)$$

For short fasteners the unthreaded area is small and so the first of the expressions in Eq. (22.2) can be used to evaluate  $k_b$ . In the case of long fasteners the threaded area is relatively small, so the second expression in Eq. (22.2) can be used to evaluate the effective stiffness of the bolt. Expressions can also be obtained for the stiffness of the members in the clamped zone [Juvinall, 1983]. Both the stiffness of the fastener and the stiffness of the members in the clamped zone must be known in order to understand what happens when the connection is subjected to an external tensile load. There may of course be more than two members included in the grip of the fastener. Taken together the members act like compressive springs in series, and hence the total spring stiffness of the members is

$$\frac{1}{k_m} = \frac{1}{k_1} + \frac{1}{k_2} + \frac{1}{k_3} + \cdots \quad (22.4)$$

If one of the members is a soft gasket, its stiffness relative to the other members is usually so small that for all practical purposes the other members can be neglected and only the gasket stiffness need be considered. If there is no gasket, the stiffness of the members is difficult to obtain, except by experimentation, because the compression spreads out between the bolt head and the nut and hence the area is not uniform. There are, however, some cases in which this area can be determined. Ultrasonic techniques have been used to determine the pressure distribution at the member interface in a bolt-flange assembly [Ito *et al.*, 1977]. The results show that the pressure stays high out to about 1.5 times the bolt radius and then falls off farther away from the bolt. Rotsher's pressure-cone method has been suggested for stiffness calculations with a variable cone angle. This method is quite complicated and a simpler approach is to use a fixed cone angle [Little, 1967].

Consider what happens when an external tensile load is applied to a bolted connection. Assuming that the preload has been correctly applied (by tightening the nut before the external tensile load is applied), the tensile load causes the connection to stretch through some distance. This elongation can be related to the stiffness of the bolts, or the members, by the equation

$$\delta = \frac{P_b}{k_b} = \frac{P_m}{k_m} \quad \text{or} \quad P_b = \frac{k_b}{k_m} P_m \quad (22.5)$$

where  $P_b$  is the portion of the external tensile load  $P$  taken by the bolt and  $P_m$  is the portion of  $P$  taken by the members. Since the external tensile load  $P$  is equal to  $P_b + P_m$ ,

$$P_b = \left( \frac{k_b}{k_b + k_m} \right) P \quad \text{and} \quad P_m = \left( \frac{k_m}{k_b + k_m} \right) P \quad (22.6)$$

The resultant bolt load is  $F_b = P_b + F_i$  and the resultant load on the members is  $F_m = P_m - F_i$ , where  $F_i$  is the preload. Therefore, the resultant bolt load can be written as

$$F_b = \left( \frac{k_b}{k_b + k_m} \right) P + F_i, \quad F_m < 0 \quad (22.7)$$

and the resultant load on the members can be written as

$$F_m = \left( \frac{k_m}{k_b + k_m} \right) P - F_i, \quad F_m < 0 \quad (22.8)$$

Equations (22.7) and (22.8) are only valid for the case when some clamping load remains in the members, which is indicated by the qualifier in the two equations. Making the grip longer causes the members to take an even greater percentage of the external load. If the external load is large enough to completely remove the compression, then the members will separate and the entire load will be carried by the bolts.

Since it is desirable to have a high preload in important bolted connections, methods of ensuring that the preload is actually developed when the parts are assembled must be considered. If the overall length of the bolt,  $L_b$ , can be measured (say with a micrometer) when the parts are assembled, then the bolt elongation due to the preload  $F_i$  can be computed from the relation

$$\delta = \frac{F_i L_b}{AE} \quad (22.9)$$

where  $A$  is the cross-sectional area of the bolt. The nut can then be tightened until the bolt elongates through the distance  $\delta$ , which ensures that the desired preload has been obtained. In many cases, however, it is not practical or possible to measure the bolt elongation. For example, the elongation of a screw cannot be measured if the threaded end is in a blind hole. In such cases the wrench torque that is required to develop the specified preload must be estimated. Torque wrenching, pneumatic-impact wrenching, or the **turn-of-the-nut method** can be used [Blake and Kurtz, 1965]. The torque wrench has a built-in dial that indicates the proper torque. With pneumatic-impact wrenching, the air pressure is adjusted so that the wrench stalls when the proper torque is obtained or, in some cases, the air shuts off automatically at the desired torque.

The **snug-tight condition** is defined as the tightness attained by a few impacts of an impact wrench or the full effort of a person using an ordinary wrench. When the snug-tight condition is attained, all additional turning develops useful tension in the bolt. The turn-of-the-nut method requires that fractional number of turns necessary to develop the required preload from the snug-tight condition be computed. For example, for heavy hexagon structural bolts, the turn-of-the-nut specification requires that under optimum conditions the nut should be turned a minimum of  $180^\circ$  from the snug-tight condition. A good estimate of the torque required to produce a given preload  $F_i$  can be obtained from the relation [Shigley and Mischke, 1989]

$$T = \frac{F_i d_m}{2} \left( \frac{L + \pi \mu d_m \sec \alpha}{\pi d_m - \mu L \sec \alpha} \right) + \frac{F_i \mu_c d_c}{2} \quad (22.10)$$

where  $d_m$  is the mean diameter of the bolt,  $L$  is the lead of the thread,  $\alpha$  is half the thread angle,  $\mu_c$  is the coefficient of thread friction,  $\mu_c$  is the coefficient of collar friction, and  $d_c$  is the mean collar diameter. The coefficients of friction depend upon the surface smoothness, the accuracy, and the degree of lubrication. Although these items may vary considerably, it is interesting to note that on the average both  $\mu$  and  $\mu_c$  are approximately 0.15.

## 22.2 Clutches and Brakes

---

A clutch is a coupling that connects two shafts rotating at different speeds and brings the output shaft smoothly and gradually to the same speed as the input shaft. Clutches and brakes are machine elements associated with rotation and have in common the function of storing or transferring rotating energy [Remling, 1983]. When the rotating members are caused to stop by means of a brake, the kinetic energy of rotation must be absorbed by the brake. In the same way, when the members of a machine that are initially at rest are brought up to speed, slipping must occur in the clutch until the driven members have the same speed as the driver. Kinetic energy is absorbed during slippage of either a clutch or a brake, and this energy appears in the form of heat. The important features in the performance of these devices are (1) the actuating force, (2) the transmitted torque, (3) the energy loss, and (4) the temperature rise. The torque that is transmitted is related to the actuating force, the coefficient of friction, and the geometry of the device. Essentially this is a problem in statics and can be studied separately for each geometric configuration. The rise in temperature, however, can be studied without regard to the type of device because the heat-dissipating surfaces are the geometry of interest. An approximate guide to the rise in temperature in a drum brake is the horsepower per square inch [Spotts, 1985].

The torque capacity of a clutch or brake depends upon the coefficient of friction of the material and a safe normal pressure. The character of the load may be such, however, that if this torque value is permitted, the clutch or brake may be destroyed by the generated heat. Therefore, the capacity of a clutch is limited by two factors: (a) the characteristics of the material, and (b) the ability of the clutch to dissipate the frictional heat. The temperature rise of a clutch or brake assembly can be approximated by the relation

$$\Delta T = \frac{H}{CW} \quad (22.11)$$

where  $\Delta T$  is in  $^{\circ}\text{F}$ ,  $H$  is the heat generated in Btu,  $C$  is the specific heat in Btu/(lbm  $^{\circ}\text{F}$ ), and  $W$  is the mass of the clutch or brake assembly in lbm. If SI units are used, then

$$\Delta T = \frac{E}{Cm} \quad (22.12)$$

where  $\Delta T$  is in  $^{\circ}\text{C}$ ,  $E$  is the total energy dissipated during the clutching operation or the braking cycle in J,  $C$  is in J/kg  $^{\circ}\text{C}$ , and  $m$  is the mass of the clutch or brake assembly in kg. Equation (22.11) or (22.12) can be used to explain what happens when a clutch or a brake is operated. However, there are so many variables involved that it is most unlikely that the analytical results

would approximate experimental results. For this reason such analyses are only useful, for repetitive cycling, in pinpointing the design parameters that have the greatest effect on performance.

The friction material of a clutch or brake should have the following characteristics, to a degree that is dependent upon the severity of the service: (a) a high and uniform coefficient of friction, (b) imperviousness to environmental conditions, such as moisture, (c) the ability to withstand high temperatures, as well as a good heat conductivity, (d) good resiliency, and (e) high resistance to wear, scoring, and galling. The manufacture of friction materials is a highly specialized process, and the selection of a friction material for a specific application requires some expertise. Selection involves a consideration of all the characteristics of a friction material as well as the standard sizes that are available. The woven-cotton lining is produced as a fabric belt, which is impregnated with resins and polymerized. It is mostly used in heavy machinery and can be purchased in rolls up to 50 feet in length. The thicknesses that are available range from 0.125 to 1 in. and the width may be up to 12 in. A woven-asbestos lining is similar in construction to the cotton lining and may also contain metal particles. It is not quite as flexible as the cotton lining and comes in a smaller range of sizes. The woven-asbestos lining is also used as a brake material in heavy machinery.

Molded-asbestos linings contain asbestos fiber and friction modifiers; a thermoset polymer is used, with heat, to form a rigid or a semirigid molding. The principal use is in drum brakes. Molded-asbestos pads are similar to molded linings but have no flexibility; they are used for both clutches and brakes. Sintered-metal pads are made of a mixture of copper and/or iron particles with friction modifiers, molded under high pressure and then heated to a high temperature to fuse the material. These pads are used in both brakes and clutches for heavy-duty applications. Cermet pads are similar to the sintered-metal pads and have a substantial ceramic content. Typical brake linings may consist of a mixture of asbestos fibers to provide strength and ability to withstand high temperatures; various friction particles to obtain a degree of wear resistance and higher coefficient of friction; and bonding materials. Some clutch friction materials may be run wet by allowing them to dip in oil or to be sprayed by oil. This reduces the coefficient of friction, but more heat can be transferred and higher pressure can be permitted.

The two most common methods of coupling are the frictional-contact clutch and the positive-contact clutch. Other methods include the overrunning or freewheeling clutch, the magnetic clutch, and the fluid coupling. In general, the types of frictional-contact clutches and brakes can be classified as rim type or axial type [Marks, 1987]. The analysis of all types of frictional-clutches and brakes follows the same general procedure, namely, (a) determine the pressure distribution on the frictional surfaces, (b) find a relation between the maximum pressure and the pressure at any point, and (c) apply the conditions of static equilibrium to find the actuating force, the torque transmitted, and the support reactions. The analysis is useful when the dimensions are known and the characteristics of the friction material are specified. In design, however, synthesis is of more interest than analysis. Here the aim is to select a set of dimensions that will provide the best device within the limitations of the frictional material that is specified by the designer [Proctor, 1961].



## Rim-Type Clutches and Brakes

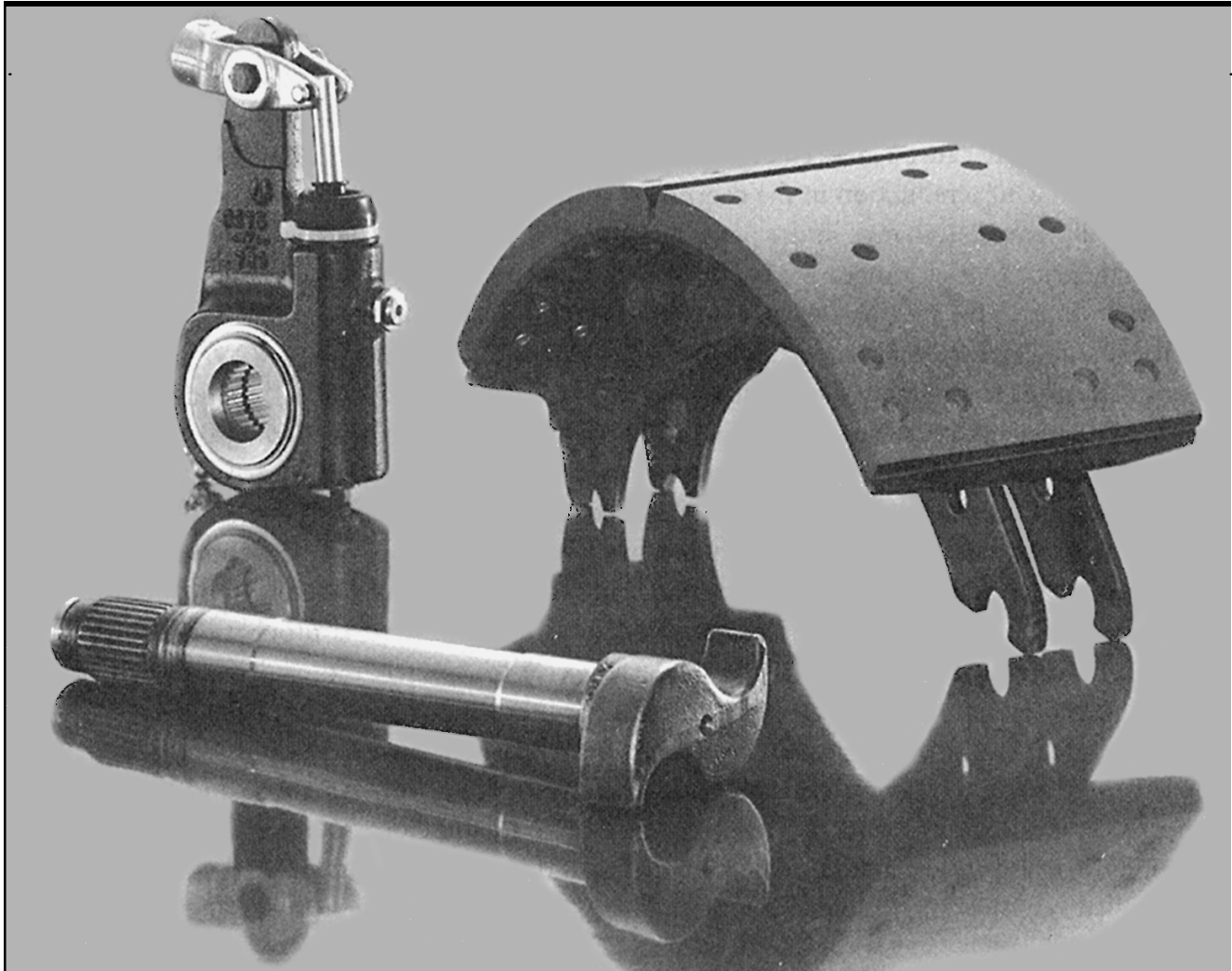
The rim-type brake can be designed for **self-energizing**, that is, using friction to reduce the actuating force. Self-energization is important in reducing the required braking effort; however, it also has a disadvantage. When rim-type brakes are used as vehicle brakes, a small change in the coefficient of friction will cause a large change in the pedal force required for braking. For example, it is not unusual for a 30% reduction in the coefficient of friction (due to a temperature change or moisture) to result in a 50% change in the pedal force required to obtain the same braking torque that was possible prior to the change.

The rim types may have internal expanding shoes or external contracting shoes. An internal shoe clutch consists essentially of three elements: (1) a mating frictional surface, (2) a means of transmitting the torque to and from the surfaces, and (3) an actuating mechanism. Depending upon the operating mechanism, such clutches can be further classified as expanding-ring, centrifugal, magnetic, hydraulic, or pneumatic. The expanding-ring clutch benefits from centrifugal effects, transmits high torque even at low speeds, and requires both positive engagement and ample release force. This type of clutch is often used in textile machinery, excavators, and machine tools in which the clutch may be located within the driving pulley. The centrifugal clutch is mostly used for automatic operation. If no spring is present, the torque transmitted is proportional to the square of the speed [Beach, 1962]. This is particularly useful for electric motor drives in which, during starting, the driven machine comes up to speed without shock. Springs can be used to prevent engagement until a certain motor speed has been reached, but some shock may occur. Magnetic clutches are particularly useful for automatic and remote-control systems and are used in drives subject to complex load cycles. Hydraulic and pneumatic clutches are useful in drives having complex loading cycles, in automatic machinery, and in manipulators. Here the fluid flow can be controlled remotely using solenoid valves. These clutches are available as disk, cone, and multiple-plate clutches.

In braking systems the internal-shoe or drum brake is used mostly for automotive applications. The actuating force of the device is applied at the end of the shoe away from the pivot. Since the shoe is usually long, the distribution of the normal forces cannot be assumed to be uniform. The mechanical arrangement permits no pressure to be applied at the heel; therefore, frictional material located at the heel contributes very little to the braking action. It is standard practice to omit the friction material for a short distance away from the heel, which also eliminates interference. In some designs the hinge pin is allowed to move to provide additional heel pressure. This gives the effect of a floating shoe. A good design concentrates as much frictional material as possible in the neighborhood of the point of maximum pressure. Typical assumptions made in an analysis of the shoe include the following: (1) the pressure at any point on the shoe is proportional to the distance from the hinge pin (zero at the heel); (2) the effect of centrifugal force is neglected (in the case of brakes, the shoes are not rotating and no centrifugal force exists; in clutch design, the effect of this force must be included in the equations of static equilibrium); (3) the shoe is rigid (in practice, some deflection will occur depending upon the load, pressure, and stiffness of the shoe; therefore, the resulting pressure distribution may be different from the assumed distribution); and (4) the entire analysis is based upon a coefficient of friction that does not vary with pressure. Actually, the coefficient may vary with a number of conditions, including temperature, wear, and the environment.

For pivoted external shoe brakes and clutches, the operating mechanisms can be classified as

solenoids, levers, linkages or toggle devices, linkages with spring loading, hydraulic devices, and pneumatic devices. It is common practice to concentrate on brake and clutch performance without the extraneous influences introduced by the need to analyze the statics of the control mechanisms. The moments of the frictional and normal forces about the hinge pin are the same as for the internal expanding shoes. It should be noted that when external contracting designs are used as clutches, the effect of the centrifugal force is to decrease the normal force. Therefore, as the speed increases, a larger value of the actuating force is required. A special case arises when the pivot is symmetrically located and also placed so that the moment of the friction forces about the pivot is zero.



## AFTERMARKET BRAKE PRODUCTS

The genuine OEM quality brake replacement parts by Rockwell are the exact components that are used for new vehicles' original equipment. Shown above are non-asbestos lined brake shoes, automatic slack adjusters, and cold-rolled 28-tooth spline camshafts. Rockwell genuine replacement parts are reliable and offer long-lasting quality. Other original OEM aftermarket brake



products include major and minor overhaul kits, unlined brake shoes, manual slack adjusters, a variety of s-cam shafts, and air dryers. (Photo courtesy of Rockwell Automotive.)

## Axial-Type Clutches and Brakes

In an axial clutch the mating frictional members are moved in a direction parallel to the shaft. One of the earliest axial clutches was the cone clutch, which is simple in construction and, yet, quite powerful. Except for relatively simple installations, however, it has been largely replaced by the disk clutch, which employs one or more disks as the operating members. Advantages of the disk clutch include (1) no centrifugal effects, (2) a large frictional area that can be installed in a small space, (3) more effective heat dissipation surfaces, and (4) a favorable pressure distribution. There are two methods in general use to obtain the axial force necessary to produce a certain torque and pressure (depending upon the construction of the clutch). The two methods are (1) uniform wear, and (2) uniform pressure. If the disks are rigid then the greatest amount of wear will first occur in the outer areas, since the work of friction is greater in those areas. After a certain amount of wear has taken place, the pressure distribution will change so as to permit the wear to be uniform. The greatest pressure must occur at the inside diameter of the disk in order for the wear to be uniform. The second method of construction employs springs to obtain a uniform pressure over the area.

## Disk Clutches and Brakes

There is no fundamental difference between a disk clutch and a disk brake [Gagne, 1953]. The disk brake has no self-energization and, hence, is not as susceptible to changes in the coefficient of friction. The axial force can be written as

$$F_a = 0.5\pi p D_1 (D_2 - D_1) \quad (22.13)$$

where  $p$  is the maximum pressure, and  $D_1$  and  $D_2$  are the inner and outer diameters of the disk, respectively. The torque transmitted can be obtained from the relation

$$T = 0.5\mu F_a D_m \quad (22.14)$$

where  $\mu$  is the coefficient of friction of the clutch material, and the mean diameter

$$D_m = 0.5(D_2 + D_1) \quad \text{or} \quad D_m = \frac{2(D_2^3 - D_1^3)}{3(D_2^2 - D_1^2)} \quad (22.15)$$

for uniform wear or for uniform pressure distribution, respectively.

A common type of disk brake is the floating caliper brake. In this design the caliper supports a single floating piston actuated by hydraulic pressure. The action is much like that of a screw

clamp, with the piston replacing the function of the screw. The floating action also compensates for wear and ensures an almost constant pressure over the area of the friction pads. The seal and boot are designed to obtain clearance by backing off from the piston when the piston is released.

## Cone Clutches and Brakes

A cone clutch consists of (1) a cup (keyed or splined to one of the shafts), (2) a cone that slides axially on the splines or keys on the mating shaft, and (3) a helical spring to hold the clutch in engagement. The clutch is disengaged by means of a fork that fits into the shifting groove on the friction cone. The axial force, in terms of the clutch dimensions, can be written as

$$F_a = \pi D_m p b \sin \alpha \quad (22.16)$$

where  $p$  is the maximum pressure,  $b$  is the face width of the cone,  $D_m$  is the mean diameter of the cone, and  $\alpha$  is one-half the cone angle in degrees. The mean diameter can be approximated as  $0.5(D_2 + D_1)$ . The torque transmitted through friction can be obtained from the relation

$$T = \frac{\mu F_a D_m}{2 \sin \alpha} \quad (22.17)$$

The cone angle, the face width of the cone, and the mean diameter of the cone are the important geometric design parameters. If the cone angle is too small, say, less than about  $8^\circ$ , the force required to disengage the clutch may be quite large. The wedging effect lessens rapidly when larger cone angles are used. Depending upon the characteristics of the friction materials, a good compromise can usually be found using cone angles between  $10^\circ$  and  $15^\circ$ . For clutches faced with asbestos, leather, or a cork insert, a cone angle of  $12.5^\circ$  is recommended.

## Positive-Contact Clutches

A positive-contact clutch does not slip, does not generate heat, cannot be engaged at high speeds, sometimes cannot be engaged when both shafts are at rest, and, when engaged at any speed, is accompanied by shock. The greatest differences among the various types of positive-contact clutches are concerned with the design of the jaws. To provide a longer period of time for shift action during engagement, the jaws may be ratchet shaped, spiral shaped, or gear-tooth shaped. The square-jaw clutch is another common form of a positive-contact clutch. Sometimes a great many teeth or jaws are used, and they may be cut either circumferentially, so that they engage by cylindrical mating or on the faces of the mating elements. Positive-contact clutches are not used to the same extent as the frictional-contact clutches.

## Defining Terms

**Snug-tight condition:** The tightness attained by a few impacts of an impact wrench, or the full effort of a person using an ordinary wrench.

**Turn-of-the-nut method:** The fractional number of turns necessary to develop the required preload from the snug-tight condition.

**Self-energizing:** A state in which friction is used to reduce the necessary actuating force. The design should make good use of the frictional material because the pressure is an allowable maximum at all points of contact.

**Self-locking:** When the friction moment assists in applying the brake shoe, the brake will be self-locking if the friction moment exceeds the normal moment. The designer must select the dimensions of the clutch, or the brake, to ensure that self-locking will not occur unless it is specifically desired.

**Fail-safe and dead-man:** These two terms are often encountered in studying the operation of clutches and brakes. Fail-safe means that the operating mechanism has been designed such that, if any element should fail to perform its function, an accident will not occur in the machine or befall the operator. Dead-man, a term from the railroad industry, refers to the control mechanism that causes the engine to come to a stop if the operator should suffer a blackout or die at the controls.

## References

- Beach, K. 1962. Try these formulas for centrifugal clutch design. *Product Eng.* 33(14): 56–57.
- Blake, A. 1986. *What Every Engineer Should Know about Threaded Fasteners: Materials and Design*, p. 202. Marcel Dekker, New York.
- Blake, J. C. and Kurtz, H. J. 1965. The uncertainties of measuring fastener preload. *Machine Design*. 37(23): 128–131.
- Gagne, A. F., Jr. 1953. Torque capacity and design of cone and disk clutches. *Product Eng.* 24(12): 182–187.
- Ito, Y., Toyoda, J., and Nagata, S. 1977. Interface pressure distribution in a bolt-flange assembly. *Trans. ASME*. Paper No. 77-WA/DE-11, 1977.
- Juvinal, R. C. 1983. *Fundamentals of Machine Component Design*, p. 761. John Wiley & Sons, New York.
- Little, R. E. 1967. Bolted joints: How much give? *Machine Design*. 39(26): 173–175.
- Marks, L. S. 1987. *Marks' Standard Handbook for Mechanical Engineers*, 9th ed. McGraw-Hill, New York.
- Proctor, J. 1961. Selecting clutches for mechanical drives. *Product Eng.* 32(25): 43–58.
- Remling, J. 1983. *Brakes*, 2nd ed., p. 328. John Wiley & Sons, New York.
- Shigley, J. E. and Mischke, C. R. 1986. *Standard Handbook of Machine Design*. McGraw-Hill, New York.
- Shigley, J. E. and Mischke, C. R. 1989. *Mechanical Engineering Design*, 5th ed., p. 779. McGraw-Hill, New York.
- Spotts, M. F. 1985. *Design of Machine Elements*, 6th ed., p. 730. Prentice Hall, Englewood Cliffs, NJ.

## Further Information

- ASME Publications Catalog. 1985. *Codes and Standards: Fasteners*. American Society of Mechanical Engineers, New York.
- Bickford, J. H. 1981. *An Introduction to the Design and Behavior of Bolted Joints*, p. 443. Marcel Dekker, New York.
- Burr, A. H. 1981. *Mechanical Analysis and Design*, p. 640. Elsevier Science, New York.
- Crouse, W. H. 1971. *Automotive Chassis and Body*, 4th. ed., pp. 262–299. McGraw-Hill, New York.
- Fazekas, G. A. 1972. On circular spot brakes. *Journal of Engineering for Industry, Transactions of ASME*, vol. 94, series B, no. 3, August 1972, pp. 859–863.
- Ferodo, Ltd. 1968. *Friction Materials for Engineers*. Chapel-en-le-Frith, England.
- Fisher, J. W. and Struik, J. H. A. 1974. *Guide to Design Criteria for Bolted and Riveted Joints*, p. 314. John Wiley & Sons, New York.
- ISO Metric Screw Threads*. 1981. Specifications BS 3643: Part 2, p. 10. British Standards Institute, London.
- Lingaiah, K. 1994. *Machine Design Data Handbook*. McGraw-Hill, New York.
- Matthews, G. P. 1964. *Art and Science of Braking Heavy Duty Vehicles*. Special Publication SP-251, Society of Automotive Engineers, Warrendale, PA.
- Motosh, N. 1976. Determination of joint stiffness in bolted connections. *Journal of Engineering for Industry, Transactions of ASME*, vol. 98, series B, no. 3, August 1976, pp. 858–861.
- Neale, M. J. (ed.), 1973. *Tribology Handbook*. John Wiley & Sons, New York.
- Osgood, C. C. 1979. Saving weight in bolted joints. *Machine Design*, vol. 51, no. 24, 25 October 1979, pp. 128–133.
- Rodkey, E. 1977. Making fastened joints reliable—ways to keep 'em tight. *Assembly Engineering*, March 1977, pp. 24–27.
- Screw Threads*. 1974. ANSI Specification B1.1-1974, p. 80. American Society of Mechanical Engineers, New York.
- Viglione, J. 1965. Nut design factors for long bolt life. *Machine Design*, vol. 37, no. 18, 5 August 1965, pp. 137–141.
- Wong, J. Y. 1993. *Theory of Ground Vehicles*, 2nd ed., p. 435. John Wiley & Sons, New York.

## Dedication

This article is dedicated to the late Professor Joseph Edward Shigley who authored and coauthored several outstanding books on engineering design. The Standard Handbook of Machine Design and the Mechanical Engineering Design text (both with C. R. Mischke, see the references above) are widely used and strongly influenced the direction of this article.

Subramanyan, P. K. "Crankshaft Journal Bearings"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Crankshaft Journal Bearings

---

- 23.1 Role of the Journal Bearings in the Internal Combustion Engine
- 23.2 Construction of Modern Journal Bearings
- 23.3 The Function of the Different Material Layers in Crankshaft Journal Bearings
- 23.4 The Bearing Materials
- 23.5 Basics of Hydrodynamic Journal Bearing Theory
  - Load-Carrying Ability
- 23.6 The Bearing Assembly
  - Housing • The Bearing Crush • Other Factors Affecting Bearing Assembly
- 23.7 The Design Aspects of Journal Bearings
- 23.8 Derivations of the Reynolds and Harrison Equations for Oil Film Pressure

### **P. K. Subramanyan**

*Glacier Clevite Heavywall Bearings*

In modern internal combustion engines, there are two kinds of bearings in the category of crankshaft journal bearings—namely, the main bearings and the connecting rod bearings. Basically, these are wraparound, semicylindrical shell bearings. Two of them make up a set and, depending on the position in the assembly, one is called the upper and the other the lower bearing. They are of equal sizes. The main bearings support the crankshaft of the engine and the forces transmitted to the crankshaft from the cylinders. The connecting rod bearings (or, simply, rod bearings) are instrumental in transferring the forces from the cylinders of the internal combustion engine to the crankshaft. These connecting rod bearings are also called big end bearings or crank pin bearings. Supporting the crankshaft and transferring the pressure-volume work from the cylinders to the pure rotational mechanical energy of the crankshaft are accomplished elegantly with minimal energy loss by shearing a suitable lubricating medium between the bearings and the journals. The segment of the crankshaft within the bounds of a set of bearings, whether main bearings or rod bearings, is called the journal. Consequently, these bearings are called journal bearings.

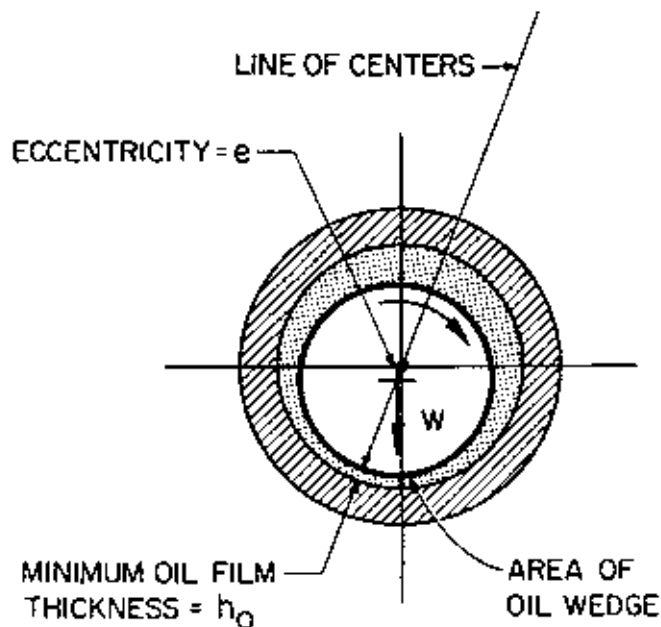
### **23.1 Role of the Journal Bearings in the Internal Combustion Engine**

---

The crankshafts of internal combustion engines of sizes from small automotive to large slow-speed engines run at widely varying rpm (e.g., 72 to 7700). When the internal combustion engine

continues to run after the start-up, the crankshaft, including the crank pins, is suspended in the lubricating oil—a fluid of very low friction. In such a condition, it is conceivable that precision-machined, semicylindrical steel shells can function as good bearings. However, there are stressful conditions, particularly in the case of automotive, truck, and medium-speed engines, when the crankshaft remains in contact with the bearings and there is little or no lubricating oil present. This condition corresponds to the initial and subsequent start-ups. The oil pump is driven directly by the engine and it takes several revolutions of the crankshaft before a good oil film is developed, as shown in Fig. 23.1, so that the journals are completely lifted and suspended. During the revolutions prior to the formation of a sufficiently thick oil film, the journal contacts the bearing surface. In such situations, the bearings provide sufficient lubrication to avoid scuffing and **seizure**. Another stressful situation, but not as critical as the start-up, is the slowing down and shutting off of the engine when the oil film reduces to a **boundary layer**.

**Figure 23.1** Schematic representation of the hydrodynamic lubricant film around a rotating journal in its bearing assembly. (Source: Slaymaker, R. R. 1955. *Bearing Lubrication Analysis*. John Wiley & Sons, New York. With permission.)



In the case of slow-speed engines, the oil pump, which is electrically driven, is turned on to prelubricate the bearings. This provides some lubrication. Nonetheless, bearings with liners and overlays are used to avoid seizure, which can result in costly damage.

Essentially, the function of journal bearings can be stated as follows: Development of the **hydrodynamic lubricating oil films** in the journal bearings lifts the journals from the surfaces of the bearings and suspends the entire crankshaft on the oil films by the journals. [Theoretical aspects of this will be considered later.] The lifting of the crankshaft or, equivalently, lifting of the journals is in the range of 30 to 1000 micro-inch in the entire range of IC engines. This process

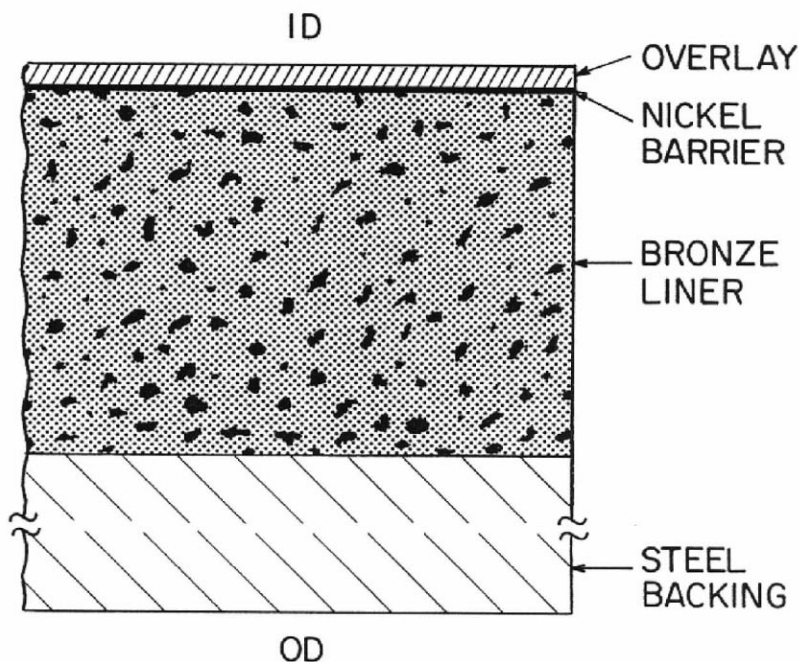
allows the crankshaft to rotate with minimal energy loss. The journal bearings make it possible so that the internal combustion engine can be started, utilized, and stopped as many times as needed.

## 23.2 Construction of Modern Journal Bearings

The majority of modern crankshaft journal bearings have three different layers of metallic materials with distinct characteristics and functions. Conventionally, these are called trimetal bearings. The remaining bearings belong to the class of bimetal bearings and have two different metallic material layers. Bimetallic bearings are becoming very popular in the automotive industry.

All crankshaft journal bearings have a steel backing, normally of low-carbon steels. Steel backing is the thickest layer in the bearing. The next layer bonded to the steel backing is the bearing liner. This is the layer that supports the load and determines the life of the bearing. The third layer bonded to the bearing liner is the overlay. Generally, this is a precision electrodeposited layer of (1) lead, tin, and copper, (2) lead and tin, or (3) lead and indium. A very thin electrodeposited layer of nickel (0.000 05 in.) is used as a bonding layer between the liner and the lead-tin-copper overlay. This nickel layer is considered a part of the overlay, not a separate layer. Construction of a trimetal bronze bearing is illustrated in [Fig. 23.2](#).

**Figure 23.2** Schematic representation of the construction of a trimetal bearing.





There are two classes of bearing liners in widespread use nowadays. These are the leaded bronzes and aluminum-based (frequently precipitation-strengthened) materials, such as aluminum-tin and aluminum-silicon. Bimetallic bearings have the advantage of being slightly more precise (about 0.0002 to 0.0003 in.) than the trimetal bearings. The bimetal bearings have a bored or broached internal diametral (ID) surface. The electrodeposited layer in the trimetal bearings is applied onto the bored or broached surface. The nickel bonding layer is applied first onto the liner, followed by the deposition of the lead-tin-copper overlay. The electrodeposited overlay introduces a certain degree of variation in the wall thickness of the bearings. In a limited application, babbitt overlays are centrifugally cast on bronze liners for slow-speed diesel engine journal bearings.

Another class of bearings is the single layer solid metal bearings—namely, solid bronze and solid aluminum bearings. These bearings are not generally used as crankshaft journal bearings. However, solid aluminum is used in some of the medium-speed and slow-speed diesel engines.

The most popular copper-tin-based leaded bearing liner in current use has 2 to 4% tin, 23 to 27% lead, and 69 to 75% copper (all by weight). This material is applied directly on mild steel by casting or sintering. The aluminum materials are roll-bonded to steel. The material as such is produced by powder rolling as a strip or by casting and rolling.

## **23.3 The Function of the Different Material Layers in Crankshaft Journal Bearings**

---

The bulk of modern crankshaft journal bearings is mild steel (1008 to 1026 low-carbon steels). This is the strongest of the two or three layers in the bearing. It supports the bearing liner, with or without the overlay. The bearing liner derives a certain degree of strength from the steel backing. The function of the steel backing is to carry the bearing liner, which on its own is weaker, much thinner, and less ductile. With the support of the steel backing, the bearings can be seated with a high degree of conformance and good interference fit in the housing bore (steel against steel).

The bearing liners in automotive and truck bearings have a thickness in the range of 0.006 to 0.030 in. In the case of the medium-speed and slow-speed engines, the thickness of the liner ranges from 0.010 to 0.080 in. The liner material contains sufficient amounts of antifriction elements, such as lead and tin. Lead is the most valuable antifriction element in the current materials and is present as a separate phase in the matrix of copper-tin alloy in the leaded bronze materials. Similarly, tin is present as an insoluble second phase in the matrix of aluminum-based materials. Lead is also insoluble in the aluminum matrix. The liner materials play the most critical role in the bearings. Once the liner material is damaged significantly, the bearing is considered unfit for further use. In a trimetal bearing, when the overlay is lost due to wear or fatigue, the bronze liner will continue to support the load and provide adequate lubrication in times of stress. The friction coefficient of liner materials is designed to be low. Besides, the soft phases of lead (in bronze) and tin (in aluminum) function as sites for embedment of dirt particles.

The overlay, which by definition is the top layer of the bearing surface, is the softest layer in the bearing. Its functions are to provide lubrication to the journal in the initial start-up situations, adjust to any misalignment or out-of-roundness of the journal, and capture dirt particles by embedment. The overlay provides sufficient lubrication during the subsequent start-up and shut-down conditions also. The journal makes a comfortable running environment in the bearing assembly during the initial runs by "bedding in." As a result of this, the wear rate of the overlay is higher in the beginning. As long as the overlay is present, the phenomenon of seizure will not occur. Once the wear progresses through the overlay, the bearing liner will provide adequate lubrication during start-up and shut-down conditions. However, if the oil supply is severely compromised or cut off for more than several seconds to a minute or so, seizure can take place once the overlay is gone, depending on the nature of the bearing liner and the load.

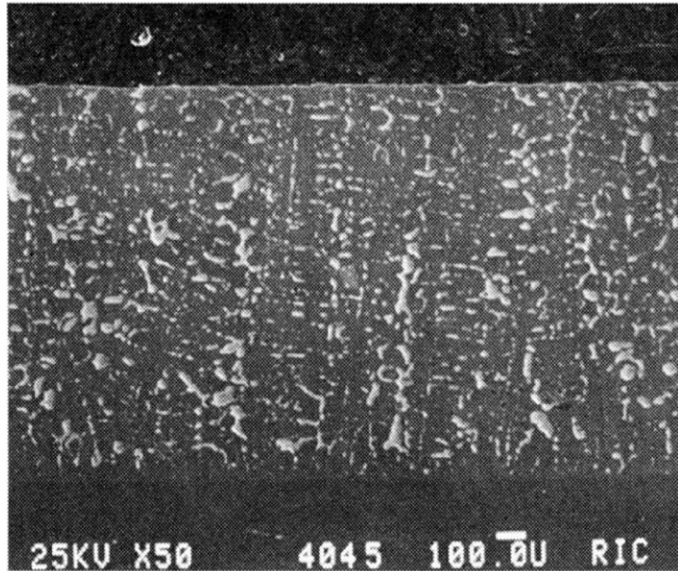
## 23.4 The Bearing Materials

---

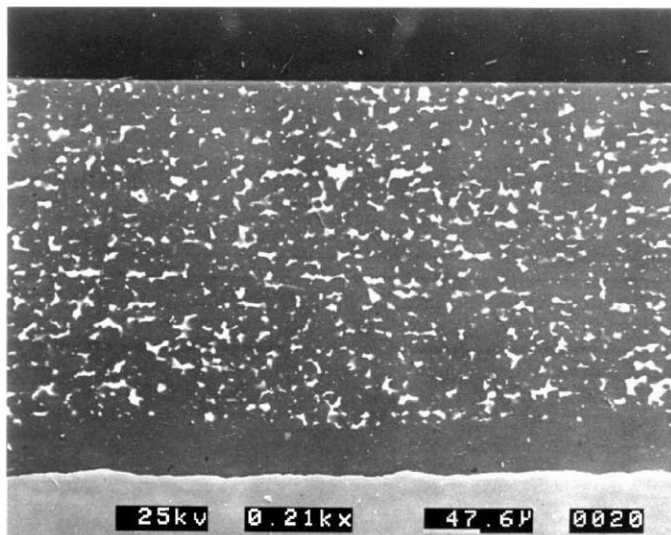
All modern crankshaft journal bearing materials are mainly composed of five elements—namely, copper, aluminum, lead, tin, and silicon. These elements account for the leaded bronze and aluminum-tin, aluminum-lead, and aluminum-silicon materials. Indium is used as a constituent of the overlays. Antimony is used in babbitts. Silver is a bearing material with good tribological properties, but it is too expensive to use as a bearing liner in journal bearings. However, it is used in special applications in some locomotive engines. An important characteristic of a good bearing material is its ability to conduct heat. Silver, copper, and aluminum are, indeed, good conductors of heat. Silver has no affinity for iron, cobalt, and nickel [Bhushan and Gupta, 1991]. Therefore, it is expected to run very well against steel shafts. Both copper and aluminum possess a certain degree of affinity for iron. Therefore, steel journals can bond to these metals in the absence of antifriction elements, such as lead and tin, or lubricating oil. Aluminum spontaneously forms an oxide layer, which is very inert, in the presence of air or water vapor. This suppresses the seizure or the bonding tendency of aluminum. Besides, the silicon particles present in the aluminum-silicon materials keep the journals polished to reduce friction.

The microstructure of the most widely used cast leaded bronze bearing liner is shown in Fig. 23.3. This has a composition of 2 to 4% tin, 23 to 27% lead, and 69 to 75% copper. Another material in widespread use, especially in automotive applications, is aluminum with 20% tin. A typical microstructure of this material is shown in Fig. 23.4. It can be used as the liner for both bimetal and trimetal bearings. The copper-tin-lead material shown in Fig. 23.3 is mainly used in trimetal bearings.

**Figure 23.3** SEM photomicrograph of a typical cross section of the cast leaded bronze diesel locomotive engine bearing material manufactured by Glacier Clevite Heavywall Bearings. The nominal composition is 3% tin, 25% lead, and 72% copper. The light gray, irregular spots represent lead in a matrix of copper-tin. This material is bonded to mild steel at the bottom. (Magnification 50 $\times$ .)]



**Figure 23.4** SEM photomicrograph of a typical cross section of aluminum-tin material roll bonded to mild steel, manufactured by Glacier Vandervell Ltd. The nominal composition is 20% tin, 1% copper, and 79% aluminum. The light gray, irregular spots represent tin in the aluminum-copper matrix. Below the aluminum-tin layer is a layer of pure aluminum which functions as a bonding layer to the mild steel underneath. (Magnification 210 $\times$ .)

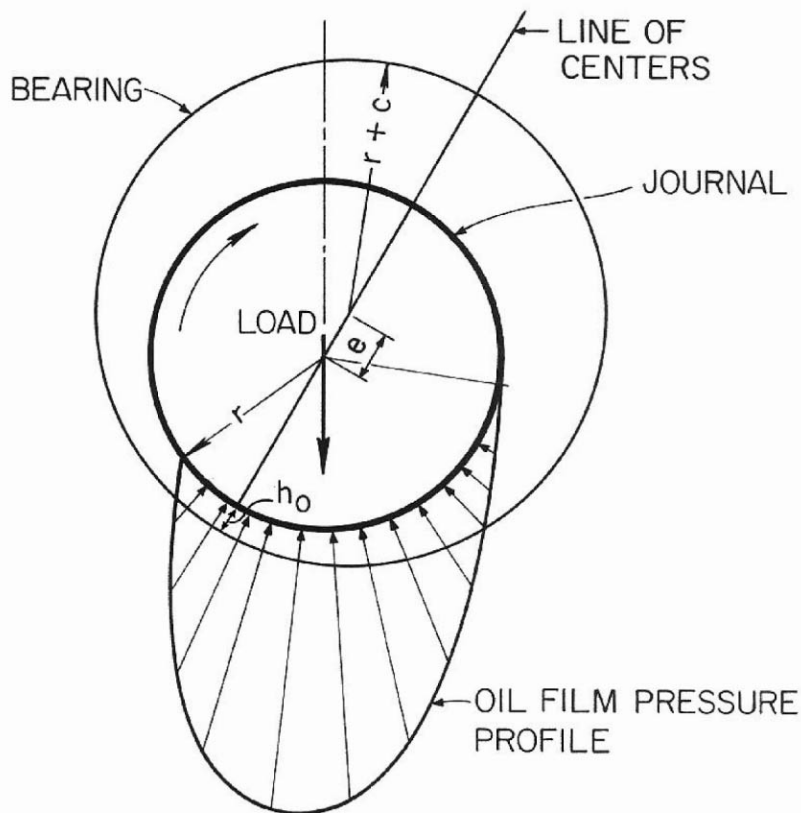


## 23.5 Basics of Hydrodynamic Journal Bearing Theory

### Load-Carrying Ability

As mentioned previously, when running in good condition, the journal which was initially lying on the surface of the bearing is lifted and surrounded by the lubricant. It becomes suspended in the surrounding film of lubricating oil. If the engine keeps running, the journal will remain in its state of suspension indefinitely. The inertial load of the crankshaft and the forces transmitted from the cylinders to the crankshaft are supported by the lubricant films surrounding the main bearing journals. The oil film surrounding the rod bearing journal supports the gas forces developed in the cylinder and the inertial load of the piston and connecting rod assembly. Around each journal, a segment of the oil film develops a positive pressure to support the load, as shown in Fig. 23.5. In the following brief theoretical consideration, the process that develops this load-carrying positive pressure will be illustrated.

**Figure 23.5** Schematic representation of the profile of the load supporting pressure in the oil film. (Source: Slaymaker, R. R. 1955. *Bearing Lubrication Analysis*. John Wiley & Sons, New York. By permission.)



As a background to the theoretical considerations, the following assumptions are made. The flow of the lubricating oil around the journal at all speeds is assumed to be laminar. The length of the bearing  $L$  is assumed to be infinite, or the flow of the lubricant from the edges of the bearing is negligible. The lubricant is assumed to be incompressible.

Consider a very small volume element of the lubricant moving in the direction of rotation of the journal—in this case, the  $x$  direction. The forces that act on this elemental volume and stabilize it are shown in Fig. 23.6. Here,  $P$  is the pressure in the oil film at a distance  $x$ . It is independent of the thickness of the oil film or the  $y$  dimension.  $S$  is the shear stress in the oil film at a distance  $y$  above the bearing surface, which is at  $y = 0$ . The length  $L$  of the bearing is in the  $z$  direction. The equilibrium condition of this volume element gives us the following relationship [Slaymaker, 1955; Fuller, 1984]:

$$\left[ P + \left( \frac{dP}{dx} \right) dx \right] dy dz + S dx dz - \left[ S + \left( \frac{dS}{dy} \right) dy \right] dx dz - P dy dz = 0 \quad (23.1)$$

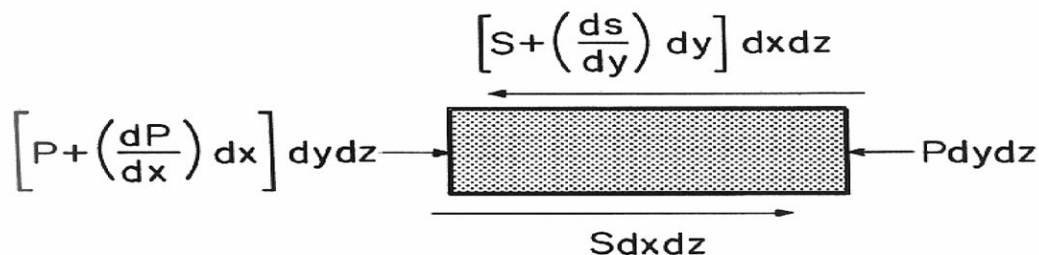
Therefore,

$$\left( \frac{dS}{dy} \right) = \left( \frac{dP}{dx} \right) \quad (23.2)$$

Equation (23.2) represents a very important, fundamental relationship. It clearly shows how the load-carrying pressure  $P$  is developed. It is the rate of change of the shear stress in the direction of the oil film thickness that generates the hydrostatic pressure  $P$ . As we shall see from Eq. (23.3), the shear stress is directly proportional to the shearing rate of the oil film ( $dv/dy$ )—as ( $dv/dy$ ) increases, ( $dS/dy$ ) must increase. Since the thickness of the oil film decreases in the direction of rotation of the journal, a progressive increase in the shearing rate of the oil film automatically occurs because the same flow rate of oil must be maintained through diminishing cross sections (i.e., decreasing  $y$  dimension). This progressive increase in the shearing rate is capable of generating very high positive hydrostatic pressures to support very high loads. A profile of the pressure generated in the load-supporting segment of the oil film is shown in Fig. 23.5. By introducing the definition of the coefficient of viscosity, we can relate the shear stress to a more measurable parameter, such as the velocity,  $v$ , of the lubricant, as

$$S = \mu \left( \frac{dv}{dy} \right) \quad (23.3)$$

**Figure 23.6** Schematic representation of the forces acting on a tiny volume element in the hydrodynamic lubricant film around a rotating journal.



Substituting for  $(dS/dy)$  from Eq. (23.3) in Eq. (23.2), we obtain a second order partial differential equation in  $v$ . This is integrated to give the velocity profile as a function of  $y$ . This is then integrated to give  $Q$ , the total quantity of the lubricant flow per unit time. Applying certain boundary conditions, one can deduce the well-known Reynolds equation for the oil film pressure:

$$\left( \frac{dP}{dx} \right) = \frac{6\mu V}{h^3} (h - h_1) \quad (23.4)$$

where  $h$  is the oil film thickness,  $h_1$  is the oil film thickness at the line of maximum oil film pressure, and  $V$  is the peripheral velocity of the journal. The variable  $x$  in the above equation can be substituted in terms of the angle of rotation  $\theta$  and then integrated to obtain the Harrison equation for the oil film pressure. With reference to the diagram in Fig. 23.7, the thickness of the oil film can be expressed as

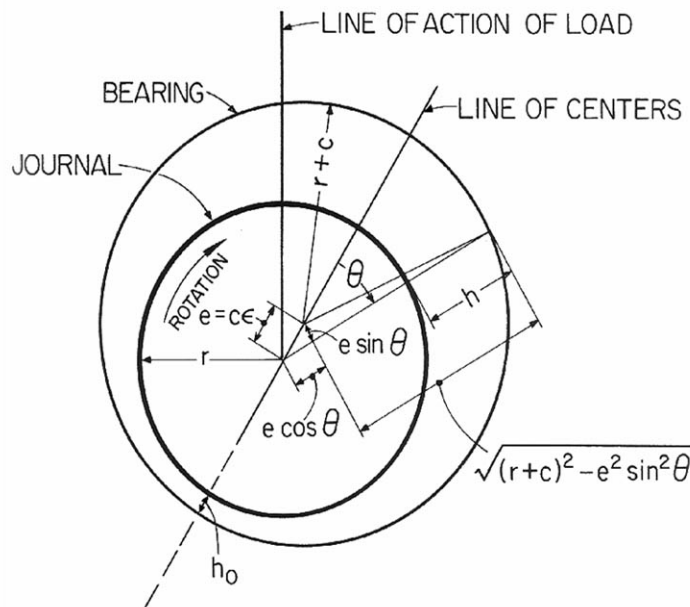
$$h = c(1 + \varepsilon \cos \theta) \quad (23.5)$$

where  $c$  is the radial clearance and  $\varepsilon$  is the eccentricity ratio. The penultimate form of the Harrison equation can be expressed as

$$\int_0^{2\pi} dP = \int_0^{2\pi} \frac{6\mu V r \varepsilon}{c^2} \left[ \frac{\cos \theta - \cos \theta_1}{(1 + \varepsilon \cos \theta)^3} \right] d\theta = P - P_0 \quad (23.6)$$

where  $P_0$  is the pressure of the lubricant at  $\theta = 0$  in Fig. 23.7, and  $\theta_1$  is the angle at which the oil film pressure is a maximum. Brief derivations of the Reynolds equation and the Harrison equation are given in section 23.8.

**Figure 23.7** Illustration of the geometric relationship of a journal rotating in its bearing assembly.  
(Source: Slaymaker, R. R. 1955. Bearing Lubrication Analysis. John Wiley and Sons, New York. By permission.)





For practical purposes, it is more convenient to carry out the integration of Eq. (23.6) numerically rather than using Eq. (23.14) in section 23.8. This is done with good accuracy using special computer programs. The equations presented above assume that the end leakage of the lubricating oil is equal to zero. In all practical cases, there will be end leakage and, hence, the oil film will not develop the maximum possible pressure profile. Therefore, its load-carrying capability will be diminished. The flow of the lubricant in the  $z$  direction needs to be taken into account. However, the Reynolds equation for this case has no general solution [Fuller, 1984]. Hence, a correction factor between zero and one is applied, depending on the length and diameter of the bearing (L/D ratio) and the eccentricity ratio of the bearing. Indeed, there are tabulated values available for the side leakage factors for bearings with various L/D ratios and eccentricity ratios [Fuller, 1984]. Some of these values are given in Table 23.1.

**Table 23.1** Side Leakage Correction Factors for Journal Bearings

L/D Ratio	Eccentricity Ratio						
	0.80	0.90	0.92	0.94	0.96	0.98	0.99
0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	—	0.867	0.88	0.905	0.937	0.97	0.99
1	0.605	0.72	0.745	0.79	0.843	0.91	0.958
0.5	0.33	0.50	0.56	0.635	0.732	0.84	0.908
0.3	0.17	0.30	0.355	0.435	0.551	0.705	0.81
0.1	—	0.105	0.115	0.155	0.220	0.36	0.53

Booker [1965] has done considerable work in simplifying the journal center orbit calculations without loss of accuracy by introducing new concepts, such as dimensionless journal center velocity/force ratio (i.e., mobility) and maximum film pressure/specific load ratio (i.e., maximum film pressure ratio). This whole approach is called the *mobility method*. This has been developed into computer programs which are widely used in the industry to calculate film pressures and thicknesses. Further, this program calculates energy loss due to the viscous shearing of the lubricating oil. These calculations are vital for optimizing the bearing design and selecting the appropriate bearing liner with the required fatigue life. This is determined on the basis of the **peak oil film pressure** (POFP). In Booker's mobility method, the bearing assembly, including the housing, is assumed to be rigid. In reality, the bearings and housings are flexible to a certain degree, depending on the stiffness of these components. Corrections are now being made to these deviations by the elastohydrodynamic theory, which involves finite element modeling of the bearings and the housing. Also, the increase in viscosity as a function of pressure is taken into account in this calculation. The elastohydrodynamic calculations are presently done only in very special cases and have not become part of the routine bearing analysis.

## 23.6 The Bearing Assembly

---

### Housing

The housing into which a set of bearings is inserted and held in place is a precision-machined cylindrical bore with close tolerance. The surface finishes of the housing and the backs of the bearings must be compatible. Adequate contact between the backs of the bearings and the surface of the housing bore is a critical requirement to ensure good heat transfer through this interface. The finish of the housing bore is expected to be in the range of 60 to 90  $\mu\text{in.}$  ( $R_a$ ) (39.4  $\mu\text{in.}$  = 1 micron). The finish on the back of the bearings is generally set at 80  $\mu\text{in.}$  maximum. Nowadays, the finishes on the housing bore and the backs of the bearings are becoming finer. The finish at the parting line face of bearings of less than 12 in. gage size is expected to be less than 63  $\mu\text{in.}$  For larger bearings, this is set at a maximum of 80  $\mu\text{in.}$  The bearing backs may be rolled, turned, or ground. All the automotive and truck bearings have rolled steel finish at the back. The housing can be bored, honed, or ground, but care must be taken to avoid circumferential and axial banding.

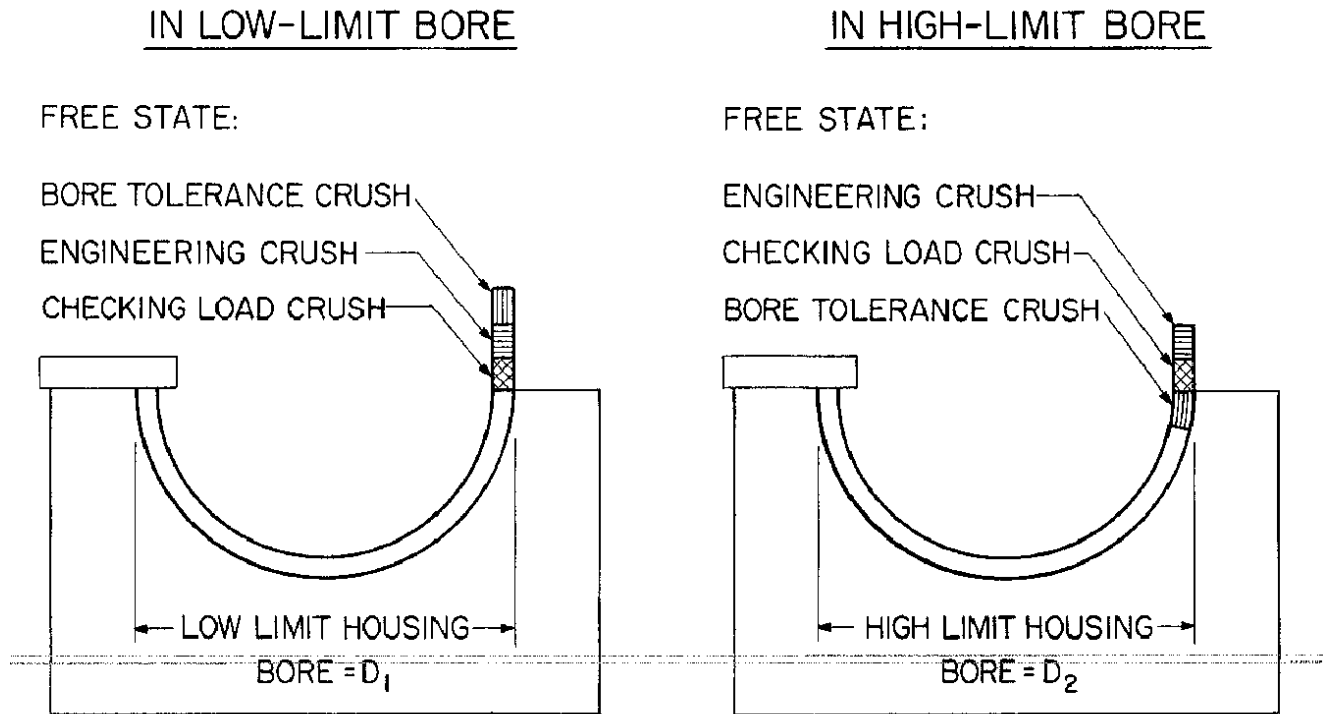
### The Bearing Crush

The term **crush** is not used in a literal sense in this context. A quantitative measure of the crush of a bearing is equal to the excess length of the exterior circumference of the bearing over half the interior circumference of the bearing housing. Effectively, this is equal to the sum of the two parting line heights. When the bearing assembly is properly torqued, the parting line height of each bearing in the set is reduced to zero. In that state, the back of the bearing makes good contact with the housing and applies a radial pressure in the range of 800 to 1200 psi (5.5 to 8.24 MPa). Thereby, a good interference fit is generated. If the bearings are taken out of the assembly, they are expected to spring back to their original state. Therefore, nothing is actually crushed.

The total crush or the parting line height of a bearing has three components—namely, the housing bore tolerance crush, the checking load crush, and the engineering crush. The housing bore tolerance crush is calculated as  $0.5\pi(D_2 - D_1)$ , where  $D_1$  and  $D_2$  are the lower and upper limits of the bore diameter, respectively. Suppose a bearing is inserted in its own inspection block (the diameter of which corresponds to the upper limit of the diameter of the bearing housing). The housing bore tolerance crush does not make a contribution to the actual crush, as shown in [Fig. 23.8](#) (high limit bore). If load is applied on its parting lines in increasing order and the values of these loads are plotted as a function of the cumulative decrease in parting line height, one may expect it to obey Hooke's law. Initially, however, it does not obey Hooke's law, but it does so thereafter. The initial nonlinear segment corresponds to the checking load crush. The checking load corresponds to the load required to conform the bearing properly in its housing. The final crush or the parting line height of the bearing is determined in consultation with the engine manufacturer.



**Figure 23.8** Schematic illustration of the components of crush of a bearing in the thinwall bearing inspection block, before application of load (i.e., in the free state). The magnitude of the crush components is exaggerated.



## Other Factors Affecting Bearing Assembly

These factors are (1) freespread, (2) bore distortion, (3) cap offset or twist, (4) misalignment of the crankshaft, (5) out-of-roundness of the journal, and (6) deviation of the bearing clearance. The outside diameter of the bearing at the parting lines must be slightly greater than the diameter of the housing bore. This is called the freespread. It helps to snap the bearings into the housing. The required degree of freespread is determined by the wall thickness and the diameter. In the case of wall thickness, the freespread is inversely proportional to it. For a wide range of bearings, the freespread is in the range of 0.025 to 0.075 in. Bearings with negative freespread are not used because, when bolted, the side of the parting lines could rub against the journal and lead to possible seizure while running. It is possible to change the freespread from negative to positive by reforming the bearing. Bore distortion, cap offset or twist, and misalignment of the crankshaft can lead to the journal making rubbing contacts with the bearing surface. The conformability of the bearings can take care of these problems to a certain degree by local wearing of the overlay in a trimetal bearing or by melting the soft phase in a bimetal bearing, which results in the two-phase structure crushing and conforming. In severe cases, the liner materials in both cases are damaged.

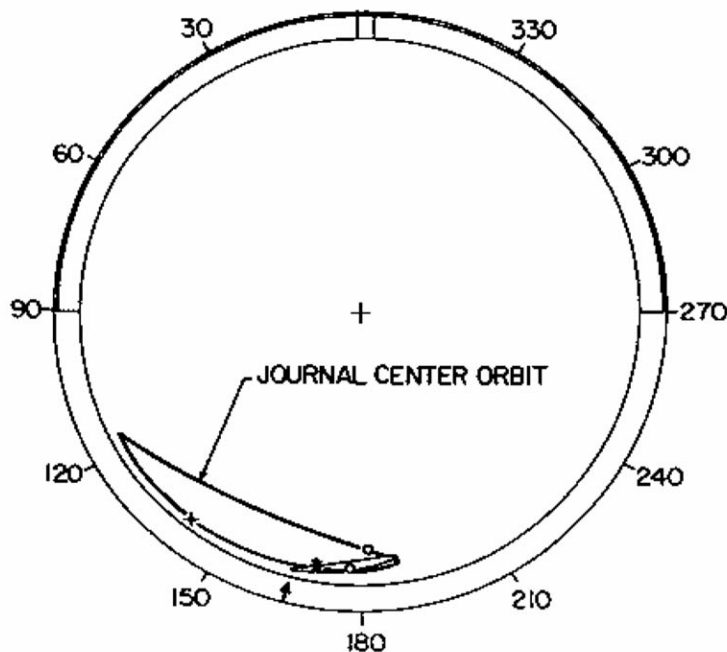
By developing high oil film pressures on the peaks of the lobes, out-of-roundness in the journal can accelerate fatigue of the bearing.

If the clearance is not adequate, the bearing will suffer from oil starvation and the temperature will rise. In extreme cases, this will lead to bearing seizure and engine damage. On the other hand, if the clearance is excessive, there will be increased noise and increased peak oil film pressure, which will bring about premature fatigue of the loaded bearing.

## 23.7 The Design Aspects of Journal Bearings

Even though the journal bearings are of simple semicylindrical shape and apparently of unimpressive features, there are important matters to be taken into account in their design. The bearing lengths, diameters, and wall thicknesses are generally provided by the engine builder or decided in consultation with the bearing manufacturer. A journal orbit study must be done to optimize the clearance space between the journal and the bearing surface. This study also provides the **minimum oil film thickness (MOFT)** and the POFP (Fig. 23.9). Values of these parameters for the optimized clearance are important factors. The MOFT is used in the calculation of the oil flow, temperature rise, and heat balance. According to Conway-Jones and Tarver [1993], about 52% of the heat generated in connecting rod bearings in automobile engines is carried away by the oil flow. Approximately 38% of the remaining heat flows into the adjacent main bearings via the crankshaft. The remaining 10% is lost by convection and radiation. In the case of main bearings, about 95% of the total heat is carried away by the oil flow, which is estimated to be more than five times the flow through the connecting rod bearings, which were fed by a single oil hole drilled in the crank pin. The POFP is the guiding factor in the selection of a bearing liner with adequate fatigue strength or fatigue life.

**Figure 23.9** Journal center orbit diagram of two-stroke cycle medium-speed (900 rpm) diesel engine main bearings (no. 1 position). The inner circle represents the clearance circle of the bearings. It also represents the bearing surface. The entire cross section of the journal is reduced to a point coinciding with the center of the journal. The upper main bearing has an oil hole at the center with a circumferential groove at the center of the bearing represented by the dark line. Maximum unit load: 1484 psi. MOFT: 151  $\mu\text{in.}$  @ 70/166. POFP: 11 212 psi @ 55/171. Oil: SAE 30W. Cylinder pressure data given by the manufacturer of the engine. Clockwise rotation. The journal orbit analysis done at Glacier Clevite Heavywall Bearings. —\*—0–180 crank angle, —+—180–360 crank angle, @ crank angle/bearing angle. Arrow indicates the location of MOFT.



The bearing must be properly located in the housing bore. This is achieved by having a notch at one end of the bearing at the parting line. There must be provisions to bring in the lubricant and remove it. Therefore, appropriate grooves and holes are required. The best groove to distribute the

lubricant is a circumferential groove with rounded edges, centrally placed in both bearings. If this is a square groove, the flow will be diminished by 10%. If these grooves are in the axial direction, the oil flow is decreased by 60% with respect to the circumferential ones. Having a circumferential groove in the loaded half of the bearings does increase the POFP. In the case of large slow-speed diesel engines, the POFPs are generally very low compared to the pressures in automotive, truck, and medium-speed diesel engines. Therefore, central circumferential grooves are best suited for slow-speed engines.

In the automotive, truck, and medium-speed engines, the loaded halves of the bearings do not have circumferential grooves. However, the other halves have the circumferential grooves. Some of the loaded bearings have partial grooves. Otherwise, some type of oil spreader machined in the location below the parting line is desirable in the case of larger bearings. If the oil is not spread smoothly, the problems of cavitation and erosion may show up. The end of the partial groove or the oil spreader must be blended.

The edges of all the bearings must be rounded or chamfered to minimize the loss of the lubricant. Edges are also chamfered to eliminate burrs. A sharp edge acts as an oil scraper and thereby enhances oil flow in the axial direction along the edges, which is harmful. Finally, bearings have a small relief just below the parting lines along the length on the inside surface. This is meant to protect the bearings in case of slight misalignment or offset at the parting lines.

## 23.8 Derivations of the Reynolds and Harrison Equations for Oil Film Pressure

The background for deriving these equations is given in section 23.5 of the text. The equilibrium condition of a tiny volume element of the lubricating oil (Fig. 23.6) is represented by the following equation [Slaymaker, 1955; Fuller, 1984]:

$$\left[ P + \left( \frac{dP}{dx} \right) dx \right] dy dz + S dx dz - \left[ S + \left( \frac{dS}{dy} \right) dy \right] dx dz - P dy dz = 0$$

(23.7)

Therefore,

$$\left( \frac{dS}{dy} \right) = \left( \frac{dP}{dx} \right) \quad (23.8)$$

Now, by introducing the definition of the coefficient of viscosity  $\mu$ , we can relate the shear stress to a more measurable parameter, like the velocity  $v$  of the lubricant, as

$$S = \mu \left( \frac{dv}{dy} \right) \quad (23.9)$$

Substituting for  $(dS/dy)$  from Eq. (23.9) in Eq. (23.8), a second order partial differential equation in  $v$  is obtained. This is integrated to give an expression for the velocity profile as

$$v = \frac{V}{h}y - \frac{1}{2\mu} \left( \frac{dP}{dx} \right) (hy - y^2) \quad (23.10)$$

In Eq. (23.10),  $V$  is the peripheral velocity of the journal and  $h$  is the oil film thickness. The boundary conditions used to derive Eq. (23.10) are (1)  $v = V$  when  $y = h$ , and (2)  $v = 0$  when  $y = 0$  (at the surface of the bearing). Now applying the relationship of continuity, the oil flowing past any cross section in the  $z$  direction of the oil film around the journal must be equal. The quantity  $Q$  of oil flow per second is given by

$$Q = L \int_0^h v \, dy \quad (23.11)$$

where  $L$  is the length of the bearing which is in the  $z$  direction. Now substituting for  $v$  from Eq. (23.10) in Eq. (23.11) and integrating,

$$Q = L \left[ \frac{Vh}{2} - \frac{h^3}{12\mu} \left( \frac{dP}{dx} \right) \right] \quad (23.12)$$

The pressure  $P$  varies as a function of  $x$  in the oil film, which is in the direction of rotation of the journal. At some point, it is expected to reach a maximum. At that point,  $(dP/dx)$  becomes zero. Let  $h_1$  represent the oil film thickness at that point. Therefore,

$$Q = \frac{LV}{2} h_1 \quad (23.13)$$

Now we can use Eq. (23.13) to eliminate  $Q$  from Eq. (23.12). Hence,

$$\left( \frac{dP}{dx} \right) = \frac{6\mu V}{h^3} (h - h_1) \quad (23.14)$$

Equation (23.14) is the Reynolds equation for the oil film pressure as a function of distance in the direction of rotation of the journal. The variable  $x$  in Eq. (23.14) can be substituted in terms of the angle of rotation  $\theta$  and then integrated to obtain the Harrison equation for the oil film pressure. With reference to the diagram in Fig. 23.7, the oil film thickness  $h$  can be expressed as

$$h = e \cos \theta + \sqrt{(r + c)^2 - e^2 \sin^2 \theta} - r \quad (23.15)$$

Here,  $e$  is the eccentricity,  $c$  is the radial clearance, and  $e = c\varepsilon$ , where  $\varepsilon$  is the eccentricity ratio. The quantity  $e^2 \sin^2 \theta$  is much smaller compared to  $(r + c)^2$ . Therefore,

$$h = c(1 + \varepsilon \cos \theta) \quad (23.16)$$

Now,  $(dP/dx)$  is converted into polar coordinates by substituting  $rd\theta$  for  $dx$ . Therefore, Eq. (23.14) can be expressed as

$$\left( \frac{dP}{d\theta} \right) = \frac{6\mu V r \varepsilon}{c^2} \left[ \frac{\cos \theta - \cos \theta_1}{(1 + \varepsilon \cos \theta)^3} \right] \quad (23.17)$$

where  $\theta_1$  is the angle at which the oil film pressure is a maximum. Integration of Eq. (23.17) from  $\theta = 0$  to  $\theta = 2\pi$  can be expressed as

$$\int_0^{2\pi} dP = \int_0^{2\pi} \frac{6\mu V r \varepsilon}{c^2} \left[ \frac{\cos \theta - \cos \theta_1}{(1 + \varepsilon \cos \theta)^3} \right] d\theta = P - P_0 \quad (23.18)$$

where  $P_0$  is the pressure of the lubricant at the line of centers ( $\theta = 0$ ) in Fig. 23.7. If  $(P - P_0)$  is assumed to be equal to zero at  $\theta = 0$  and  $\theta = 2\pi$ , the value of  $\cos \theta_1$ , upon integration of Eq. (23.18), is given by

$$\cos \theta_1 = -\frac{3\varepsilon}{2 + \varepsilon^2} \quad (23.19)$$

and the Harrison equation for the oil film pressure for a full journal bearing by

$$P - P_0 = \frac{6\mu V r \varepsilon}{c^2} \frac{\sin \theta (2 + \varepsilon \cos \theta)}{(2 + \varepsilon^2)(1 + \varepsilon \cos \theta)^2} \quad (23.20)$$

## Acknowledgment

The author wishes to express his thanks to David Norris, President of Glacier Clevite Heavywall Bearings, for his support and interest in this article, and to Dr. J. M. Conway-Jones (Glacier Metal Company, Ltd., London), George Kingsbury (Consultant, Glacier Vandervell, Inc.), Charles Latreille (Glacier Vandervell, Inc.), and Maureen Hollander (Glacier Vandervell, Inc.) for reviewing this manuscript and offering helpful suggestions.

## Defining Terms

**Boundary layer lubrication:** This is a marginally lubricating condition. In this case, the surfaces of two components (e.g., one sliding past the other) are physically separated by an oil film that has a thickness equal to or less than the sum of the heights of the asperities on the surfaces. Therefore, contact at the asperities can occur while running in this mode of lubrication. This is also described as "mixed lubrication." In some cases, the contacting asperities will be polished out. In other cases, they can generate enough frictional heat to destroy the two components. Certain additives can be added to the lubricating oil to reduce asperity friction drastically.

**Crush:** This is the property of the bearing which is responsible for producing a good interference fit in the housing bore and preventing it from spinning. A quantitative measure of the crush is equal to the excess length of the exterior circumference of the bearing over half the interior circumference of the housing. This is equal to twice the parting line height, if measured in an equalized half height measurement block.

**Hydrodynamic lubrication:** In this mode of lubrication, the two surfaces sliding past each other (e.g., a journal rotating in its bearing assembly) are physically separated by a liquid lubricant of suitable viscosity. The asperities do not come into contact in this case and the friction is very low.

**Minimum oil film thickness (MOFT):** The hydrodynamic oil film around a rotating journal develops a continuously varying thickness. The thickness of the oil film goes through a

minimum. Along this line, the journal most closely approaches the bearing. The maximum wear in the bearing is expected to occur around this line. Therefore, MOFT is an important parameter in designing bearings.

**Peak oil film pressure (POFP):** The profile of pressure in the load-carrying segment of the oil film increases in the direction of rotation of the journal and goes through a maximum (Fig. 23.5). This maximum pressure is a critical parameter because it determines the fatigue life of the bearing. This is also called maximum oil film pressure (MOFP).

**Positive freespread:** This is the excess in the outside diameter of the bearing at the parting line over the inside diameter of the housing bore. As a result of this, the bearing is clipped in position in its housing upon insertion. Bearings with negative freespread will be loose and lead to faulty assembly conditions.

**Seizure:** This is a critical phenomenon brought about by the breakdown of lubrication. At the core of this phenomenon is the occurrence of metal-to-metal bonding, or welding, which can develop into disastrous levels, ultimately breaking the crankshaft. With the initiation of seizure, there will be increased generation of heat, which will accelerate this phenomenon. Galling and adhesive wear are terms which mean the same basic phenomenon. The term *scuffing* is used to describe the initial stages of seizure.

## References

- Bhushan, B. and Gupta, B. K. 1991. *Handbook of Tribology*. McGraw-Hill, New York.
- Booker, J. F. 1965. Dynamically loaded journal bearings: Mobility method of solution. *J. Basic Eng. Trans. ASME*, series D, 87:537.
- Conway-Jones, J. M. and Tarver, N. 1993. Refinement of engine bearing design techniques. *SAE Technical Paper Series, 932901, Worldwide Passenger Car Conference and Exposition*, Dearborn, MI, October 25–27.
- Fuller, D. D. 1984. *Theory and Practice of Lubrication for Engineers*, 2nd ed. John Wiley & Sons, New York.
- Slaymaker, R. R. 1955. *Bearing Lubrication Analysis*. John Wiley & Sons, New York.

## Further Information

- Yahraus, W. A. 1987. Rating sleeve bearing material fatigue life in terms of peak oil film pressure. *SAE Technical Paper Series, 871685, International Off-Highway & Powerplant Congress and Exposition*, Milwaukee, WI, September 14–17.
- Booker, J. F., 1971. Dynamically loaded journal bearings: Numerical application of the mobility method. *J. of Lubr. Technol. Trans. ASME*, 93:168.
- Booker, J. F., 1989. Squeeze film and bearing dynamics. *Handbook of Lubrication*, ed. E. R. Booser. CRC Press, Boca Raton, FL.
- Hutchings, I. M. 1992. *Tribology*. CRC Press, Boca Raton, FL.
- Transactions of the ASME, Journal of Tribology.*
- STLE Tribology Transactions.*
- Spring and Fall Technical Conferences of the ASME/ICED.

Lebeck, A. O. "Fluid Sealing in Machines, Mechanical Devices..."  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Fluid Sealing in Machines, Mechanical Devices, and Apparatus

---

## 24.1 Fundamentals of Sealing

### 24.2 Static Seals

Gaskets • Self-Energized Seals • Chemical Compound or Liquid Sealants as Gaskets

### 24.3 Dynamic Seals

Rotating or Oscillating Fixed-Clearance Seals • Rotating Surface-Guided Seals—Cylindrical Surface • Rotating Surface-Guided Seals—Annular Surface • Reciprocating Fixed-Clearance Seals • Reciprocating Surface-Guided Seals • Reciprocating Limited-Travel Seals

### 24.4 Gasket Practice

### 24.5 O-Ring Practice

### 24.6 Mechanical Face Seal Practice

## Alan O. Lebeck

*Mechanical Seal Technology, Inc.*

The passage of fluid (leakage) between the mating parts of a machine and between other mechanical elements is prevented or minimized by a fluid seal. Commonly, a gap exists between parts formed by inherent roughness or misfit of the parts—where leakage must be prevented by a seal. One may also have of necessity gaps between parts that have relative motion, but a fluid seal is still needed. The fluid to be sealed can be any liquid or gas. Given that most machines operate with fluids and must contain fluids or exclude fluids, most mechanical devices or machines require a multiplicity of seals.

Fluid seals can be categorized as *static* or *dynamic* as follows.

Static:

- Gap to be sealed is generally very small.
- Accommodates imperfect surfaces, both roughness and out-of-flatness.
- Subject to very small relative motions due to pressure and thermal cyclic loading.
- Allows for assembly/disassembly.

Dynamic:

- Gap to be sealed is much larger and exists of necessity to permit relative motion.
- Relatively large relative motions between surfaces to be sealed.
- Motion may be continuous (rotation) in one direction or large reciprocating or amount of



motion may be limited.

- Seal must not constrain motion (usually).

Although there is some crossover between static and dynamic seal types, by categorizing based on the static and dynamic classification, the distinction between the various seal types is best understood.

## 24.1 Fundamentals of Sealing

---

Sealing can be accomplished by causing the gap between two surfaces to become small but defined by the geometric relationship between the parts themselves. In this case one has a fixed-clearance seal. One may also force two materials into contact with each other, and the materials may be either sliding relative to each other or static. In this case one has a surface-guided seal where the **sealing clearance** now becomes defined by the materials themselves and the dynamics of sliding in the case of a sliding seal.

There are two broad classes of surface-guided material pairs. The first and most common involves use of an **elastomeric**, plastic, or other soft material against a hard material. In this case the soft material deforms to conform to the details of the shape of the harder surface and will usually seal off completely in the static case and nearly completely in the dynamic case. A rubber gasket on metal is an example. The second class, far less common, is where one mates a hard but wearable material to a hard material. Here the sealing gap derives from a self-lapping process plus the alignment of the faces of the material. Since both materials are relatively hard, if one material develops a roughness or grooves, the seal will leak. A mechanical face seal is an example.

## 24.2 Static Seals

---

Static seals can be categorized as follows:

### Gaskets

- Single or composite compliant material

- Metal encased

- Wrapped and spiral wound

- Solid metal

### Self-energized elastomeric rings

- Circular cross section (O-ring)

- Rectangular cross section

### Chemical compound or liquid sealants as gaskets

- Rubbers

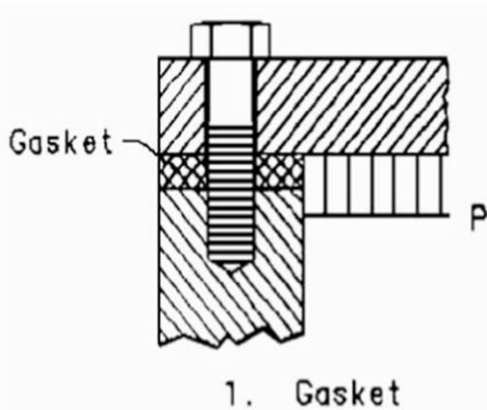
- Plastics

## Gaskets

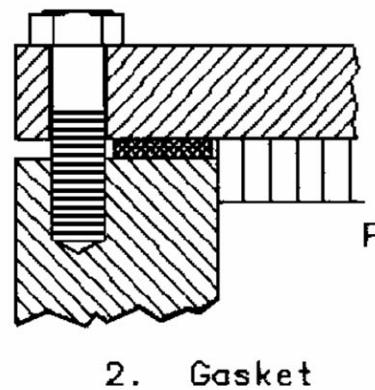
Within the category of static seals, gaskets comprise the greatest fraction. The sealing principle common to gaskets is that a material is clamped between the two surfaces being sealed. Clamping force is large enough to deform the gasket material and hold it in tight contact even when the pressure attempts to open the gap between the surfaces.

A simple single-material gasket clamped between two surfaces by bolts to prevent leakage is shown in Fig. 24.1. Using a compliant material the gasket can seal even though the sealing surfaces are not flat. As shown in Fig. 24.2, the gasket need not cover the entire face being sealed. A gasket can be trapped in a groove and loaded by a projection on the opposite surface as shown in Fig. 24.3. Composite material gaskets or metal gaskets may be contained in grooves as in Fig. 24.4. Gaskets are made in a wide variety of ways. A spiral-wound metal/fiber composite, metal or plastic clad, solid metal with sealing projections, and a solid fiber or rubber material are shown in Fig. 24.5.

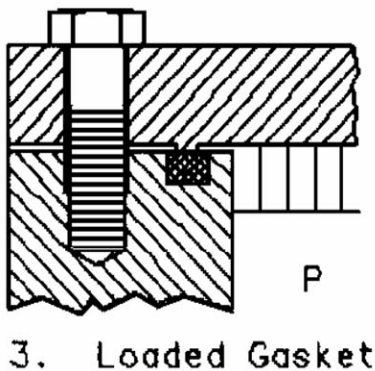
**Figure 24.1** Gasket.



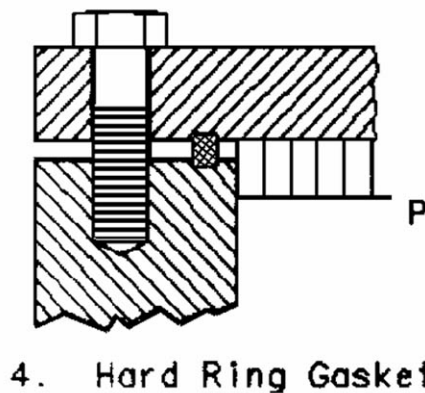
**Figure 24.2** Gasket.



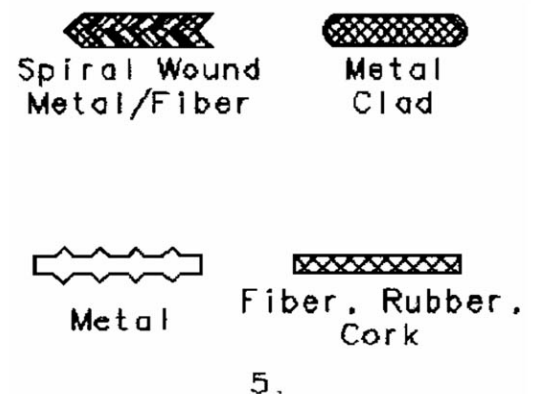
**Figure 24.3** Loaded gasket.



**Figure 24.4** Hard ring gasket.



**Figure 24.5** Varieties of gaskets.



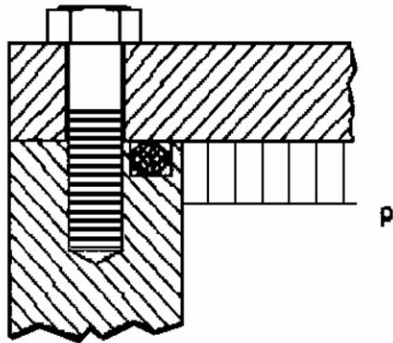
Gaskets can be made of relatively low-stiffness materials such as rubber or cork for applications at low pressures and where the surfaces are not very flat. For higher pressures and loads, one must utilize various composite materials and metal-encased materials as in Fig. 24.5.

For the highest pressures and loads a gasket may be retained in a groove and made either of very strong composite materials or even metal, as shown in Fig. 24.4.

## Self-Energized Seals

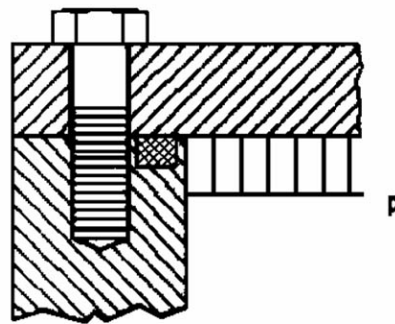
Elastomeric or **self-energized** rings can seal pressures to 20 MPa or even higher. As shown in Figs. 24.6 and 24.7, the two metal parts are clamped tightly together and they are not supported by the elastomer. As the pressure increases, the rubber is pushed into the corner through which leakage would otherwise flow. An elastomer acts much like a fluid so that the effect of pressure on one side is to cause equal pressure on all sides. Thus, the elastomer pushes tightly against the metal walls and forms a seal. The limitation of this type of seal is that the rubber will flow or extrude out of the clearance when the pressure is high enough. This is often not a problem for static seals, since the gap can be made essentially zero as shown in Fig. 24.6, which represents a typical way to utilize an elastomeric seal for static sealing.

**Figure 24.6** Elastomeric O-ring.



6. Elastomeric O-Ring

**Figure 24.7** Elastomeric rectangular ring.



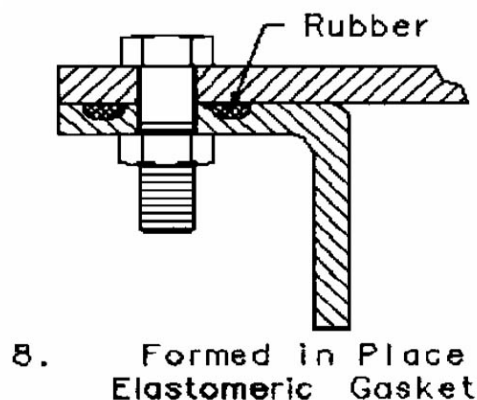
7. Elastomeric Rectangular Ring

Although the O-ring (circular cross section) is by far the most common elastomeric seal, one can also utilize rectangular cross sections (and even other cross sections) as shown in Fig. 24.7.

## Chemical Compound or Liquid Sealants as Gaskets

Formed-in-place gaskets such as in Fig. 24.8 are made by depositing a liquid-state compound on one of the surfaces before assembly. After curing, the gasket retains a thickness and flexibility, allowing it to seal very much like a separate gasket. Such gaskets are most commonly created using room temperature vulcanizing rubbers (RTV), but other materials including epoxy can be used.

**Figure 24.8** Formed-in-place elastomeric gasket.



8. Formed in Place Elastomeric Gasket

While formed-in-place gaskets retain relatively high flexibility, there are other types of plastic materials (including epoxy and anaerobic hardening fluids) that can be used to seal two surfaces. These fluids are coated on the surfaces before assembly. Once the joint is tightened and the material hardens, it acts like a form-fitted plastic gasket, but it has the advantage that it is also bonded to the sealing surfaces. Within the limits of the ability of the materials to deform, these types of gaskets make very tight joints. But one must be aware that relative expansion of dissimilar materials so bonded can weaken the bond. Thus, such sealants are best utilized when applied to tight-fitting assemblies. These same materials are used to lock and seal threaded assemblies, including pipe fittings.

There have been many developments of chemical compounds for sealing during the past 25 years, and one is well advised to research these possibilities for sealing/assembly solutions.

## 24.3 Dynamic Seals

---

Dynamic seals can be categorized as follows:

- Rotating or oscillating shaft

  - Fixed clearance seals

    - Labyrinth

    - Clearance or bushing

    - Visco seal

    - Floating-ring seal

    - Ferrofluid seal

- Surface-guided seals

  - Cylindrical surface

    - Circumferential seal

    - Packing

    - Lip seal

    - Elastomeric ring

  - Annular** surface (radial face)

    - Mechanical face seal

    - Lip seal

    - Elastomeric ring

- Reciprocating

  - Fixed clearance seals

    - Bushing seal

    - Floating-ring seal

    - Clearance or bushing

  - Surface-guided seals

    - Elastomeric rings

    - Solid cross section

    - U-cups, V-rings, chevron rings

    - Split piston rings

  - Limited-travel seals

    - Bellows

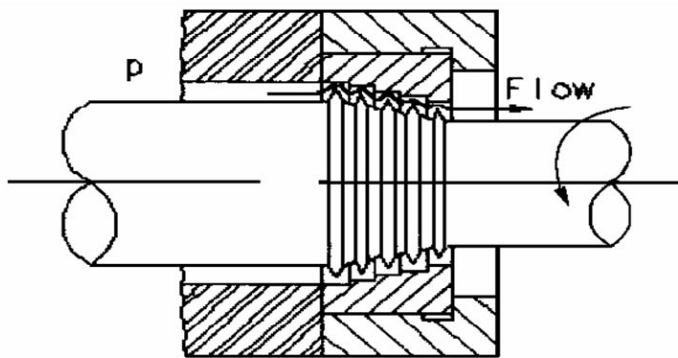
    - Diaphragm

One finds considerable differences between dynamic seals for rotating shaft and dynamic seals for reciprocating motion, although there is some crossover. One of the largest differences in seal types is between fixed-clearance seals and surface-guided seals. Fixed-clearance seals maintain a sealing gap by virtue of the rigidity of the parts and purposeful creation of a fixed sealing clearance. Surface-guided seals attempt to close the sealing gap by having one of the sealing surfaces actually (or nearly) touch and rub on the other, so that the position of one surface becomes guided by the other. Fixed-clearance seals leak more than surface-guided seals as a rule, but each has its place. Finally, dynamic seals usually seal to either cylindrical surfaces or annular (radial) surfaces. Sealing to cylindrical surfaces permits easy axial freedom, whereas sealing to radial surfaces permits easy radial freedom. Many seals combine these two motions to give the needed freedom of movement in all directions.

## Rotating or Oscillating Fixed-Clearance Seals

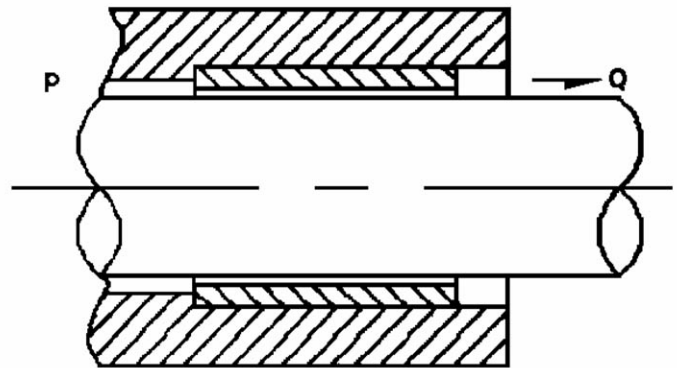
The labyrinth seal is shown in [Fig. 24.9](#). This seal has a calculable leakage depending on the exact shape, number of stages, and clearance and is commonly used in some compressors and turbomachinery as interstage seals and sometimes as seals to atmosphere. Its components can be made of readily wearable material so that a minimum initial clearance can be utilized.

**Figure 24.9** Labyrinth seal. (Source: Lebeck, A. O. 1991. *Principles and Design of Mechanical Face Seals*. John Wiley & Sons, New York. With permission.)



9. Labyrinth Seal

**Figure 24.10** Bushing seal. (Source: Lebeck, A. O. 1991. *Principles and Design of Mechanical Face Seals*. John Wiley & Sons, New York. With permission.)



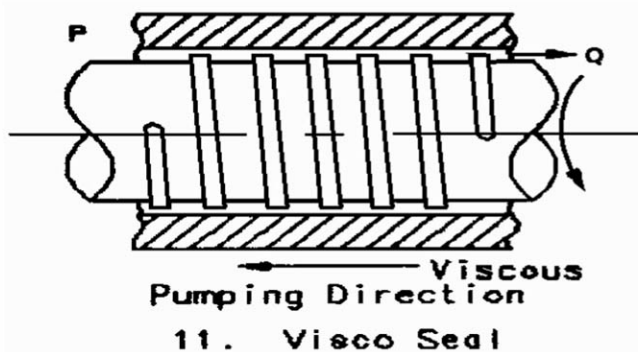
10. Bushing Seal

The clearance or bushing seal in [Fig. 24.10](#) may leak more for the same clearance, but this represents the simplest type of clearance seal. Clearance bushings are often used as backup seals to limit flow in the event of failure of yet other seals in the system. As a first approximation, flow can be estimated using flow equations for fluid flow between parallel plates. Clearance-bushing leakage increases significantly if the bushing is eccentric.

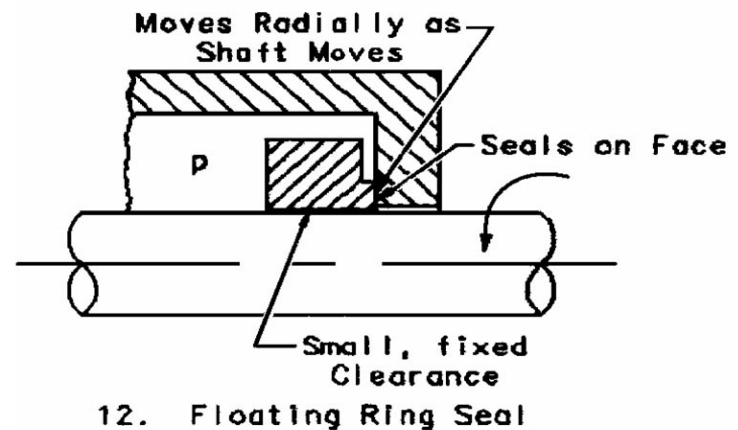
In high-speed pumps and compressors, bushing seals interact with the shaft and bearing system dynamically. Bushing seals can utilize complex shapes and patterns of the shaft and seal surfaces to minimize leakage and to modify the dynamic stiffness and damping characteristics of the seal.

The visco seal or windback seal in Fig. 24.11 is used to seal highly viscous substances where it can be fairly effective. It acts like a screw conveyor, extruder, or spiral pump to make the fluid flow backward against sealed pressure. It can also be used at no differential pressure to retain oil within a shaft seal system by continuously pumping leaked oil back into the system.

**Figure 24.11** Visco seal. (Source: Lebeck, A. O. 1991. *Principles and Design of Mechanical Face Seals*. John Wiley & Sons, New York. With permission.)



**Figure 24.12** Floating-ring seal. (Source: Lebeck, A. O. 1991. *Principles and Design of Mechanical Face Seals*. John Wiley & Sons, New York. With permission.)

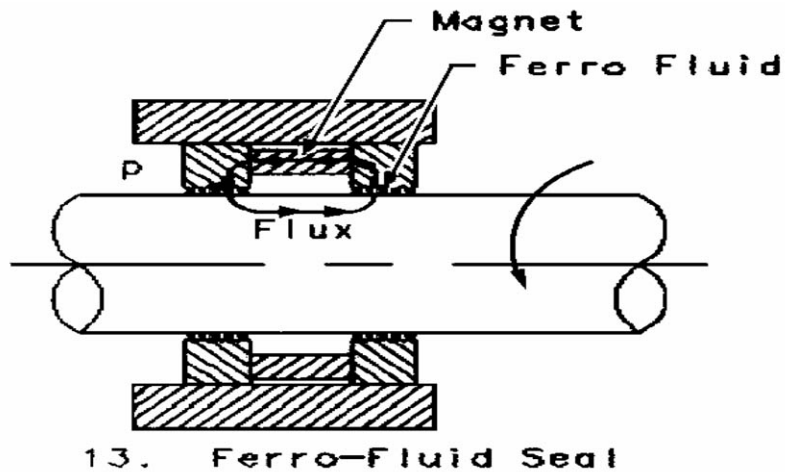


The floating-ring seal in Fig. 24.12 is used in gas compressors (can be a series of floating rings). It can be used to seal oil where the oil serves as a barrier to gas leakage or it can seal product directly. This seal can be made with a very small clearance around the shaft because the seal can float radially to handle larger shaft motions. The floating-ring seal is a combination of a journal bearing where it fits around the shaft and a face seal where it is pressed against the radial face. Most of the leakage is between the shaft and the bore of the bushing, but some leakage also occurs at the face. This seal can be used in stages to reduce leakage. It can be balanced to reduce the load on the radial face. Leakage can be less than with a fixed-bushing seal.

The **ferrofluid** seal in Fig. 24.13 has found application in computer disk drives where a true "positive seal" is necessary to exclude contaminants from the flying heads of the disk. The ferrofluid seal operates by retaining a **ferrofluid** (a suspension of iron particles in a special liquid) within the magnetic flux field, as shown. The fluid creates a continuous bridge between the rotating and nonrotating parts at all times and thus creates a positive seal. Each stage of a ferrofluid seal is capable of withstanding on the order of 20000 Pa (3 psi), so although these seals can be staged they are usually limited to low-differential pressure applications.



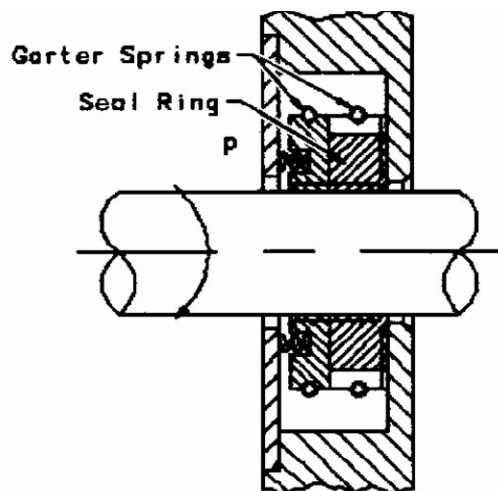
**Figure 24.13** Ferrofluid seal. (Source: Lebeck, A. O. 1991. *Principles and Design of Mechanical Face Seals*. John Wiley & Sons, New York. With permission.)



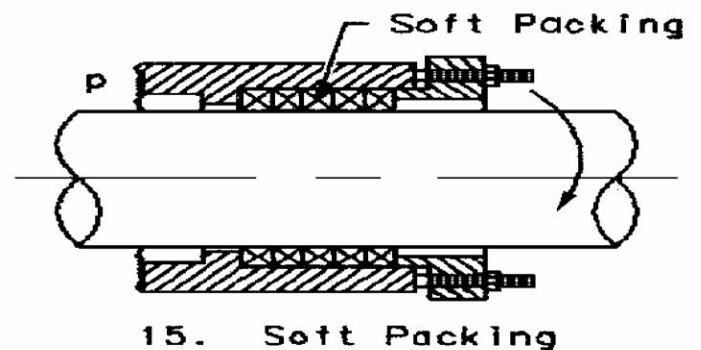
## Rotating Surface-Guided Seals—Cylindrical Surface

Figure 24.14 shows a segmented circumferential seal. The seal consists of angular segments with overlapping ends, and the segments are pulled radially inward by garter spring force and the sealed pressure. The seal segments are pushed against the shaft and thus are surface guided. They are also pushed against a radial face by pressure. This seal is similar to the floating-ring seal except that the seal face is pushed tight against the shaft because the segments allow for circumferential contraction. Circumferential segmented seals are commonly used in aircraft engines to seal oil and gas.

**Figure 24.14** Circumferential seal.

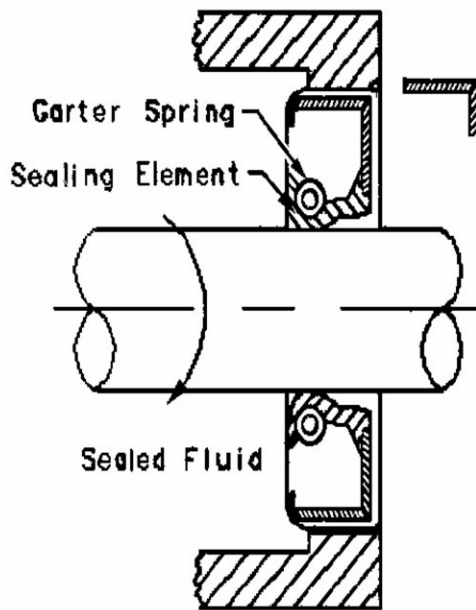


**Figure 24.15** Soft packing.



There are many types of soft packing used in the manner shown in [Fig. 24.15](#). The packing is composed of various types of fibers and is woven in different ways for various purposes. It is often formed into a rectangular cross section so it can be wrapped around a shaft and pushed into a packing gland as shown. As the packing nut is tightened the packing deforms and begins to press on the shaft (or sleeve). Contact or near contact with the shaft forms the seal. If the packing is overtightened the packing material will generate excessive heat from friction and burn. If it is too loose, leakage will be excessive. At the point where the packing is properly loaded, there is some small leakage which acts to lubricate between the shaft and the packing material. Although other types of sealing devices have replaced soft packing in many applications, there are still many applications (e.g., pump shafts, valve stems, and hot applications) that utilize soft packing, and there has been a continuous development of new packing materials. Soft packing for continuously rotating shafts is restricted to moderate pressures and speeds. For valve stems and other reciprocating applications, soft packing can be used at high pressure and temperature.

**Figure 24.16 Lip seal.**



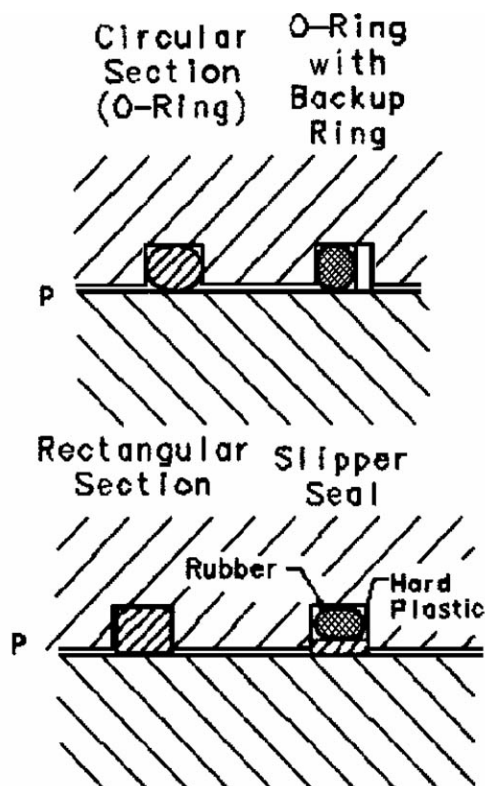
**16. Lip Seal**

The lip seal (oil seal) operating on a shaft surface represents one of the most common sealing arrangements. The lip seal is made of rubber (or, much less commonly, a plastic) or similar material that can be readily deflected inward toward the shaft surface by a garter spring. The lip is very lightly loaded, and, in operation in oils with rotation, a small liquid film thickness develops between the rubber lip and the shaft. The shape of the cross section determines which way the seal will operate. As shown in [Fig. 24.16](#) the seal will retain oil to the left. Lip seals can tolerate only moderate pressure (100000 Pa maximum). The normal failure mechanism is deterioration (stiffening) of the rubber, so lip seals have a limited speed and temperature of service. Various elastomers are best suited for the variety of applications.

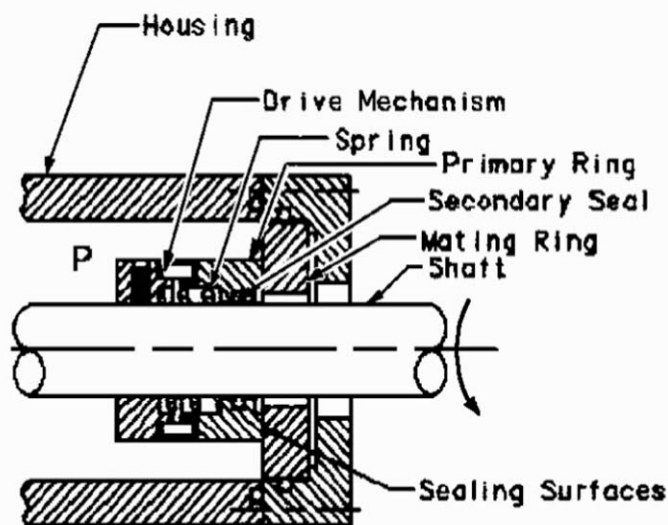


The elastomeric ring as described for static seals can also be used to seal continuous or oscillating rotary motion, given low-pressure and low-speed applications. As shown in Fig. 24.17, the control of the pressure on the rubber depends on the squeeze of the rubber itself, so that compression set of the rubber will cause a loss of the seal. But, yet, if the squeeze is too high, the seal will develop too much friction heat. The use of a backup ring under high-pressure or high-gap conditions and the slipper seal to reduce friction are also shown in Fig. 24.17.

**Figure 24.17** Elastomeric ring seals for rotating and reciprocating motion.



**Figure 24.18** Mechanical face seal. (Source: Lebeck, A. O. 1991. *Principles and Design of Mechanical Face Seals*. John Wiley & Sons, New York. With permission.)



## Rotating Surface-Guided Seals—Annular Surface

The mechanical face seal, as shown in Fig. 24.18, has become widely used to seal rotating and oscillating shafts in pumps and equipment. The mechanical face seal consists of a self-aligning primary ring, a rigidly mounted mating ring, a secondary seal such as an O-ring or bellows that gives the primary ring freedom to self-align without permitting leakage, springs to provide loading of the seal faces, and a drive mechanism to flexibly provide the driving torque. It is common to have the pressure to be sealed on the outside, but in some cases the pressure is on the inside. The flexibly mounted primary ring may be either the rotating or the nonrotating member.

Face seal faces are initially lapped very flat (1 micrometer or better) so that when they come into contact only a very small leakage gap results. In fact, using suitable materials, such faces lap themselves into conformity so that such a seal can leak as little as a drop of liquid per hour. Face seals also can be used for sealing gas.

One may also utilize a lip seal or an elastomeric ring to seal rotationally on an annular face.

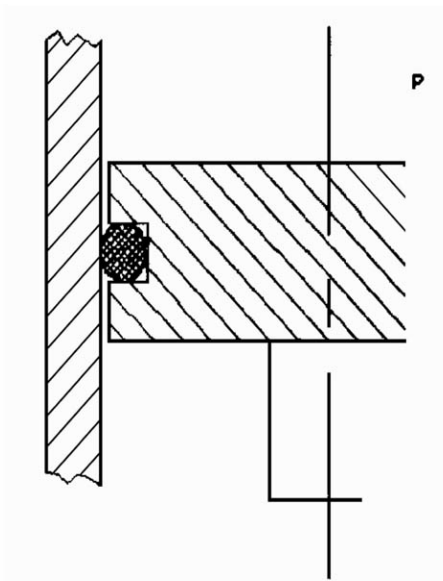
## Reciprocating Fixed-Clearance Seals

The clearance or bushing seal (Fig. 24.10) and the floating-ring seal (Fig. 24.12) can also be used for reciprocating motion, such as sealing piston rods. In fact, the bushing can be made to give a near-zero clearance by deformation in such applications.

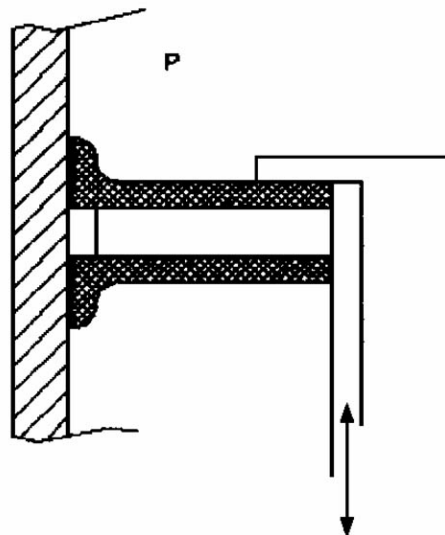
## Reciprocating Surface-Guided Seals

An elastomeric ring can be used to seal the reciprocating motion of a piston, as shown in Fig. 24.19. But more commonly used for such applications are cup seals (Fig. 24.20), U-cups, V- or chevron rings, or any of a number of specialized shapes (Fig. 24.21). Various types of these seals are used to seal piston rods, hydraulic cylinders, air cylinders, pumping rods, and pistons.

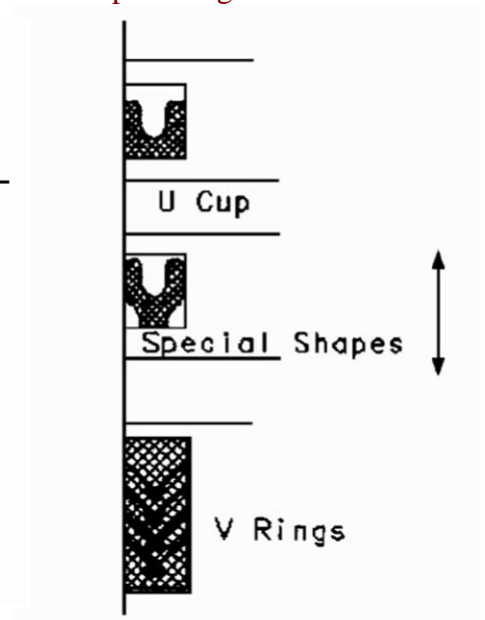
**Figure 24.19** Elastomeric ring seal.



**Figure 24.20** Cup seal.



**Figure 24.21** Elastomeric ring reciprocating seals.

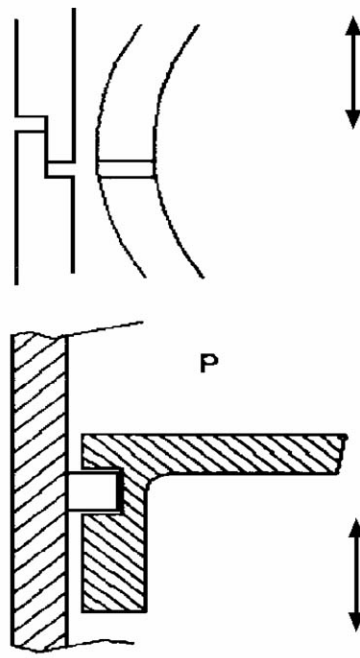


Split rings such as shown in Fig. 24.22 can be made of rigid materials. They are split for installation and so that they are loaded tightly against the wall by fluid pressure. Metal piston rings can be used in very hot environments. Plastic piston rings are suited to lower-temperature compressors.

## Reciprocating Limited-Travel Seals

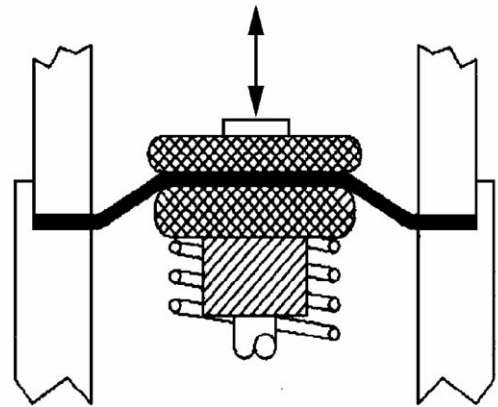
Most commonly used in pressure regulator and other limited-travel devices is the diaphragm shown in Fig. 24.23. Properly designed, this seal can be absolute and have significant travel. It can also allow for angular misalignment. In Fig. 24.24 is shown a metal bellows and in Fig. 24.25 is a rubber bellows. Both of these permit limited axial and angular motion. They have the advantage of being absolute seals because they do not rely on a sealing interface or suffer from wear and have no significant friction. Metal bellows may be made from edge-welded disks as shown or formed from a thin metal tube.

**Figure 24.22** Split ring seal (piston ring).

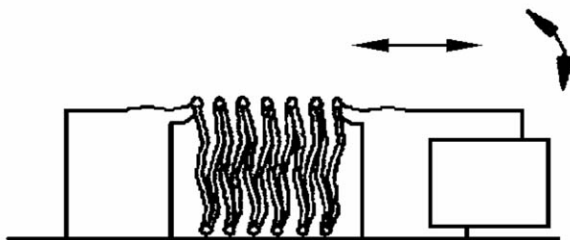


22. Split Ring Seal  
(Piston Ring)

**Figure 24.23** Diaphragm.

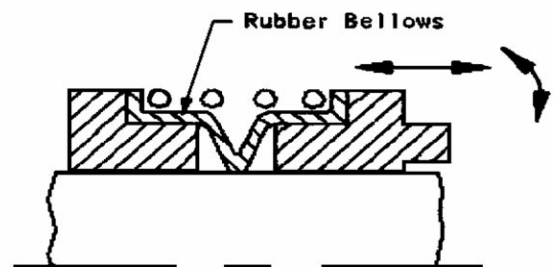


**Figure 24.24** Welded metal bellows.



24. Welded Metal Bellows

**Figure 24.25** Rubber bellows.



25. Rubber Bellows

## 24.4 Gasket Practice

For a gasket to seal, certain conditions must be met. There must be enough bolt or clamping force initially to seat the gasket. Then there also must be enough force to keep the gasket tightly clamped as the joint is loaded by pressure.

One may take the ASME Pressure Vessel Code [1980] formulas and simplify the gasket design procedure to illustrate the basic ideas. The clamping force, to be applied by bolts or other suitable means, must be greater than the larger of the following:

$$W_1 = \frac{\pi}{4} D^2 P + \pi 2b D m P \quad (24.1)$$

$$W_2 = \pi D b y \quad (24.2)$$

where

$D$  = effective diameter of gasket (m)

$b$  = effective seating width of gasket (m)

$2b$  = effective width of gasket for pressure (m)

$P$  = maximum pressure (Pa)

$m$  = gasket factor

$y$  = seating load (Pa)

Equation (24.1) is a statement that the clamping load must be greater than the load created by pressure plus a factor  $m$  times the same pressure applied to the area of the gasket in order to keep the gasket tight. Equation (24.2) is a statement that the initial clamping load must be greater than some load associated with a seating stress on the gasket material. To get some idea of the importance of the terms, a few  $m$  and  $y$  factors are given in Table 24.1. One should recognize that the procedure presented here is greatly simplified, and the user should consult one of the comprehensive references cited for details.

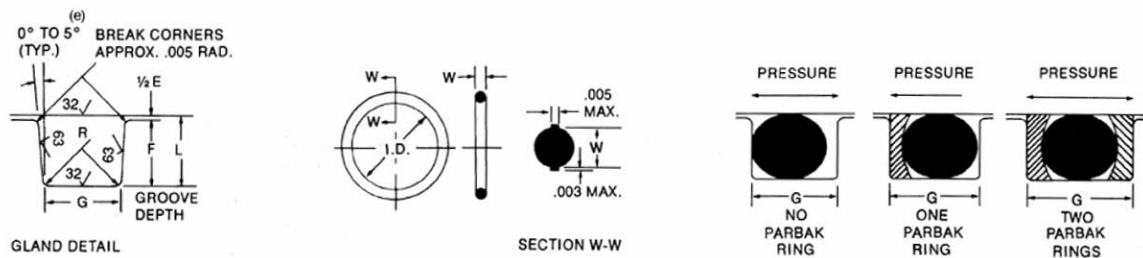
**Table 24.1** Gasket Factors

Type	$m$	$y$ (MPa)
Soft elastometer	0.5	0
Elastometer with fabric insertion	2.5	20
Metal jacketed and filled	3.5	55
Solid flat soft copper	4.8	90

## 24.5 O-Ring Practice

To seal properly, an O-ring must have the proper amount of squeeze or **preload**, have enough room to thermally expand, not have to bridge too large a gap, have a rubber hardness suitable to the job, and be made of a suitable rubber. Table 24.2 shows an abbreviated version of recommendations for static O-rings and Table 24.3 for reciprocating O-rings. In many cases one will want to span gaps larger or smaller than those recommended in the tables, so Fig. 24.26 shows permissible gap as a function of pressure and hardness based on tests.

**Table 24.2** Static O-Ring Grooves—Design Chart A5-1 for Industrial O-Ring Static Seal Glands



Refer to design chart A5-1 (below) and table A5-1 for dimensions.

**Table 24.2** Static O-Ring Grooves—Design Chart A5-1 for Industrial O-Ring Static Seal Glands

O-Ring Size Parker 2-	W Cross Section		L Gland Depth <sup>1</sup>	Squeeze <sup>1</sup>		E Diametral Clearance <sup>2,3</sup>	G Groove Width			R Groove Radius	Eccentricity Max. <sup>4</sup>
	Nominal	Actual		Actual	%		No Parbak Rings	One Parbak Ring	Two Parbak Rings		
004 through 050	1/16	.070 ±.003	.050 to .052	.015 to .023	22 32	.002 to .005	.093 to .098	.138 to .143	.205 to .210	.005 to .015	.002
102 through 178	3/32	.103 ±.003	.081 to .083	.017 to .025	17 24	.002 to .005	.140 to .145	.171 to .176	.238 to .243	.005 to .015	.002
201 through 284	1/8	.139 ±.004	.111 to .113	.022 to .032	16 23	.003 to .006	.187 to .192	.208 to .213	.275 to .280	.010 to .025	.003
309 through 395	3/16	.210 ±.005	.170 to .173	.032 to .045	15 21	.003 to .006	.281 to .286	.311 to .316	.410 to .415	.020 to .035	.004
425 through 475	1/4	.275 ±.006	.226 to .229	.040 to .055	15 20	.004 to .007	.375 to .380	.408 to .413	.538 to .543	.020 to .035	.005

<sup>1</sup>For ease of assembly when Parbaks are used, gland depth may be increased up to 5%.

<sup>2</sup>Clearance gap must be held to a minimum consistent with design requirements for temperature range variation.

<sup>3</sup>Reduce maximum diametral clearance 50% when using silicone or fluorosilicone O-rings.

<sup>4</sup>Total indicator reading between groove and adjacent bearing surface.

(e) 0° preferred.

Source: Parker Hannifin Corporation. 1990. *Parker O-Ring Handbook*. Parker Hannifin Corporation. Cleveland, OH. With permission.

**Table 24.3** Reciprocating O-Ring Grooves—Design Chart A6-5 for Industrial Reciprocating O-Ring Packing Glands

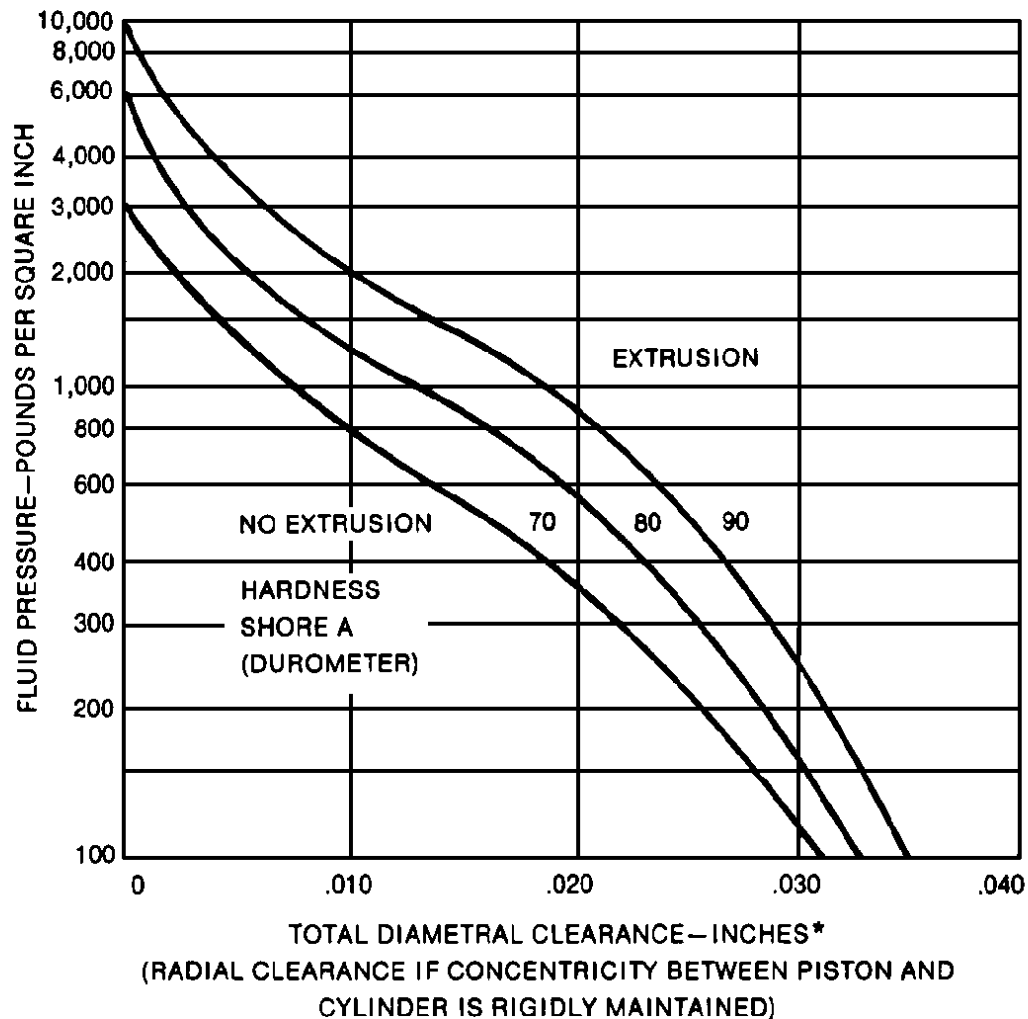
O-Ring Size Parker 2-	W Cross Section		L Gland Depth	Squeeze		E Diametral Clearance <sup>1</sup>	G Groove Width			R Groove Radius	Eccentricity Max. <sup>2</sup>
	Nominal	Actual		Actual	%		No Parbak Rings	One Parbak Ring	Two Parbak Rings		
006 through 012	1/16	.070 ±.003	.055 to .057	.010 to .018	15 25	.002 to .005	.093 to .098	.138 to .143	.205 to .210	.005 to .015	.002
104 through 116	3/32	.103 ±.003	.088 to .090	.010 to .018	10 17	.002 to .005	.140 to .145	.171 to .176	.238 to .243	.005 to .015	.002
201 through 222	1/8	.139 ±.004	.121 to .123	.012 to .022	9 16	.003 to .006	.187 to .192	.208 to .213	.275 to .280	.010 to .025	.003
309 through 349	3/16	.210 ±.005	.185 to .188	.017 to .030	8 14	.003 to .006	.281 to .286	.311 to .316	.410 to .415	.020 to .035	.004
425 through 460	1/4	.275 ±.006	.237 to .240	.029 to .044	11 16	.004 to .007	.375 to .380	.408 to .413	.538 to .543	.020 to .035	.005

<sup>1</sup>Clearance (extrusion gap) must be held to a minimum consistent with design requirements for temperature range variation.

<sup>2</sup>Total indicator reading between groove and adjacent bearing surface.

Source: Parker Hannifin Corporation. 1990. *Parker O-Ring Handbook*. Parker Hannifin Corporation. Cleveland, OH. With permission.

**Figure 24.26** Limits for extrusion. (Source: Parker Hannifin Corporation. 1990. *Parker O-Ring Handbook*. Parker Hannifin Corporation. Cleveland, OH. With permission.)



\*REDUCE THE CLEARANCE SHOWN BY 50% WHEN USING SILICONE OR FLUROSILICONE ELASTOMERS.

#### **FIGURE A4-2 LIMITS FOR EXTRUSION**

##### **BASIS FOR CURVES**

1. 100,000 pressure cycles at the rate of 60 per minute from zero to the indicated pressure.
2. Maximum temperature (i.e. test temperature) 160°F.
3. No back-up rings.
4. Total diametral clearance must include cylinder expansion due to pressure.
5. Apply a reasonable safety factor in practical applications to allow for excessively sharp edges and other imperfections and for higher temperatures.

Whereas nitrile rubber is most common and suitable for oils and aqueous solutions, fluorocarbon is excellent for hot oils. Many of the elastomer materials are made into O-rings and find application in certain chemical environments. Proper O-ring elastomer selection using one of the extensive recommendation tables [ASME, 1980; Lebeck, 1991] is essential for good performance.

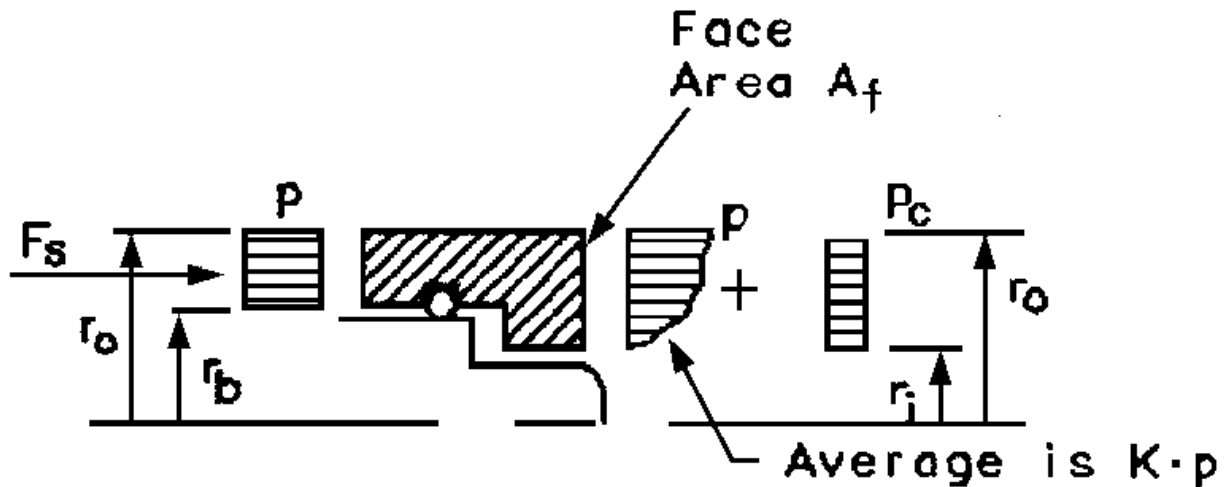
## 24.6 Mechanical Face Seal Practice

Figure 24.27 shows how, in general, the area on which the pressure is acting to load the primary ring may be smaller (or larger) than the area of the face. Thus, the balance ratio for a mechanical seal is defined as

$$B = \frac{r_o^2 - r_b^2}{r_o^2 - r_i^2} \quad (24.3)$$

where balance ratios less than 1.0 are considered to be "balanced" seals where in fact the face load pressure is made less than the sealed pressure. If balance ratio is greater than 1.0, the seal is "unbalanced."

**Figure 24.27** Mechanical seal elementary theory.



## 27. Mechanical Seal Elementary Theory

Balance radius ( $r_b$ ) of a seal is used by seal designers to change balance ratio and thus to change the load on the seal face. With reference to Fig. 24.27, and noting that the face area is

$$A_f = \pi(r_o^2 - r_i^2) \quad (24.4)$$



the average **contact pressure** (load pressure not supported by fluid pressure) on the face is given by

$$p_c = (B - K)p + \frac{F_s}{A_f} \quad (24.5)$$

where the  $K$  factor represents the average value of the distribution of the fluid pressure across the face. For well-worn seals in liquid,  $K = 1/2$  and, for a compressible fluid,  $K$  approaches  $2/3$ .

The sliding speed of the seal is based on the average face radius, or

$$V = \frac{r_o + r_i}{2} \omega \quad (24.6)$$

The severity of service for the seal is taken as the pressure times the sliding speed, or

$$(PV)_{\text{total}} = pV \quad (24.7)$$

The severity of operating conditions for the seal materials is the contact pressure times the sliding speed, or

$$(PV)_{\text{net}} = p_c V \quad (24.8)$$

The maximum allowable net  $PV$  is materials- and environment-dependent. For liquids the limiting values of [Table 24.4](#) are generally used.

**Table 24.4** Limiting Values for Liquids

Materials	$(PV)_{\text{net}}$ (psi·ft/min)	$(PV)_{\text{net}}$ (Pa·m/s) · 10 <sup>6</sup>
Carbon graphite/alumina	100 000	$3.5 \cdot 10^6$
Carbon graphite/tungsten carbide	500 000	$17.5 \cdot 10^6$
Carbon graphite/silicon carbide	> 500 000	> $17.5 \cdot 10^6$

Friction or seal power can be estimated from

$$P = p_c A_f f_c V \quad (24.9)$$

where  $P$  is the power and  $f_c$  is the friction coefficient, with values ranging from 0.07 for carbon graphite on silicon carbide to 0.1 for carbon graphite on tungsten carbide.

## Defining Terms

**Annulus:** The radial face of a rectangular cross-section ring.



**Contact pressure:** At a seal interface a part of the force needed for equilibrium is supplied by fluid pressure and a part by contact pressure.

**Elastomer(ic):** A material having the property of recovery of shape after deformation; rubberlike materials.

**Ferrofluid:** A liquid containing a suspension of magnetic particles.

**Preload:** The clamping load before pressure is applied.

**Sealing clearance:** The effective gap between two surfaces.

**Self-energized:** The preload is supplied by the elastic behavior of the material itself.

## References

American Society of Mechanical Engineers. 1980. *Code for Pressure Vessels*, Section VIII, Div 1. ASME, New York.

Lebeck, A. O. 1991. *Principles and Design of Mechanical Face Seals*. John Wiley & Sons, New York.

Parker Hannifin Corporation. 1990. *Parker O-Ring Handbook*. Parker Hannifin Corporation. Cleveland, OH.

## Further Information

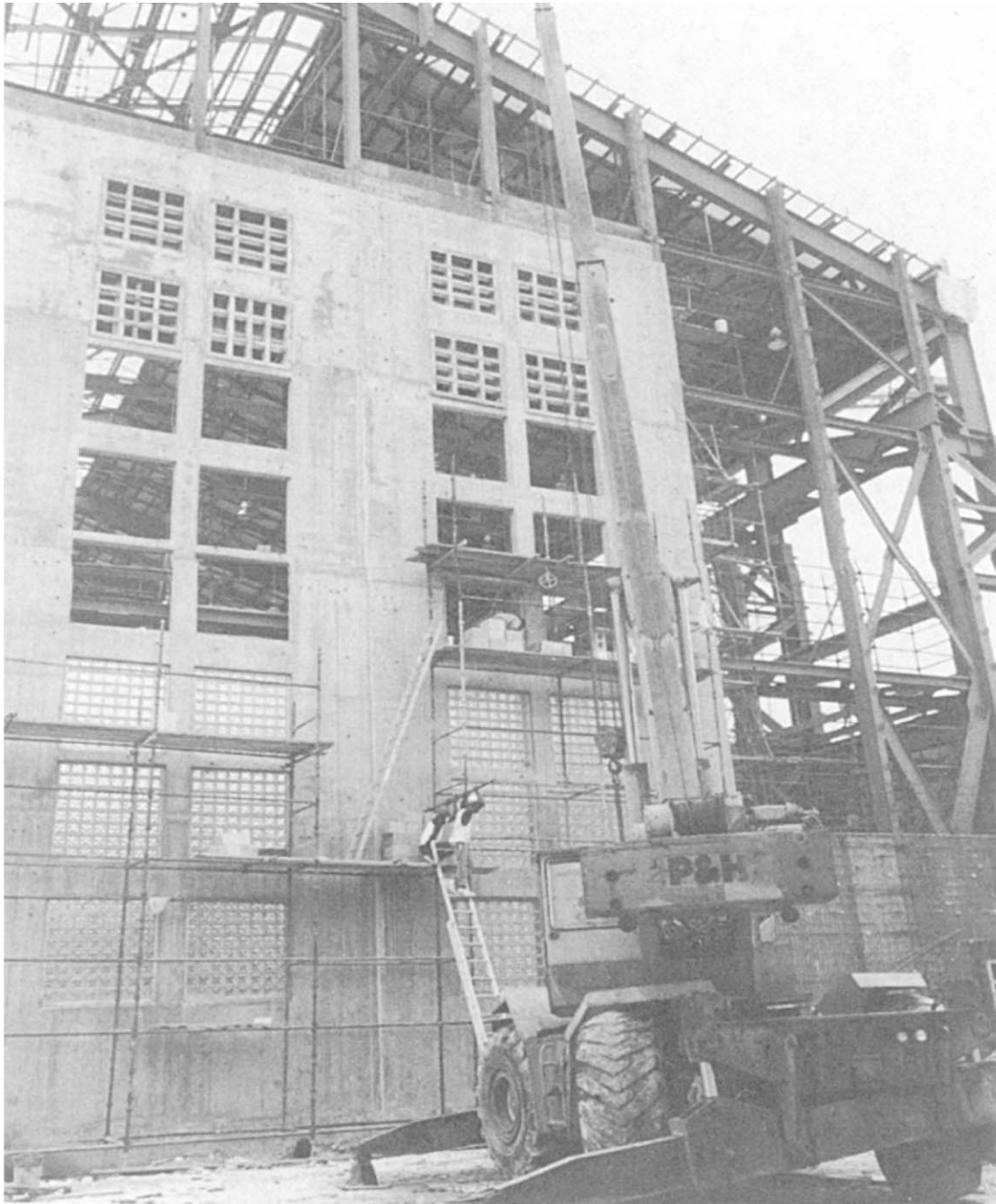
Brink, R. V., Czernik, D. E., and Horve, L. A. 1993. *Handbook of Fluid Sealing*. McGraw-Hill, New York.

Buchter, H. H. 1979. *Industrial Sealing Technology*. John Wiley & Sons, New York.

Kaydon Ring & Seals, Inc. 1987. *Engineer's Handbook—Piston Rings, Seal Rings, Mechanical Shaft Seals*. Kaydon Rings & Seals, Inc. Baltimore, MD.

Warring, R. H. 1981. *Seals and Sealing Handbook*. Gulf, Houston, TX.

McCormac, J. "Structures"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



Construction on the Clifton Pier Power Station, located in Nassau in the Bahamas, began in 1992 and was completed approximately two years later. Over 1000 tons of steel was used in the construction of this structure. The steel for this project was sent by rail from Bethlehem Steel Company in Pennsylvania to Florida where the steel was manufactured by Steel Fabricators, Inc. The fabricated steel was then trailered to barges and shipped to Nassau.

All of the steel used in this structure was painted with a zinc paint (CZ11HS) to prevent rust caused by weather conditions in the Bahamas.

The Clifton Pier Power Station, housing huge diesel generators up to 30 feet in diameter, required double-shafted columns in order to support a 300 ton overhead crane used to move the diesel generators. The double-shafted column design is unique to this structure. (Photo Courtesy of Steel Fabricators, Inc.)

**Jack McCormac***Clemson University*

- 25 **Loads** *P. Gergely*  
Dead Loads • Live Loads • Impact Loads • Snow Loads
- 26 **Wind Effects** *T. A. Reinhold and B. L. Sill*  
Wind Climate • Local Wind Exposure • Mean Wind Speed Profile • Turbulence • Pressure Coefficients and Load Factors
- 27 **Earthquake Effects** *M. K. Ravindra and J. G. Shipp*  
Why Do Earthquakes Occur? • Characteristics of Earthquakes • Damage Mechanisms • Seismic Hazard Analysis • Earthquake Damage Surveys • Earthquake-Resistant Design
- 28 **Structural Analysis** *R. G. Sexsmith and T. M. Cigic*  
Beams • Trusses • Frames • Computer-Aided Analysis
- 29 **Structural Steel** *W. T. Segui*  
Members • Connections • Composite Members • Computer Applications
- 30 **Concrete** *E. G. Nawy*  
Structural Concrete • Flexural Design of Reinforced Concrete Members • Shear and Torsion Design of Reinforced Concrete Members • Prestressed Concrete • Serviceability Checks • Computer Applications for Concrete Structures
- 31 **Timber** *D. E. Breyer*  
Durability of Wood • Wood Products • Member Design • Connections • Lateral Force Design
- 32 **Masonry Design** *J. E. Amrhein*  
Basis of Design • Masonry Materials • Masonry Units • Concrete Masonry • Mortar • Grout • Unreinforced Masonry • Strength of Masonry • Design of Reinforced Masonry Members • Design of Structural Members—Strength Design

STRUCTURAL ENGINEERING EMBRACES the analysis and design of buildings, bridges, dams, towers, cables, arches, storage tanks, concrete pavements, and other structures—a list too lengthy to enumerate.

Perhaps the most important and most difficult task faced by structural engineers is the accurate estimation of the loads that structures may have to support during their lives. Loads and their magnitudes are the first topics discussed in this part of the handbook. Particular emphasis is given to the difficult task of estimating wind and earthquake effects on structures.

When loads are applied to structures, those structures deform and stresses are caused in various parts of the structures. The calculation of the magnitudes of these stresses and deformations is called structural analysis. Reaching the present level of structural analysis has taken many centuries. It is a very complicated topic and yet one which confronts almost every branch of technology with questions of strength and deformations.

The manner in which structures are analyzed and designed has been drastically changed in recent years due to the availability of personal computers. Computers are readily available in almost every engineering school and office in the U.S. Up to the present time they have been used more for analysis than for design, but this situation is rapidly changing as more and more design software is becoming available. As a result, the structural engineer now should have more time to devote to the planning of structures and the selection of the materials to be used.

There is a gradual worldwide movement toward a limit states type of design. The term *limit state* is used to describe a condition at which a structure or some part of a structure ceases to perform its intended function. There are two categories of limit states—strength and serviceability. Not only must the load-carrying capacity of a structure equal or exceed its loading (buckling, fracture, fatigue, overturning, and so on), but the structure must perform under normal loading conditions in a satisfactory manner as to deflections, vibrations, slipping, cracking, and deterioration. This limit states philosophy is particularly reflected in the chapter on structural steel design.

Gergely, P. "Loads"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 25

## Loads

---

### 25.1 Dead Loads

### 25.2 Live Loads

Occupancy Loads • Bridge Live Loads

### 25.3 Impact Loads

### 25.4 Snow Loads

#### **Peter Gergely**

*Cornell University*

Structures are designed to carry various loads and load combinations without collapse and with an adequate margin of safety. In addition, several **serviceability** conditions (deflections, cracking) must also be satisfied for most structures. The expected maximum values of most loads can be estimated only approximately, and building codes give only estimates of the minimum design loads, based on judgment. The design loads and load combinations rarely occur for most structures.

The two main types of loads are **dead loads** and **live loads**. Dead loads include the weight of structural and most nonstructural components: beams, columns, walls, floor slabs, bridge decks, roofing, partitions, and ceiling and flooring materials. Live loads include occupancy loads (people, building contents, traffic, movable partitions) and snow. Wind forces, earthquake forces, water pressure, and blast are similar to live loads, but they are usually considered separately. The weights of movable materials in warehouses are usually considered as live loads. In addition to these loads, temperature effects can also be considered as loads.

Most designers and building codes rely on the *Minimum Design Loads for Buildings and Other Structures* [[ASCE 7-93, 1993](#)] or the *Standard Specifications for Highway Bridges* [[AASHTO, 1989](#)] for the design of buildings and bridges, respectively. Loads for special structures, such as liquid containers, towers, cranes, and power plants, are normally specified by trade or professional organizations.

Loads are combined to produce the maximum member forces. However, codes allow reduction of combined loads if the probability of simultaneous occurrence of maximum effects is low. For example, a 0.75 factor may be applied for the combined dead load, live load, and wind or earthquake. These factors are different in the working stress design approach and in the strength (or the load and resistance factor) design approach.

# 25.1 Dead Loads

Dead loads are made up almost entirely of the weights of all structural elements and permanent fixtures. Therefore, it is generally easy to calculate dead loads. However, in preliminary design the sizes of structural members (beams, columns, walls, and floor slabs) are not yet known and must be estimated. The unit weights of several common materials are listed in [Table 25.1](#).

**Table 25.1** Unit Weight of Common Construction Materials

Aluminum	165 lb/ft <sup>3</sup>
Brick	120 lb/ft <sup>3</sup>
Lightweight concrete	90–110 lb/ft <sup>3</sup>
Normal-weight concrete	150 lb/ft <sup>3</sup>
Steel	490 lb/ft <sup>3</sup>
Timber	35–40 lb/ft <sup>3</sup>
Roofing	5–10 lb/ft <sup>2</sup>
Tile	10–15 lb/ft <sup>2</sup>
6 in. hollow concrete block wall	43 lb/ft <sup>2</sup>

The maximum forces in structures sometimes occur during construction—for example, during the cantilever construction of bridges. It is important to consider the loads during various stages of construction.

# 25.2 Live Loads

There are many types of live loads: occupancy, weights in warehouses, traffic loads on bridges, construction or maintenance forces, automobiles in parking garages, and snow. These are much more variable than dead loads and require larger safety margins. (Wind and earthquake loads are considered separately as environmental loads.) These live loads are gravity loads and must be positioned (acting on all or part of the area) to cause maximum forces in the member being designed.

## Occupancy Loads

The major type of live load in buildings is caused by occupants. The minimum specified occupancy loads depend on the use and the likelihood of congregation of many people. Typical values of distributed loads are shown in [Table 25.2](#). In office buildings a 20 lb/ft<sup>2</sup> uniform load is used to account for the weight of movable partitions.

**Table 25.2** Typical Occupancy Loads

Theaters with fixed seats	60 lb/ft <sup>2</sup>
Theaters with movable seats	100 lb/ft <sup>2</sup>
Corridors and lobbies	100 lb/ft <sup>2</sup>
Garages	50 lb/ft <sup>2</sup>



Restaurants	100 lb/ft <sup>2</sup>
Library reading rooms	60 lb/ft <sup>2</sup>
Offices	50 lb/ft <sup>2</sup>
Stadium bleachers	100 lb/ft <sup>2</sup>
Stairways	100 lb/ft <sup>2</sup>

---

In addition to the distributed loads, structures are also designed for concentrated loads to account for concentrations of people or furniture. Typical values are 2000 lb on office floors and on slabs in garages for passenger cars. These are assumed to be acting on a 2.5 ft<sup>2</sup> area.

Since it is unlikely that a very large area or most floors of a building will have the full occupancy loads, most codes allow a live load reduction factor for such cases. However, reduction is not allowed for garages and areas of public assembly. In the design of a structural member, if the influence area is more than 400 ft<sup>2</sup>, the reduction factor is

$$0.25 + \frac{15}{\sqrt{A}} \quad (25.1)$$

with a minimum value of 0.5 for one floor and 0.4 for columns receiving loads from multiple floors. For columns the influence area is four times the tributary area (thus equal to the area of all four adjoining panels), and for beams it is twice the tributary area.

## Bridge Live Loads

The design forces in a bridge depend on the magnitude and distribution of the vehicle load. It is not reasonable to assume that only heavy trucks will travel at close spacing, especially on a long bridge. For short bridges the actual position of the heaviest truck is important, whereas for long bridges a uniform load could be used in design. Design codes [[AASHTO, 1989](#)] specify standard loads for short and long bridges. Several standard trucks are specified—for example, the H20-44, which has a total weight of 20 tons (the 44 signifies the year of adoption). Of this weight, 8000 lb acts under the front axle and 32000 lb under the rear wheels. Other standard truck loads are H10-44, H15-44, HS15-44, and HS20-44, where the HS designation is for semitrailers, with the weight allocation of 10%, 40%, and 40% for the cab, front trailer, and rear trailer wheels, respectively.

These concentrated loads must be placed on the bridge to produce maximum forces (shears and moments) in the member being designed. In addition to the individual truck loads, codes specify a uniform lane load combined with a single concentrated load. For an H10-44 loading the distributed load is 320 lb/ft and the concentrated load is 9000 lb; the respective numbers for HS20-44 are twice as large.

## 25.3 Impact Loads

---

Moving loads on bridges and crane girders can cause vibrations and increased stresses. Simple empirical impact formulas have been developed to account for this effect, although a large number

of variables, such as surface roughness, speed, and span, influence the impact effect.

In the AASHTO [1989] code the impact formula is

$$I = \frac{50}{L + 125} \quad (25.2)$$

where  $L$  is in feet. The maximum value of the impact factor  $I$  is 0.3. For shorter bridges the impact effect can be high, especially when a heavy vehicle travels on the bridge at high speed.

The loads created by elevators are often increased by 100% to account for impact. Likewise, impact factors have been recommended for various other types of machinery, ranging from 20 to 50%. Craneways have three factors—25%, 20%, and 10% for forces in the vertical, lateral, and longitudinal directions, respectively.

## 25.4 Snow Loads

---

The expected maximum snow accumulation in various regions is given in codes, usually for a 50-year mean recurrence interval. Values reach 50 psf in many areas but can be twice as much or more in regions with heavy snowfalls. In some regions—for example, in parts of the Rocky Mountains—local climate and topography dictate the design snow load level. The weight of snow depends on its density and typically ranges from 0.5 psf to 0.7 psf for 1 in. of snow after some compaction. Fresh dry snow has a specific gravity of only 0.2.

For flat roofs (outside Alaska) the snow load is

$$p_f = 0.7C_e C_t I p_g \quad (25.3)$$

$C_e$  is the exposure factor (0.8 for windy, open areas, 0.9 for windy areas with some shelter, 1.0 if wind does not remove snow, 1.1 with little wind, and 1.2 in a forested area). However, for large roofs (greater than 400 ft in one direction),  $C_e$  should not be less than unity [Lew *et al.*, 1987].  $C_t$  is the thermal factor (1.0 for heated structures, 1.1 just above freezing, 1.2 for unheated structures).  $I$  is the importance factor (ranges from 0.8 to 1.2 for various occupancies).  $p_g$  is the ground snow load from maps.

The flat-roof values are corrected for sloping roofs:

$$p_s = C_s p_f \quad (25.4)$$

where  $C_s$  is the slope factor, which is given in a diagram in ASCE 7-93 [1993]. For warm roofs ( $C_t = 1.0$ ) the slope factor is 1.0 for slopes less than about  $30^\circ$  and reduces linearly to zero as the slope increases to  $70^\circ$ . Thus roofs with slopes greater than  $70^\circ$  are assumed to have no snow load. Unbalanced snow load caused by a certain wind direction also must be considered.

## Defining Terms

**Dead load:** Gravity loads produced by the weight of structural elements, permanent parts of

structures such as partitions, and weight of permanent equipment.

**Impact factor:** Accounts for the increase in stresses caused by moving load effects.

**Live load:** Loads caused by occupancy and movable objects, including temporary loads.

**Serviceability:** Limit on behavior in service, such as on deflections, vibrations, and cracking.

## References

AASHTO. 1989. *Standard Specifications for Highway Bridges*, 14th ed. The American Association of State Highway and Transportation Officials, Washington, DC.

ASCE 7-93. 1993. *Minimum Design Loads for Buildings and Other Structures*. American Society of Civil Engineers, New York.

Lew, H. S., Simiu, E., and Ellingwood, B. 1987. Loads. In *Building Structural Design Handbook*, ed. R. N. White and C. G. Salmon, p. 9-43. John Wiley & Sons, New York.

## Further Information

Uniform Building Code, International Conference of Building Officials, 5360 South Workman Mill Road, Whittier, CA 90601

ASCE Standard, ASCE 7-93, *Minimum Design Loads for Buildings and Other Structures*, American Society of Civil Engineers, 345 East 47th Street, New York, NY 10017

Reinhold, T. A., Sill, B. L. "Wind Effects"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

**26.1 Wind Climate****26.2 Local Wind Exposure****26.3 Mean Wind Speed Profile**

Power Law Profile • Logarithmic Profile

**26.4 Turbulence****26.5 Pressure Coefficients and Load Factors****Timothy A. Reinhold***Clemson University***Ben L. Sill***Clemson University*

Wind is one of the two primary sources of lateral forces on land-based buildings and structures; the other is earthquake ground motion. Winds completely engulf the structure and generate complex distributions of pressures and, hence, loads on all exterior surfaces. Most surfaces experience negative pressures or suctions which tend to pull the building apart. Most roofs also experience negative pressures which act to lift the roof off of the walls and to pull roofing membranes and sheathing from the supporting structure. If the exterior surface contains openings, either because they were designed that way or because a **cladding** element fails, the interior of the structure can become exposed to some fraction of the external pressure that would have occurred at the opening. Internal pressures can also develop as a result of normal air leakage through the building skin or cladding. Internal pressures tend to be fairly uniform throughout the interior of the building and can significantly increase the loads on the walls and roof.

Wind effects on structures include the direct application of wind-induced forces, movement of the structure, and the flow of the wind around the structure, which may affect pedestrians or the function of the building. Normally, wind effects are grouped according to limit states and safety and serviceability considerations. The selection of structural systems based on their ability to resist wind-induced stresses with appropriate margins of safety is an example of a design for safety or an ultimate limit state. Limiting deflections caused by the wind loads to prevent cracking of walls or partitions and limiting the motion of the structure to prevent occupant discomfort are examples of serviceability limit state design. Following the large economic losses suffered in recent hurricanes, there have been increasing calls for protection of the building envelope against water penetration by protecting glazed openings from failure due to direct wind loads or impact loads from

wind-borne debris and by reducing the penetration of wind-driven rain.

The wind effects which should be considered in the design of a particular structure vary depending on the following factors:

1. The wind climate (the expected magnitude and frequency of wind events and consideration of the types of events—that is, hurricanes, thunderstorms, tornadoes, and extra-tropical storms).
2. The local wind exposure (siting of the building or structure, including the type of terrain and terrain features surrounding the structure and the influence of neighboring structures).
3. Pressure coefficients and load factors (coefficients that depend on the exterior shape of the building or structure and factors which relate wind loads to reference wind speeds).
4. Dynamic effects such as resonance response and aerodynamic instabilities, which can lead to failures or to significant increases in the dynamic response and loading (these dynamic effects are not covered by normal pressure coefficients and load factors; they depend on the shape of the building or structure and properties such as mass, stiffness, damping, and modes of vibration). Typically, tall, slender structures and long, suspended structures should be evaluated for these possible effects. These include long-span bridges, stacks, towers, buildings with height-to-width ratios greater than 5, and exposed long, flexible members of structures.

Within the space available, it is possible to present only a brief description of the types of wind effects which should be considered in the design of buildings and structures. Consequently, rather than reproduce a set of code-like requirements, the focus of this chapter is on describing some basic relationships which will help the engineer compare different code approaches to estimating wind loads. There are many different codes available throughout the world which use significantly different types of reference wind speeds. Often, it is not clear whether the codes will produce similar estimates of design loads if applied to the same structure (indeed, they often do produce significantly different loads). It is not the intent of this chapter to promote a particular code. The goal is to provide the tools which will allow the engineer to compare code estimates of design loads by using a consistent set of reference wind speeds, regardless of whether the code calls for a mean hourly speed, a ten-minute speed, a one-minute sustained wind speed, a fastest mile wind speed, or a gust wind speed.

In addition, the field of wind engineering remains highly empirical, which means that accurate estimates of wind loads and wind effects often require the conduct of a physical model study. The "For Further Information" section provides a list of references which can provide additional guidance on when a model study is warranted or desirable.

## 26.1 Wind Climate

---

The wind climate encompasses the range of wind events which may occur in a geographical region and the expected frequency and intensity of the event. Types of events include **extra-tropical cyclones**, thunderstorms, downbursts, microbursts, tornadoes, and hurricanes. Each type of storm

has potentially different wind characteristics of importance to buildings and structures, as well as separate occurrence rate and intensity relationships. For most engineering purposes, downbursts, microbursts, and tornadoes are not considered in the establishment of design winds and loads. Thunderstorm winds are frequently buried in the historical data records and, thus, are partially built into the design winds estimated from historical data. Recent work has been conducted to extract thunderstorm winds from the historical data at selected stations. This analysis suggests that it will be important in the future to treat thunderstorms as a separate population of wind events in much the same way that current analysis considers hurricane and tornado events as separate populations for statistical analysis.

The current approach to estimating design winds in hurricane-prone regions is to conduct Monte Carlo simulations of the events using statistical information on historical tendencies of hurricanes in the area. The probabilities of experiencing hurricane winds in coastal areas are then developed from the statistics produced from the simulation of thousands of years of storms. These occurrence probabilities are then combined with probabilities for nonhurricane events to estimate design winds for various return periods ranging from 10 to 100 years. This type of analysis has been used to produce design wind speed maps for the continental U.S., and the latest edition of ASCE-7 *Minimum Design Loads for Buildings and Structures* contains the most recent map which is generally adopted by model building codes in the U.S. This type of systematic analysis of hurricane and nonhurricane winds has not been conducted for the Hawaiian Islands or most of the rest of the world. Consequently, in hurricane-prone regions, the designer should endeavor to determine the source of the estimates for design wind speeds and the basis for the estimates. In some instances, it will be necessary to contract with a group experienced in hurricane simulations in order to produce reasonable estimates of design wind speeds.

In areas where hurricanes are not expected—which normally includes areas 100 miles inland from a hurricane coastline—a series of annual extreme wind speeds can be used to estimate the design wind speed for a specific return period using the following equation [Simiu, 1985], where it is assumed that the extreme wind climate follows a Type I extreme value distribution:

$$U_N = U_{\text{avg}} + 0.78\sigma_N (\ln N - 0.577) \quad (26.1)$$

where  $U_N$  is the design wind speed for a return period of  $N$  years,  $U_{\text{avg}}$  is the average of the annual extreme wind speeds, and  $\sigma_N$  is the standard deviation of the annual extreme wind speeds. It should be noted that the number of years of record greatly affects the reliability of the design wind speed estimate [Simiu, 1985].

Note also that an averaging time has not been specified in this discussion. The emphasis is placed on the wind speeds being extreme annual values which must all correspond to the same averaging time, regardless of whether it is a mean hourly, ten-minute mean, or shorter-duration averaging time. These maximum speeds must also correspond to a consistent set of terrain conditions and a consistent elevation. The U.S. codes currently use a fastest mile wind speed at a height of 10 m in open terrain as the reference design wind speed and reference conditions. This type of measurement has a variable averaging time since it corresponds to the time required for one mile of wind to pass a location. Thus, any calculation of a fastest mile wind speed requires

iteration. The averaging time for the fastest mile wind speed is calculated by:

$$t = 3600/U_{\text{FM}} \quad (26.2)$$

where  $t$  is given in seconds and  $U_{\text{FM}}$  is the fastest mile wind speed expressed in miles per hour.

Other countries use similar terrain conditions and usually specify a 10 m elevation, but use a wide variety of averaging times ranging from mean hourly (Canada) to peak gust (Australia), which is normally assumed to correspond to a two- to three-second averaging time. The following section provides relationships for converting maximum wind speeds from one averaging time, terrain exposure, and elevation to maximum wind speeds for a different averaging time, terrain, and elevation. These equations provide a means for converting a design wind speed in an unfamiliar code to one that can be used in a code with which the designer is more familiar, if the reference conditions for the two codes are different.

## 26.2 Local Wind Exposure

---

As wind moves over the surface of the earth, the roughness of trees, buildings, and other features reduces the wind speed and creates the **atmospheric boundary layer**. The greatest reduction occurs close to the ground, with reduced effects at greater heights. There is, in fact, a height—known as the *gradient height*—at which the wind is not affected by the surface characteristics. For engineering purposes, the gradient height is generally assumed to be between 300 and 600 m, depending on the terrain. Surface roughness also affects the air flow by creating turbulent eddies (or gusts) which can have a significant effect on buildings. The gusty nature of wind is random and is analyzed using statistical approaches.

## 26.3 Mean Wind Speed Profile

---

No single analytical expression perfectly describes the mean wind speed variation with height in the atmospheric boundary layer. The two used most often are the power law profile and the logarithmic profile. The logarithmic profile is the most widely accepted, although both can give adequate descriptions of the wind speed. Each is described below, and sufficient information is given to (a) allow transfer from one profile to the other, (b) effectively convert maximum speeds in one terrain to those in another, and (c) convert from one averaging time to another.

### Power Law Profile

Taking  $z$  as the elevation,  $U$  as the wind speed (with the bar indicating a mean hourly value), and  $g$  as conditions at the gradient height, this profile is written as:



$$\frac{\overline{U(z)}}{\overline{U_g}} = \left( \frac{z}{z_g} \right)^\alpha \quad (26.3)$$

Table 26.1 provides estimates for the gradient height and power law exponent for different terrains.

**Table 26.1** Parameters Used in Mean Velocity Profile Relations

Terrain	$\alpha$	$z_g$ (m)	$z_0$ (m)	$p$	$\beta$
Coastal	0.1	230	0.005	0.83	6.5
Open	0.14	275	0.07	1	6
Suburbs	0.22	375	0.3	1.15	5.25
Dense suburbs	0.26	400	1	1.33	4.85
City center	0.33	500	2.5	1.46	4

## Logarithmic Profile

This expression for the variation of mean wind speed with height utilizes an aerodynamic roughness length,  $z_0$ , and a shear velocity,  $u_*$ , which is a measure of the surface drag:

$$\overline{U(z)} = \frac{u_*}{k} \ln \left( \frac{z}{z_0} \right) \quad (26.4)$$

Here,  $k$  is von Karman's constant and is usually taken as 0.4. For rough surfaces, such as dense suburban and urban conditions, a displacement height,  $d$ , should be included in Eq. (26.4) by replacing  $z$  with  $z - d$ . For  $z$  values substantially greater than  $d$ , the correction is negligible. Writing the log law at two heights for two different terrain roughnesses (one of which is open country) and taking the ratio gives

$$\frac{\overline{U(z_1)}}{\overline{U(z_2)_{\text{open}}}} = p \frac{\ln(z_1/z_0)}{\ln(z_2/z_0)_{\text{open}}} \quad (26.5)$$

where  $p$  is the ratio of the shear velocities for the two different terrain conditions. Table 26.1 gives a summary of the parameters needed to use the profiles.

## 26.4 Turbulence

The wind speed can be divided into two parts—a mean or time-averaged part,  $\overline{U}$ , and a fluctuating or time varying part,  $u'$ . The long-term properties of the fluctuating part can be described by the variance or standard deviation,  $\sigma_u$ . The maximum wind speed for any averaging time,  $t$ , can be obtained by adjusting the hourly (3600 s) average as

$$U_t(z) = \overline{U}_{3600}(z) + C(t)\sigma_u(z) \quad (26.6)$$

where the coefficient  $C(t)$  is given in Table 26.2. The values of  $C(t)$  for extra-tropical winds were obtained by Simiu [1981], while the values of  $C(t)$  for hurricane winds reflect recent research which suggests that hurricane winds contain larger fluctuations than extra-tropical strong winds [Kramer, 1992].

**Table 26.2** Gust Factors for Use in Calculating Maximum Wind Speeds from Mean Speeds

Conditions	Time (s)	1	3	10	30	60	600	3600
Hurricane winds	$C(t)$	4	3.62	3.06	2.23	1.75	0.43	0
Extra-tropical winds	$C(t)$	3	2.86	2.32	1.73	1.28	0.36	0

To convert maximum wind speeds between open terrain conditions for any averaging time and elevation and another set of conditions (i.e., variations in terrain, averaging time, or elevation), the following combination of the above expressions can be used:

$$\frac{U_{t_1}(z_1)}{U_{t_2}(z_2)_{\text{open}}} = p \left[ \frac{\ln(z_1/z_0) + 0.4\sqrt{\beta}C(t_1)}{\ln(z_2/z_0)_{\text{open}} + 0.98C(t_2)_{\text{open}}} \right] \quad (26.7)$$

The examples in Table 26.3 serve to illustrate the use of this expression. The first line indicates the example case, the second through fifth rows describe the open terrain wind characteristics, the sixth through thirteenth rows describe the second terrain wind characteristics, and the last row gives the ratio of maximum speeds for the stated conditions and averaging times.

**Table 26.3** Example Conversions of Maximum Speeds for Different Conditions

Example	A	B	C	D	E Iteration #1	E Iteration #2	F	G
Terrain 2	open	open	open	open	open		open	open
$z$ (m)	10	10	10	10	10		10	10
$u$ (mph) <sup>a</sup>					90			90
$t$ (s)	60	3600	60	3600	40 <sup>b</sup>		3	40
$C(t)$	1.28	0	1.28	0	1.58		2.86	1.58
Terrain 1	coast	suburb	open	open	suburb		suburb	open
$z$ (m)	10	10	100	10	10		10	10
$u$ (mph) <sup>a</sup>				90	90 <sup>c</sup>	79	90	
$t$ (s)	3	3600	60	40 <sup>b</sup>	40 <sup>b</sup>	46	40	3
$C(t)$	2.86	0	1.58	1.58	1.58	1.49	1.58	2.58

$\beta$	6.5	5.25	6	6	5.25		5.25	6
$p$	0.82	1.15	1	1	1.15		1.15	1
$U_{t_1}(z_1)/U_{t_2}(z_2)$	1.39	0.81	1.42	1.31	0.88	0.86	0.56	1.19

<sup>a</sup>Wind speed in fastest mile.

<sup>b</sup>Averaging time calculated from  $3600/U_{FM}$ .

<sup>c</sup>Initial guess of fastest mile speed for the first iteration selected as open country value. In the second iteration, the 79 mph value was calculated from 90 mph multiplied by the velocity ratio of 0.88.

## 26.5 Pressure Coefficients and Load Factors

Pressure coefficients and load factors, such as terrain exposure factors and gust factors, are available in building codes. These factors must be consistent with the type of reference wind used in the code. Local cladding pressures in modern codes are significantly higher than those found in earlier codes because of the improved understanding of fluctuating loads on structures. Except in areas of positive mean pressures (the windward wall), there is no direct correlation between the local pressure and the occurrence of gusts in the approaching wind.

### Defining Terms

**Atmospheric boundary layer:** The lower part of the atmosphere where the wind flow is affected by the earth's surface.

**Cladding:** Parts of the exterior building surface which keep out the weather but are generally not considered part of the structural system, although they do transfer loads to the structural system.

**Extra-tropical cyclones:** Large-scale low-pressure systems which control most of the severe weather conditions and extreme winds in temperate regions.

### References

- Simiu, E. 1981. Modern developments in wind engineering: Part 1. *J. Eng. Struct.* 3:233–241.  
 Simiu, E. and Scanlan, R. H. 1985. *Wind Effects on Structures*, 2nd ed. John Wiley & Sons, New York.  
 Krayner, W. R. and Marshall, R. D. 1992. Gust factors applied to hurricane winds. *Bull. AMS* 73.

### Further Information

Information on types of physical model studies commonly performed can be obtained from *Wind Tunnel Modeling for Civil Engineering Applications*, Cambridge University Press, 1982, and ASCE Manual and Reports on Engineering Practice No. 67, *Wind Tunnel Model Studies of Buildings and Structures*, American Society of Civil Engineers, New York, 1987.

In addition to the references by Simiu listed above, a good general book on wind engineering is *The Designer's Guide to Wind Loading of Building Structures* by Cook (Butterworths, 1985). General articles, including a number of conference proceedings, are published in the *Journal of Industrial Aerodynamics and Wind Engineering*.

Ravindra, M.K., Shipp, J. G. "Earthquake Effects"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

- 27.1 Why Do Earthquakes Occur?
- 27.2 Characteristics of Earthquakes
- 27.3 Damage Mechanisms
- 27.4 Seismic Hazard Analysis
- 27.5 Earthquake Damage Surveys
- 27.6 Earthquake-Resistant Design

**M. K. Ravindra**

*EQE International*

**J. G. Shipp**

*EQE International*

This chapter gives a brief description of the causes and characteristics of earthquakes. The different sources of damage from earthquakes are discussed. Seismic hazard analysis performed to develop seismic design criteria is summarized. Earthquake damage surveys and the principles of earthquake-resistant design are described.

## 27.1 Why Do Earthquakes Occur?

---

Examples of the types of earthquakes are tectonic earthquakes, volcanic earthquakes, and reservoir-induced earthquakes. Of these, tectonic earthquakes are the most common. Plate tectonic theory explains the cause of some tectonic earthquakes. Earth's outermost part (called the lithosphere) consists of several large and fairly stable slabs called plates which are constantly moving with respect to each other. Collisions between adjacent plates result in interplate earthquakes. However, some major earthquakes have occurred within continental regions away from plate boundaries (e.g., the New Madrid earthquakes in 1811–1812, and the 1886 Charleston earthquake). These are called intraplate earthquakes. For a detailed discussion of the causes of earthquakes and the elements of seismology, the reader is referred to the book by Bolt [[1987](#)].

## 27.2 Characteristics of Earthquakes

---

Earthquakes are typically measured in terms of magnitude. Richter [[1958](#)] defined the magnitude of a local earthquake as the logarithm to base ten of the maximum seismic wave amplitude (in

microns) recorded on a Wood–Anderson seismograph located at a distance of 100 km from the earthquake epicenter. Magnitude is a measure of the energy released in an earthquake.

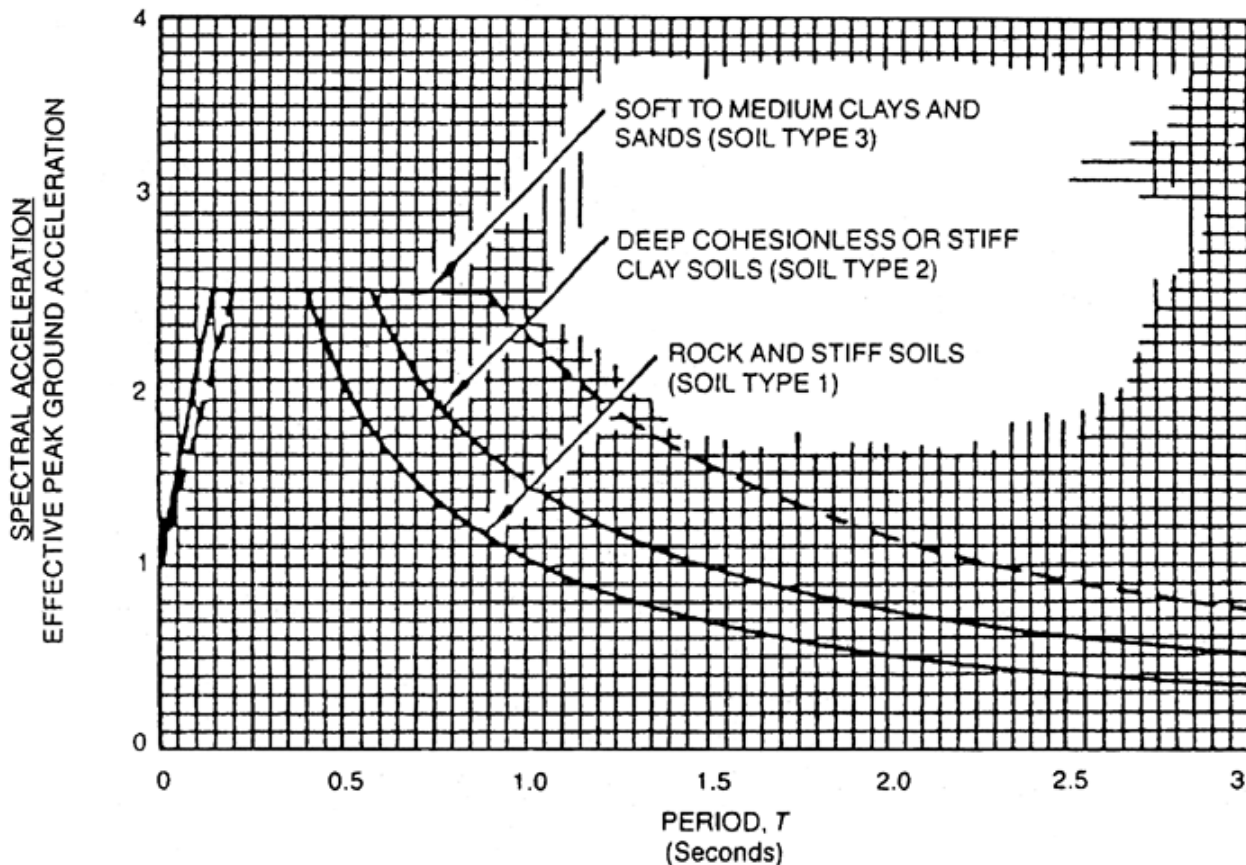
Earthquakes are also described in terms of an intensity measure. Intensity is the measure of damage to the structures and to the ground surface, and of the human response to the ground shaking. There are several intensity scales, the most common of which is the Modified Mercalli Intensity (MMI) scale of 1931 with a twelve-degree range from I to XII.

In engineering analysis, earthquake ground motion is described by the following characteristics:

1. Peak ground motion (peak ground acceleration, peak ground velocity, and peak ground displacement)
2. Duration of strong ground motion
3. Frequency content

A compact way of representing the ground motion characteristics is a **response spectrum**. It is a plot of the maximum elastic response of a series of one-degree-of-freedom structures to the given earthquake. Figure 27.1 shows the normalized response spectra shapes given in the Uniform Building Code [1994].

**Figure 27.1** Normalized response spectra shapes.



Building codes provide static and dynamic analysis methods for calculation of lateral earthquake forces to be resisted by structures. In the static analysis, earthquake forces are assumed to act in a horizontal direction against the structure. The Uniform Building Code gives the following formula for the total design base shear:

$$V = \frac{ZIC}{R_w}W$$

where

$$C = 1.25S/(T)^{2/3}$$

$Z$  = seismic zone factor

$R_w$  = ductility modification factor

$I$  = importance factor

$S$  = site coefficient for soil characteristics

$T$  = fundamental period of vibration of the structure in the direction under consideration

$W$  = total seismic dead load

The dynamic lateral force procedures given in the Uniform Building Code include response spectrum analysis and time-history analysis. The ground motion specified for these analyses is one having a 10% probability of being exceeded in 50 years and may be one of the following:

1. The normalized response spectrum
2. A site-specific response spectrum based on the geologic, tectonic, seismologic, and soil characteristics associated with the specific site
3. Ground motion time histories developed for the specific site representative of actual earthquake motions

## 27.3 Damage Mechanisms

---

Structures could suffer damage from earthquakes from one or more of the following mechanisms:

- Inertial forces generated by severe ground shaking
- Fire following earthquakes
- Soil failures (e.g., settlement, liquefaction, and landslides)
- Fault rupture
- Tsunamis

Of these mechanisms, damage by severe ground shaking is generally extensive and widespread. For example, the loss in the recent Northridge earthquake [EERI, 1994] was mostly by excessive ground shaking. Fire following an earthquake should be considered because an earthquake could initiate a fire and also destroy the mitigation systems, such as water distribution, transportation, communication, and emergency response.

Soil failures include settlement, liquefaction, and landslides. Differential settlement of buildings and its effect on nonstructural elements are of interest in the design. Liquefaction is a process by which sediments below the water table temporarily lose strength when subjected to earthquake shaking and behave as a viscous liquid. The effects of liquefaction include flow failures, lateral spreading, loss of bearing strength, differential settlement, and increased lateral pressure on retaining walls. Maps of regions with liquefaction potential have been developed. The engineer, developer, or building official should consult these maps to avoid building in these regions or provide mitigation measures to minimize damage from earthquakes.

Landslides and avalanches generated by earthquakes are localized and should be considered in zoning. Tsunamis are long water waves generated by sudden displacements underwater caused by fault rupture in a large earthquake. The wave roundup could exceed 20 meters and cause devastating damage to structures in the coastal regions around the Pacific rim.

## 27.4 Seismic Hazard Analysis

---

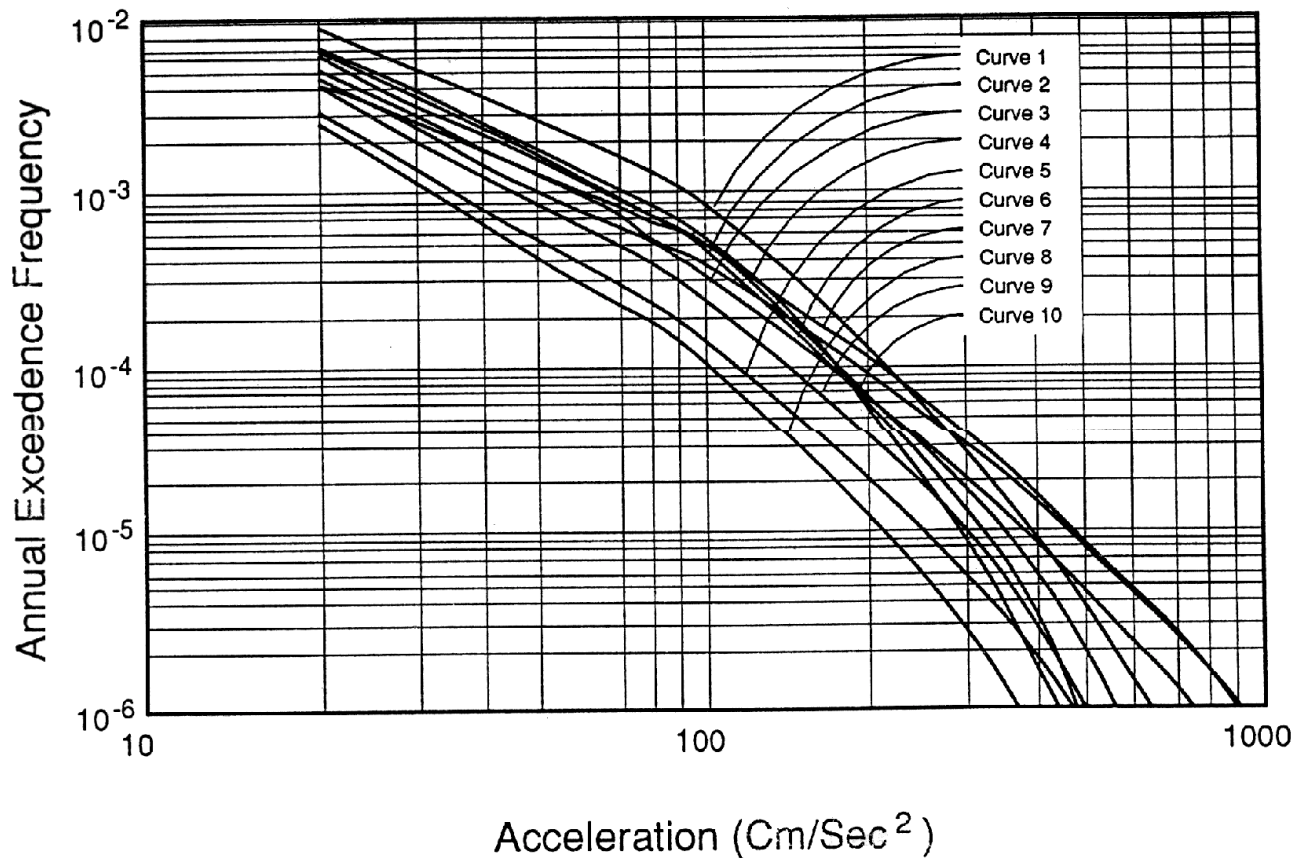
For design purposes, the engineer would like to know how often earthquakes occur near the facility and the potential severity of such earthquakes. The answers to these questions are obtained by performing a **seismic hazard analysis**. Seismic hazard is usually expressed in terms of the frequency distribution of the peak value of a ground motion parameter (e.g., peak ground acceleration, spectral velocity, and spectral acceleration) during a specified time interval. The different steps of this analysis are as follows:

1. Identification of the sources of earthquakes, such as faults and seismotectonic provinces
2. Evaluation of the earthquake history of the region to assess the frequencies of occurrence of earthquakes of different magnitudes or epicentral intensities
3. Development of attenuation relationships to estimate the intensity of earthquake-induced ground motion (e.g., peak ground acceleration) at the site
4. Integration of the above information to estimate the frequency of exceeding the selected ground motion parameter value

The hazard estimate depends on uncertain estimates of attenuation, upperbound magnitudes, and the geometry of the seismic sources. Such uncertainties are included in the hazard analysis by assigning probabilities to alternative hypotheses about these parameters. A probability distribution for the frequency of occurrence is thereby developed. The annual frequencies for exceeding specified values of the ground motion parameter are displayed as a family of curves with different probabilities (Fig. 27.2).



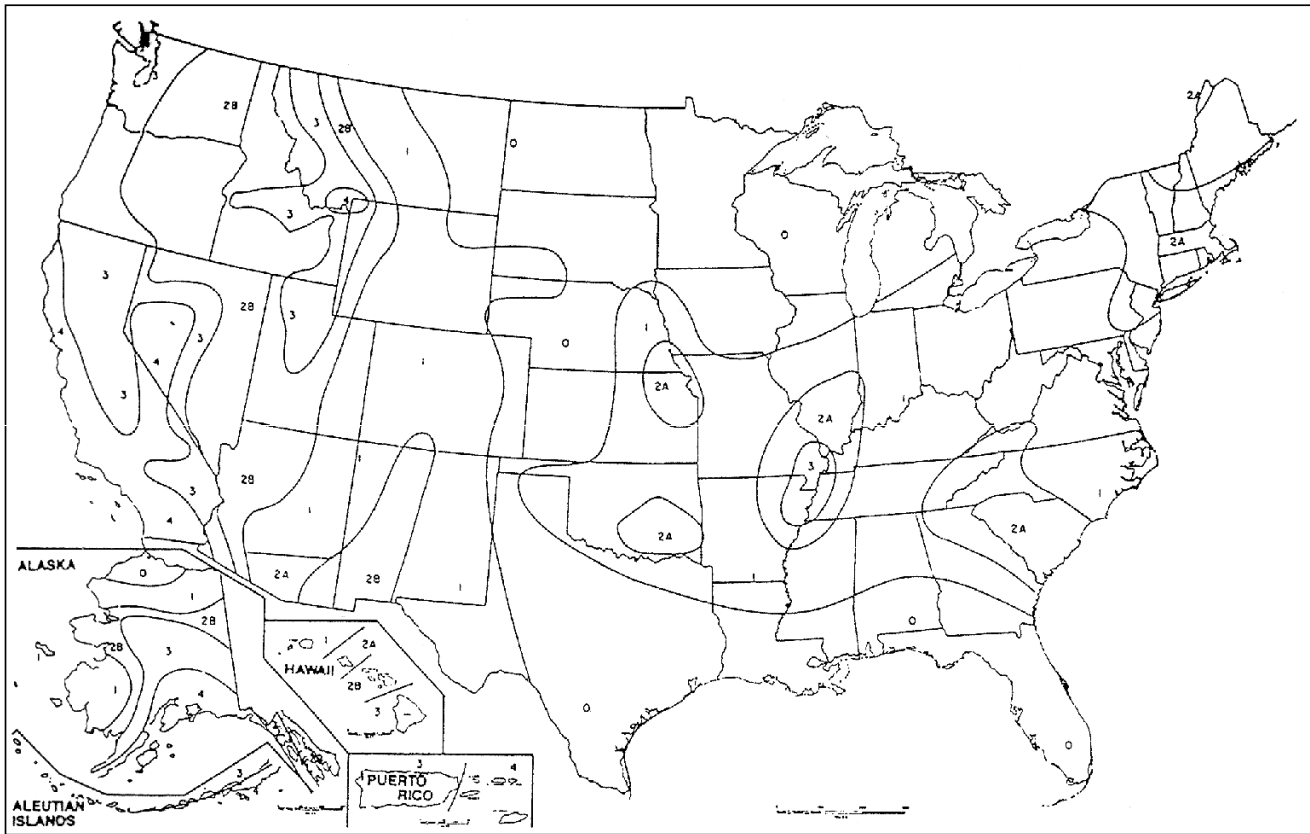
**Figure 27.2** Seismic hazard curves for a site



Two basic types of analytical models are available for probabilistic hazard analysis depending on whether faults or seismotectonic provinces are used for modeling the source of future earthquakes in the region. Many such methods have been developed over the last twenty years [Cornell, 1968; Bender and Perkins, 1987; Reiter, 1991; NEHRP, 1991]. Analyses based on seismotectonic provinces model the sources of future earthquakes as point sources that occur randomly over time and space in a given province. The occurrence rate over a source zone is considered constant and independent of occurrence in other zones. Such models are more appropriate where fault sources are not clearly identifiable (e.g., eastern U.S.). The fault models consider the location, length, and movement of faults, and are relevant for certain regions, such as California. However, the basic procedure is similar for both the models.

The above description of the seismic hazard analysis applies to a single site (e.g., nuclear power plant, chemical plant, dam, tall building). By modeling a region as consisting of a number of sites, this approach has been used to develop regional seismic hazard maps for different annual frequencies of exceedance. For example, the National Earthquake Hazard Reduction Program (NEHRP) has developed seismic hazard maps of spectral acceleration (at periods of 0.1, 0.3, and 1 second) for 10% probability of exceeding in 50 years, 100 years, and 250 years [NEHRP, 1991]. These maps have been adopted as part of the seismic design criteria for buildings (e.g., Uniform Building Code) (see Fig. 27.3).

**Figure 27.3** Seismic zone map of the U.S.



## 27.5 Earthquake Damage Surveys

There are two basic pre-earthquake damage surveys (PEDS) currently being offered by the structural engineering community. The first is a brief review of the available construction documents (i.e., drawings, specifications, geotechnical investigation reports, seismological reports, calculations) combined with a site visit to observe any obvious deficiencies and to compare the in situ conditions with the construction documents. The type of construction, age, and condition of the building are evaluated with the available documentation and field observations to determine the expected performance of the building during the postulated seismic event based on the known performance of similar structures for a similar seismic event. A common numerical technique is to express these findings as a probable maximum loss (PML). This is an indication of the probable maximum loss that the owner could expect to the property due to the postulated earthquake. For this type of survey, very few structural calculations are generated and only visually observable deficiencies are accounted for in the evaluation. This type of survey (sometimes referred to as a "Phase 1") is appropriate for regular buildings of a known construction type, age, and established track record of seismic performance.

The second type of PEDS requires a more in-depth review of the construction documents, a

thorough site visit (including some destructive or nondestructive testing of the structural materials and construction systems), and the generation of structural engineering calculations. These calculations are used to verify the existing structural systems and develop conceptual structural strengthening schemes to reduce the risk of damage due to a major earthquake. This type of survey (sometimes referred to as a "Phase 2") is appropriate for buildings of a complex geometry and unique or irregular structural configuration type with little established track record of seismic performance.

The post-earthquake damage survey, or seismic reconnaissance report (SRR), is one of the most informative and useful documents to help the structural engineer understand how buildings and structures actually behave when subjected to seismic forces [Yanev, 1991]. The findings from SRRs are used to develop changes to the building codes and determine additional areas of required research and development for future codes. The Northridge earthquake of January 17, 1994, was a watershed event for the understanding of the behavior of modern structures designed and built in accordance with the current standard of practice [EERI, 1994]. The lessons learned from past earthquakes are also used in identifying potential weaknesses in buildings and systems during pre-earthquake damage surveys.

## 27.6 Earthquake-Resistant Design

---

The documents available to the structural engineer for the design of earthquake-resistant structures can be classified into two distinct categories—voluntary codes which become municipal law when adopted by the governing jurisdiction, and guideline documents. The Uniform Building Code (UBC) published by the International Conference of Building Officials is the code that is used in most of the western states, the BOCA National Building Code published by the Building Officials and Code Administration International is used primarily in the eastern states, and the Standard Building Code (SBC) published by Southern Building Code Congress International is used primarily in the southern states. There has been discussion among these publishers to develop a single code. This would be a major achievement that would be welcomed by the structural engineering community.

The basic format of these codes is to present criteria for the design of seismic lateral forces using one of several prescriptive methods depending on the attributes of the building or structure. Regular structures (as defined by the code) are allowed to be designed using an equivalent lateral force method. This procedure allows a static lateral force to be applied to the structure to represent seismic forces. More complex or irregular structures are required to be designed using a dynamic analysis and a site specific spectrum. The most complex and important buildings or structures are required to be designed using a dynamic analysis and a series of site specific time histories. All of these code-defined procedures require that the demand on the structural elements be equal to or less than the allowable stress or strength as determined by prescriptive code equations.

The available guideline documents are too numerous to discuss in detail and consist of design criteria documents published by both public and private entities. The Federal Emergency Management Agency, the Department of Energy, the Department of Defense, the Nuclear Regulatory Commission, the Department of the Navy, and the Department of the Interior are a few

of the federal agencies that publish seismic design guidelines. The Division of the State Architect for the State of California, the City of Los Angeles, the City of San Francisco, and many other cities and counties throughout the U.S. publish seismic design criteria. Professional organizations, such as the Structural Engineers of California, the Applied Technology Council, and the American Society of Civil Engineers, continue to lead in the development and distribution of new seismic force and vertical load design criteria. Most of the current code or guideline documents are written in a prescriptive format with the protection of life as the primary focus. The direction of the future seismic design documents (i.e., the documents currently being prepared and scheduled to be implemented by the year 2000) is toward a performance-based criteria, with a probability-based limit states design philosophy.

## Defining Terms

**Equivalent lateral force method:** Procedure which allows a static lateral force to be applied to the structure to represent seismic forces.

**Response spectrum:** Plot of maximum elastic response of a series of one-degree-of-freedom structures to a given earthquake.

**Seismic hazard analysis:** Estimate of the annual frequencies of occurrence of different levels of ground motion at a site.

## References

- Bender, B. and Perkins, D. M. 1987. *SEISRISK III: A Computer Program for Seismic Hazard Estimation*. U.S. Geological Survey Bulletin 1772, Denver.
- Bolt, B. A. 1987. *Earthquakes*. W.H. Freeman, New York.
- Cornell, C. A. 1968. Engineering seismic risk analysis. *Bull. Seismological Soc. Am.* 58:1583–1606.
- Hall, J. (ed). 1994. *Northridge Earthquake, January 17, 1994: Preliminary Reconnaissance Report*. Earthquake Engineering Research Institute, Report 94-01.
- NEHRP. 1991. *NEHRP Recommended Provisions for the Development of Seismic Regulations for New Buildings*. Earthquake Hazard Reduction Series 16, Federal Emergency Management Agency, FEMA 222.
- Reiter, L. 1991. *Earthquake Hazard Analysis: Issues and Insights*. Columbia University Press, New York.
- Richter, C. F. 1958. *Elementary Seismology*. W.H. Freeman, San Francisco.
- Uniform Building Code. 1994. International Conference of Building Officials, Whittier, CA.
- Yanev, P. I. 1991. *Peace of Mind in Earthquake Country*. Chronicle Books, San Francisco.

Sexsmith, R. G., Cigic, T. M. "Structural Analysis"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 28

## Structural Analysis

---

- 28.1 Beams
- 28.2 Trusses
- 28.3 Frames
- 28.4 Computer-Aided Analysis

**Robert G. Sexsmith**

*University of British Columbia*

**Tony M. Cigic**

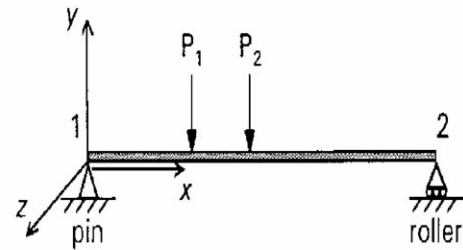
*University of British Columbia*

Structural analysis provides the key to the safe design of structures by creating reliable estimates of the internal forces, stresses, deformations, and behavior of structures under the influence of loads or imposed movements.

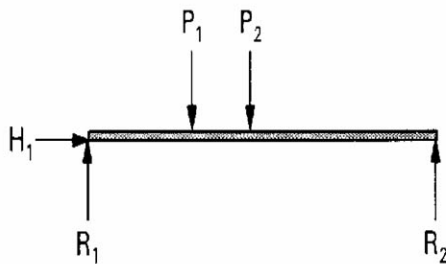
The starting point for any structural analysis is the *free-body diagram*, discussed in Section 1, Statics. Where the equilibrium equations are sufficient to determine all the forces on a free-body diagram—that is, where the number of unknowns does not exceed the number of independent equilibrium equations—the structure or segment is said to be **statically determinate**. Where this is not the case, the structure is **statically indeterminate**. In the indeterminate case, methods of analysis that include consideration of deformations are required.

Consider the beam of Fig. 28.1(a). The free-body diagram of Fig. 28.1(b) has three unknowns; thus equilibrium equations are sufficient to determine the three unknown reactions. With these now known, a cut is made through the beam at a distance  $x$  from the left end, shown in Fig. 28.1(c). Equilibrium equations can again be used, this time solving for the unknown internal bending moment,  $M(x)$ , shear,  $V(x)$ , and axial force,  $H(x)$ .

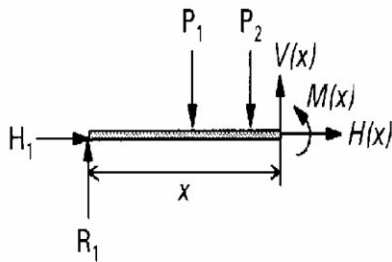
**Figure 28.1** Simply supported beam: (a) loads, (b) free-body diagram, (c) internal stress resultants.



(a)

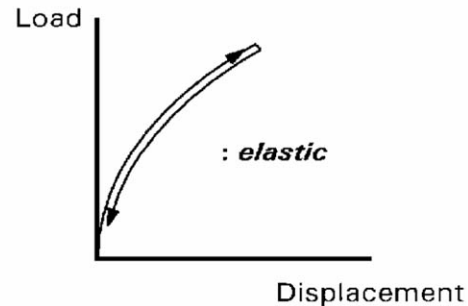


(b)

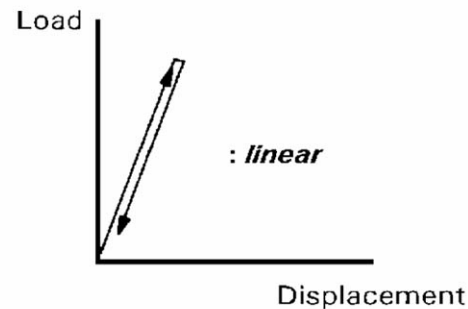


(c)

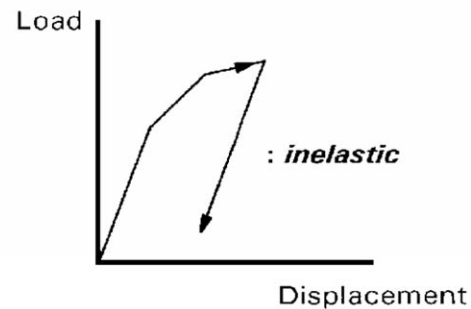
**Figure 28.2** Load-deflection curves: (a) elastic, (b) linear, (c) inelastic.



(a)



(b)



(c)

When a plot is made of load versus the deflection of some point on a structure, the structure is said to be *elastic* if the unloading curve is the same as the loading curve. When the loading and unloading curves are straight lines, the structure is said to be linear. Figure 28.2(a) illustrates an elastic load deflection curve, Fig. 28.2(b) a linear one, and Fig. 28.2(c) an *inelastic* curve. Unless otherwise indicated, most structural analysis methods are based on the assumption of linear elasticity.

When the load deflection response is linear, the **principle of superposition** holds. This means that the displacements resulting from each of a number of forces may be added to give the displacement resulting from the sum of the forces. Superposition applies equally to forces, stresses, strains, and displacements.

## 28.1 Beams

A beam spans two or more support points. The analysis of a beam usually consists of construction of the *load diagram*, *shear diagram*, and *moment diagram*, and calculation of deflections at key locations. (Examples of such analysis results appear later in Fig. 28.6.)

The beam of Fig. 28.1(a) is called a *simply supported* beam because it has a support only at its ends and is statically determinate.

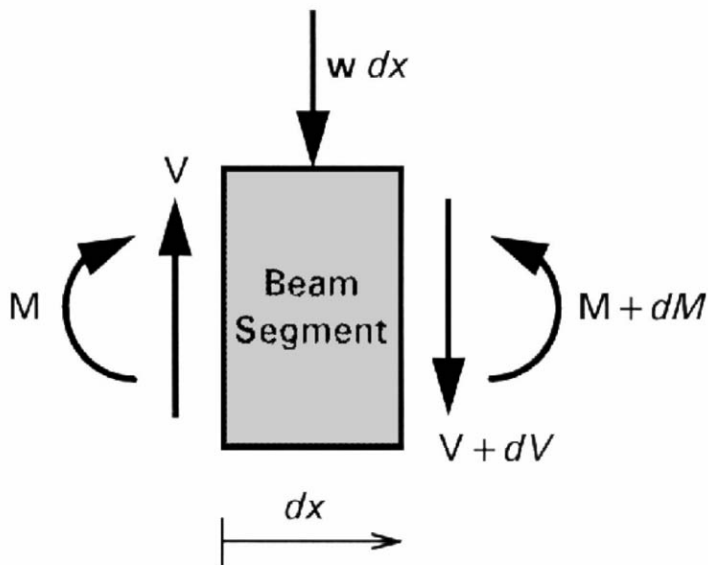
Inspection of the equilibrium of any short beam segment of length  $dx$ , as shown in Fig. 28.3, provides the useful relationships

$$w = -dV/dx \quad (28.1)$$

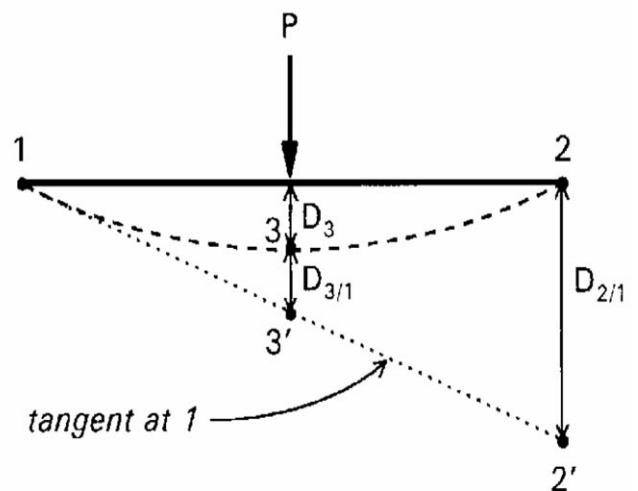
$$V = dM/dx \quad (28.2)$$

where  $w$  is the load intensity on the segment,  $V$  is the shear, and  $M$  is the bending moment.

**Figure 28.3** Elemental beam segment.



**Figure 28.4** Moment-area method applied to a simple beam.





With careful choice of free-body diagrams, Eqs. (28.1) and (28.2) facilitate the drawing of shear and moment diagrams for statically determinate beams.

When the span-to-depth ratio of a beam is in the normal range—greater than about 15—deformation and resulting deflections due to shear are relatively small. We can neglect such shear deformations and assume that only flexural deformations are important. Beam **deflections** may then be determined by establishing the relationships between bending moment and curvature. A useful way to compute beam deflections is by application of the *moment area theorems* [White *et al.*, 1976], based on relationships between bending moment and curvature in a beam. The first moment area theorem states that the change in slope from point 1 to point 2 on a beam equals the area of the  $M/EI$  diagram between points 1 and 2, or

$$\theta_{2/1} = \int M/EI \, dx \quad (28.3)$$

where  $E$  is the material modulus of elasticity and  $I$  is the cross-section moment of inertia.

The second moment area theorem states that the deflection of point 2 on a beam with respect to a tangent at point 1 is the moment of the  $M/EI$  diagram between points 1 and 2, taken about point 2, or

$$D_{2/1} = \int x M/EI \, dx \quad (28.4)$$

The key to application of the two theorems is the establishment of a known tangent and then computation of relative deflections with respect to that tangent. For example, the deflected shape of the beam in Fig. 28.4 has a tangent line 1-2'. The deflection  $D_{2/1}$  (distance 2-2') is first computed, thus establishing the tangent line. Deflection of any point on the beam, such as point 3, is then found by computing the deflection from the tangent line—that is,  $D_{3/1}$  (distance 3-3')—and then computing  $D_3$  from the fact that the sum of  $D_3$  and  $D_{3/1}$  can be found by proportion to  $D_{2/1}$ .

In the example beam of Fig. 28.4, a simple beam with concentrated load at center,

$$D_{2/1} = (PL/4EI)(L)(1/2)(L/2) = PL^3/16EI \quad (28.5)$$

$$D_{3/1} = (PL/4EI)(L/2)(1/2)(L/6) = PL^3/96EI \quad (28.6)$$

$$D_3 + D_{3/1} = D_{2/1}/2 = PL^3/32EI \quad (28.7)$$

$$D_3 = PL^3/32EI - PL^3/96EI = PL^3/48EI \quad (28.8)$$

Many beams are statically indeterminate. Consider the beam of Fig. 28.5(a). Statics alone cannot determine the three reactions  $X_1$ ,  $X_2$ , and  $X_3$ . We create a *primary structure* by removing the center support and turning the beam into a single span 3-2. We choose the reaction near midspan and define it as an unknown  $X_1$ . The deflection of this primary beam (which is statically determinate) due to the loads, corresponding to the location and direction of  $X_1$ , is denoted  $\delta_{10}$ , shown in Fig. 28.5(b). In this case  $\delta_{10}$  is shown in the figure in the opposite direction to the direction of  $X_1$ , and it is therefore shown having a negative value. The value  $\delta_{10}$  can be found by application of the moment area theorems. An additional load case on the same beam is shown in Fig. 28.5(c), consisting of a unit load corresponding to  $X_1$ . The resulting deflection is  $\delta_{11}$ , which may also be determined by the moment area method. The actual deflection at this location is zero; we can therefore write

$$\delta_1 = \delta_{10} + \delta_{11} X_1 = 0 \quad (28.9)$$

from which we can solve for  $X_1$ . Once this is determined, the remaining reactions can be determined from the equations of statics. Shear and moment diagrams can be drawn, and deflections can be determined from the now known moment diagram using the moment area method.

**Figure 28.5** Statically indeterminate beam analysis: (a) primary structure, (b) deflection due to applied loads, (c) deflection under unit load that corresponds to unknown reaction.

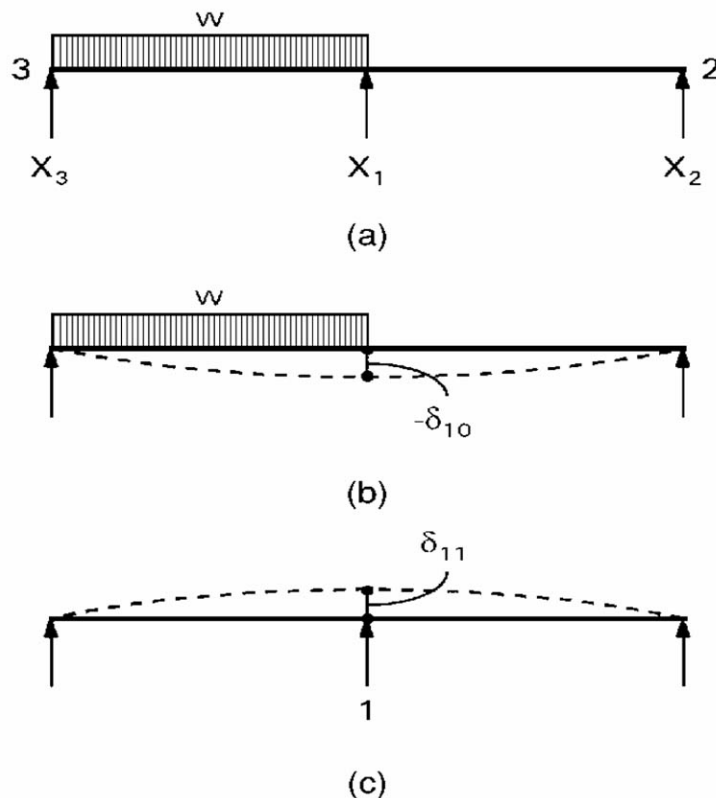
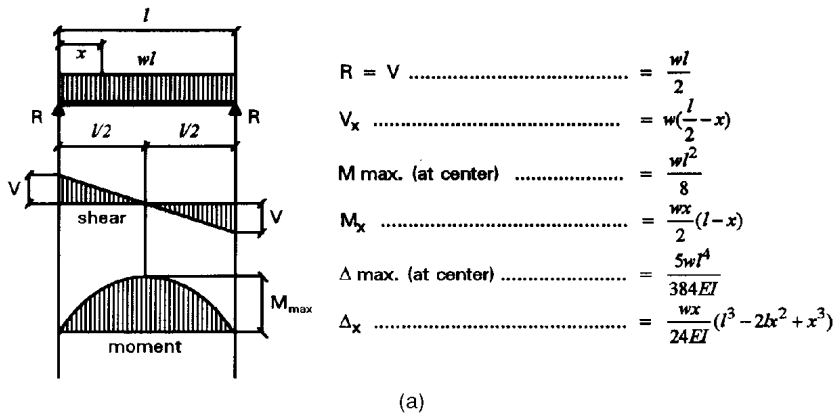


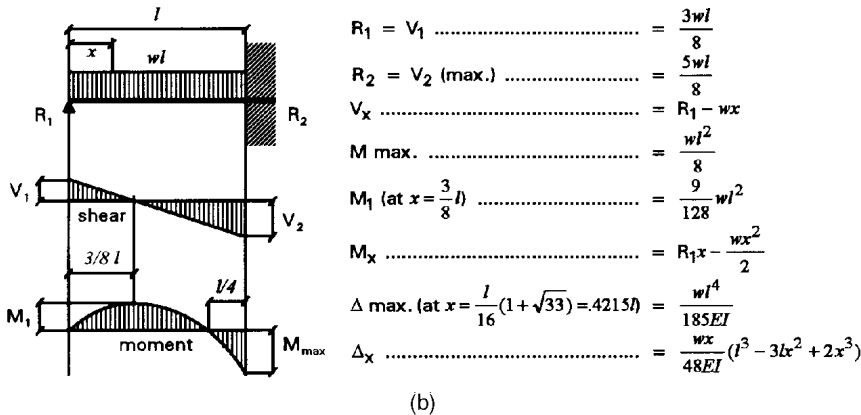
Figure 28.6 gives shear and moment diagrams and deflections for three commonly encountered beam cases. Similar and more extensive tables are widely available [AISC, 1986; Young, 1989].

**Figure 28.6** Shear and moment diagrams: (a) simple beam, (b) propped cantilever beam with uniform load, (c) propped cantilever with concentrated load.

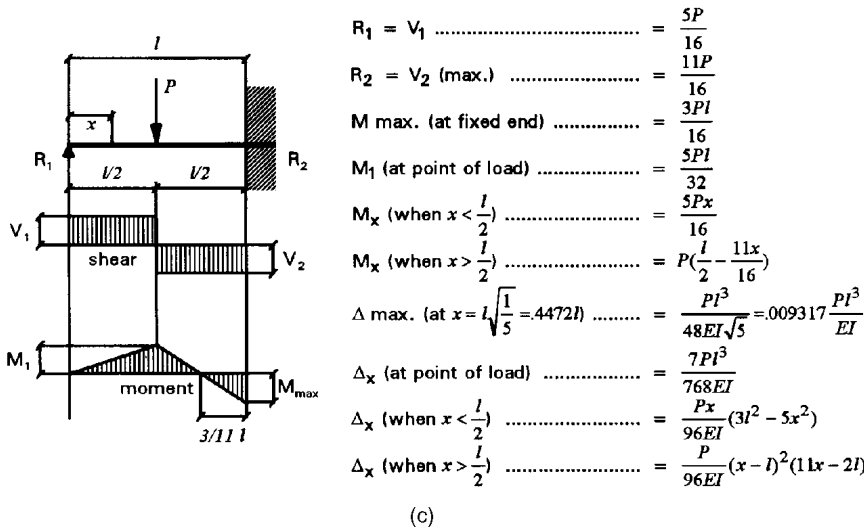
**SIMPLE BEAM: UNIFORMLY DISTRIBUTED LOAD**



**BEAM FIXED AT ONE END, SUPPORTED AT OTHER: UNIFORMLY DISTRIBUTED LOAD**



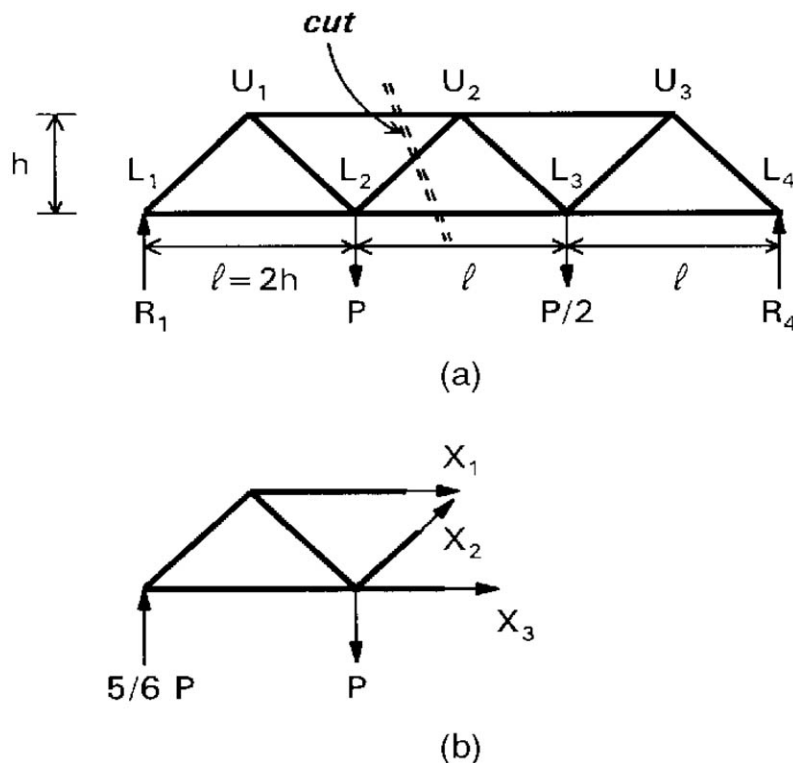
**BEAM FIXED AT ONE END, SUPPORTED AT OTHER: CONCENTRATED LOAD AT CENTER**



## 28.2 Trusses

Most trusses are assemblies of straight bars into a set of triangles. The joints are considered to be hinged and hence do not transmit moments. This is usually a very good assumption, even for trusses with near rigid joints. If we start with a single triangle, consisting of three bars and three joints, and form successive triangles by addition of two bars and one new joint for each new triangle, we have a **simple truss**, as in Fig. 28.7(a). If such a truss is supported in a statically determinate manner, then the truss is statically determinate and *stable*, and we can find all the bar forces using free-body diagrams and equilibrium. Trusses that do not meet the requirements for a simple truss are sometimes encountered. These may be *unstable* or statically indeterminate and this must be identified. Such cases are discussed in textbooks on structural analysis [White *et al.*, 1976].

**Figure 28.7** Simple truss: (a) imaginary "cut" through bars for which forces are required, (b) free-body diagram with the required unknown forces.



Structural analysis of a truss is aimed at the determination of the axial forces in all of the truss bars. The loading is assumed to consist only of concentrated loads applied at the joints of the truss. The analysis of a simple truss is usually carried out by first solving for the external reactions ( $R_1$  and  $R_4$  in Fig. 28.7), then defining a free-body diagram with cuts through the members where forces are required. This is called the **method of sections**. If the cut defines three or fewer unknowns, the three equilibrium equations are sufficient to determine the unknown forces. Because the truss bars are assumed hinged at their ends, the only unknown in a truss bar is the axial force. The truss of Fig. 28.7(a) is cut to define a free-body diagram in Fig. 28.7(b) that exposes several bar forces as unknowns  $X_1$ ,  $X_2$ , and  $X_3$ . It is usually possible to establish equations that involve only one or two unknowns. A consistent sign convention is adopted. In this example, tension is taken as positive. The solution proceeds as follows:

$$\begin{aligned}\Sigma M_{U_2} = 0: \quad & 5/6P(3h) - Ph - X_3(h) = 0 & (28.10) \\ & X_3 = 3/2P\end{aligned}$$

$$\begin{aligned}\Sigma M_{L_2} = 0: \quad & 5/6P(2h) + X_1(h) = 0 & (28.11) \\ & X_1 = -5/3P\end{aligned}$$

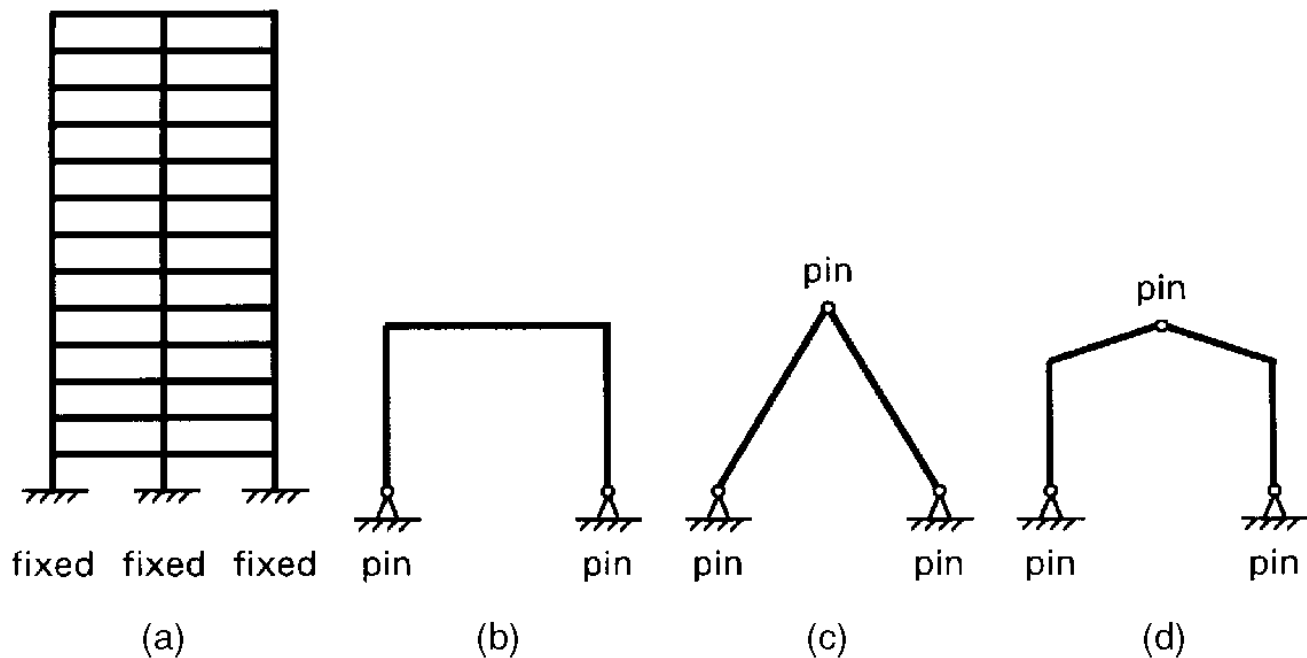
$$\begin{aligned}\Sigma F_y = 0: \quad & 5/6P - P + \sqrt{2}/2 X_2 & (28.12) \\ & X_2 = 1/3\sqrt{2}P\end{aligned}$$

An alternative to the method of sections is the **method of joints**. In this case a free-body diagram of each truss joint is drawn. Each joint has forces acting on it that correspond to each of the members framing into the joint. There are two equilibrium equations for each joint since the *concurrent* forces acting on the joint cannot have a moment resultant. For a statically determinate truss there will be  $m = 2j - 3$  members, where  $j$  is the number of joints. Thus the  $2j$  equilibrium equations are sufficient to solve for the  $m$  unknown member forces and the three external reactions. After first solving for the support reactions, it is usually possible to choose an analysis sequence that avoids or minimizes the need to solve simultaneous equations.

## 28.3 Frames

Frames are systems of bending members or beams connected together. For example, a building frame as in Fig. 28.8(a) may consist of several vertical columns and horizontal beams at each floor and roof level. The result is a rectangular array of members that carry axial forces, shear forces, and bending moments.

**Figure 28.8** Frames: (a) tall building frame, (b) portal frame, (c) statically determinate "A" frame, (d) statically determinate gable frame.



Frames are usually statically indeterminate; thus, we require advanced analysis methods such as moment distribution [White *et al.*, 1976] or one of the computer-based methods. Fig. 28.8(b) shows a statically indeterminate portal frame. There are also statically determinate frames, usually single-story frames with three hinges, as shown in Figs. 28.8(c) and 28.8(d).

Usually bending dominates the behavior of low frames, and the effects of axial forces are neglected in the analysis. It is significant to note that the reactions generally include horizontal reactions at the base supports.

## 28.4 Computer-Aided Analysis

The routine aspects of structural analysis can be performed by digital computers after the assumptions as to material behavior, member properties, geometry, and loading have been made. The **direct stiffness method** [White *et al.*, 1976] is the most widely used approach for the analysis of linear elastic structures. The structure is defined by a number of *nodes* connected by members. The members are usually assumed straight between the nodes. Each of the nodes can, in general, displace. The displacements in directions defined by the coordinate system are the *degrees of freedom* of the structure. For a two-dimensional structure there are in general two displacements and one rotation for each node. The direct stiffness method is a *displacement method* because it aims to determine the displacements corresponding to the degrees of freedom of the structure. The direct stiffness method is based on the matrix equilibrium equation

$$KD = Q \quad (28.13)$$

where  $K$  is the **stiffness matrix**, and  $K_{ij}$  is the force corresponding to degree of freedom  $i$  caused by a unit displacement corresponding to degree of freedom  $j$ . Thus  $K$  is a matrix of stiffness-influence coefficients. The stiffness matrix is symmetric and positive definite. It becomes nonpositive definite if the structure is unstable.  $D$  is a vector of the nodal displacements corresponding to the degrees of freedom of the structure, and  $Q$  is the vector of applied loads corresponding to the degrees of freedom. If there are  $n$  degrees of freedom, the stiffness matrix  $K$  is a  $n \times n$  array, and  $D$  and  $Q$  are  $n \times 1$  arrays.

$K$  is first assembled from the member properties. Equation (28.13) is then solved for  $D$ . Once the displacements are known, the member properties can be utilized to determine internal forces.

Although the method is explained here for framed structures, the same principles can be applied to plate and shell structures, where the "members" are then **finite elements**.

The computer programs currently available provide a rapid format for entering the required data and obtaining results. The data input starts by defining the nodes and their coordinates, assigning consecutive numbers to the nodes and defining restraints to displacement where there are supports. Then the members that connect the nodes are defined. Required member properties include area, moment of inertia, and modulus of elasticity. Loadings are then defined in their coordinate directions. The programs typically provide output in the form of internal member forces, and displacements at the nodes in the coordinate directions.

There are many commercially available programs that perform structural analysis. They include SAP90 and ETABS from Computers and Structures, Inc., Berkeley, CA 94704, and Staad-III from Research Engineers, Inc., Yorba Linda, CA 92687.

While static analysis of linear structures is the most common application of the stiffness method, it extends to dynamic analysis [Clough and Penzien, 1975] and analysis of nonlinear structures. The analysis results depend entirely on the quality of the assumptions used to define the problem; they should therefore be used only by structural engineers who have a full knowledge of the modeling issues that must be addressed in order to achieve a valid analysis.

## Defining Terms

**Deflection:** The movement of a point on a structure to its stressed position from the unstressed position.

**Direct stiffness method:** A method of analysis that uses stiffness-influence coefficients as the terms in a set of equilibrium equations to determine the unknown displacements that correspond to the degrees of freedom of a structure.

**Finite element:** An element, usually a small segment of a plate or shell, that plays the role of a member joining several nodes in the direct stiffness method.

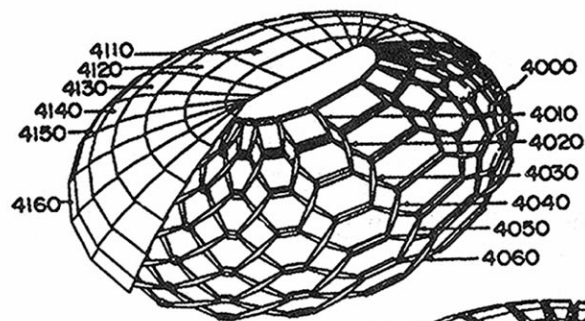


FIG. 31

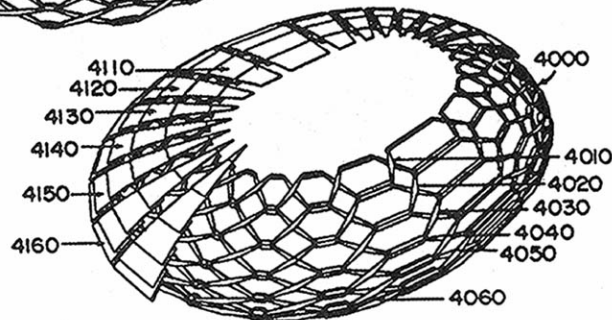


FIG. 32

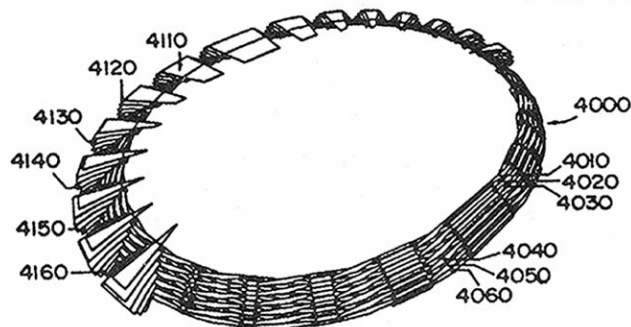


FIG. 33

## RADIAL EXPANSION/RETRACTION TRUSS STRUCTURES

*Charles Hoberman*

*Patented June 18, 1991*

*#5,024,031*

An excerpt:

I have discovered a method for constructing reversibly expandable truss-structures that provides for an extremely wide variety of geometries. Trusses formed by this method will collapse and expand in a controlled, smooth and synchronized manner. Such structures require no complex joints. Connections are limited to simple pivots. A unique characteristic of one embodiment of the present invention is that it provides a three-dimensional folding truss whose overall shape and geometry is constant and unchanging during the entire folding process. Only its size changes between a compact bundle and an extended self-supporting structure.

Based on simple scissors tongs from childhood, Hoberman envisions a retractable stadium roof could be built using his invention. He and his structures were recently featured in Discovery magazine (©1993, DewRay Products, Inc. Used with permission.)



**Linear:** The stress-strain or load deflection relationship is a straight line for loading and for unloading. This allows the principle of superposition to hold.

**Method of joints:** An analysis method for trusses in which each joint is successively isolated, applying conditions of equilibrium to determine the unknown forces.

**Method of sections:** An analysis method for trusses in which a section is cut through the truss, introducing a number of unknowns less than or equal to the number of available equations of equilibrium.

**Principle of superposition:** The individual displacements resulting from each of several load cases can be added to give the displacement corresponding to the sum of the loads.

**Simple truss:** A truss made up of triangles, formed from an initial triangle of three bars and three joints by successively adding two bars and one joint to the existing form. Such a truss is statically determinate and stable.

**Statically determinate:** A part of a structure for which the equations of equilibrium are sufficient to determine the unknown forces.

**Statically indeterminate:** A part of a structure for which compatibility of deformations is required in addition to equilibrium conditions to determine the unknown forces.

**Stiffness matrix:** The matrix of stiffness-influence coefficients in the equilibrium equations of the direct stiffness method.

## References

AISC. 1986. *Manual of Steel Construction—Load and Resistance Factor Design*, 1st ed. American Institute of Steel Construction, Chicago.

Clough, R. W. and Penzien, J. 1975. *Dynamics of Structures*. McGraw-Hill, New York.

Young, W. C. 1989. *Roark's Formulas for Stress and Strain*, 6th ed. McGraw-Hill, New York.

White, R. W., Gergely, P., and Sexsmith, R. G. 1976. *Structural Engineering, Combined Edition*. John Wiley & Sons, New York.

## Further Information

In addition to the references cited, the reader is encouraged to refer to one of the study guides for the structural engineering license examinations, such as *246 Solved Structural Engineering Problems*, by Dale Buckner, Professional Publications Inc., Belmont, CA. The publications of the American Society of Civil Engineers, New York, provide a rich source of additional detailed information on structural analysis.

Segui, W. T. "Structural Steel"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## 29.1 Members

Tension Members • Compression Members • Beams • Beam-Columns

## 29.2 Connections

Bolts • Welds

## 29.3 Composite Members

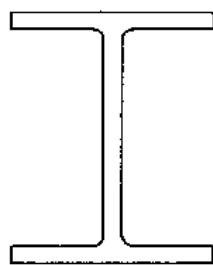
## 29.4 Computer Applications

### William T. Segui

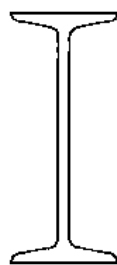
University of Memphis

Structural steel is used for the framework of buildings and bridges, either alone or in combination with other materials such as reinforced concrete. Steel buildings are usually constructed with standard shapes produced by hot-rolling (Fig. 29.1), although custom shapes can be fabricated from plate material. Various grades of steel, as classified by the American Society for Testing and Materials [ASTM, 1994], are suitable for building and bridge construction. The most commonly used steel is ASTM A36, with a minimum tensile yield stress  $F_y$  of 36 ksi and an ultimate tensile stress  $F_u$  between 58 and 80 ksi. (The actual yield stress of most A36 steel currently being produced is close to 50 ksi.)

**Figure 29.1** Examples of standard hot-rolled shapes (cross-sectional views).



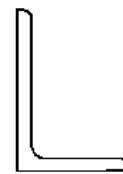
Wide flange  
(W-shape)



American Standard  
(S-shape)



Channel  
(C-shape)



Angle  
(L-shape)

Several design approaches are available to the structural engineer. In *allowable stress design*, (ASD), structural members are proportioned so that the maximum computed stress is less

than a permissible, or allowable, stress. This approach is also called *working stress design* or *elastic design*.

In *load and resistance factor design*, (LRFD), members are proportioned so that the resistance (strength) of the member is greater than the applied load. This approach can be represented by the following relationship:

$$\sum \gamma_i Q_i \leq \phi R_n \quad (29.1)$$

where  $\gamma_i$  is a **load factor**,  $Q_i$  is a load effect (force or moment),  $\phi$  is a **resistance factor**, and  $R_n$  is the nominal resistance, or **nominal strength**. The summation indicates that the total factored load effect is the sum of the products of individual load effects (such as dead, live, and snow) and corresponding load factors—which are a function of not only the type of load effect, but also the combination of loads under consideration. The nominal strength is a theoretical strength, and the resistance factor reduces it to a practical value. This reduced value,  $\phi R_n$ , is called the **design strength**. Equation (29.1) states that the sum of the factored load effects must not exceed the design strength.

A third approach, *plastic design*, also uses load factors but is primarily a structural analysis method of obtaining failure loads by a consideration of collapse mechanisms.

Although all three approaches are acceptable, the current trend is toward load and resistance factor design, in part because it can result in a more efficient use of material. As a consequence, the focus of this part of the handbook will be on the LRFD approach.

The design of structural steel buildings in the U.S. is usually based on the provisions of the specification of the American Institute of Steel Construction [AISC, 1993] and the *Manual of Steel Construction* [AISC, 1994]. All of the requirements covered herein will be based on the AISC specification.

The AISC specification gives the loading conditions to be investigated in conjunction with Eq. (29.1). These load combinations, along with the associated load factors, are the same as those in ASCE 7-93 [ASCE, 1994]. The value of the resistance factor depends upon the type of member or connecting element being investigated and will be covered in the following sections.

## 29.1 Members

---

### Tension Members

Tension members are used in trusses, bracing systems, and in building and bridge suspension systems. The load and resistance factor relationship of Eq. (29.1) can be expressed in the following way for a tension member:

$$P_u \leq \phi_t P_n \quad (29.2)$$

where  $P_u$  is the sum of the factored axial tension loads,  $\phi_t$  is the resistance factor for tension, and  $P_n$  is the nominal tensile strength.

There are two possible failure modes, or **limit states**, for tension members: yielding of the gross cross section and fracture of the net cross section. The net cross-sectional area is the gross area minus any area removed by bolt holes. For yielding of the gross section, the design strength is given by

$$\phi_t P_n = \phi_t F_y A_g = 0.90 F_y A_g \quad (29.3)$$

and the design strength based on fracture of the net section is

$$\phi_t P_n = \phi_t F_u A_e = 0.75 F_u A_e \quad (29.4)$$

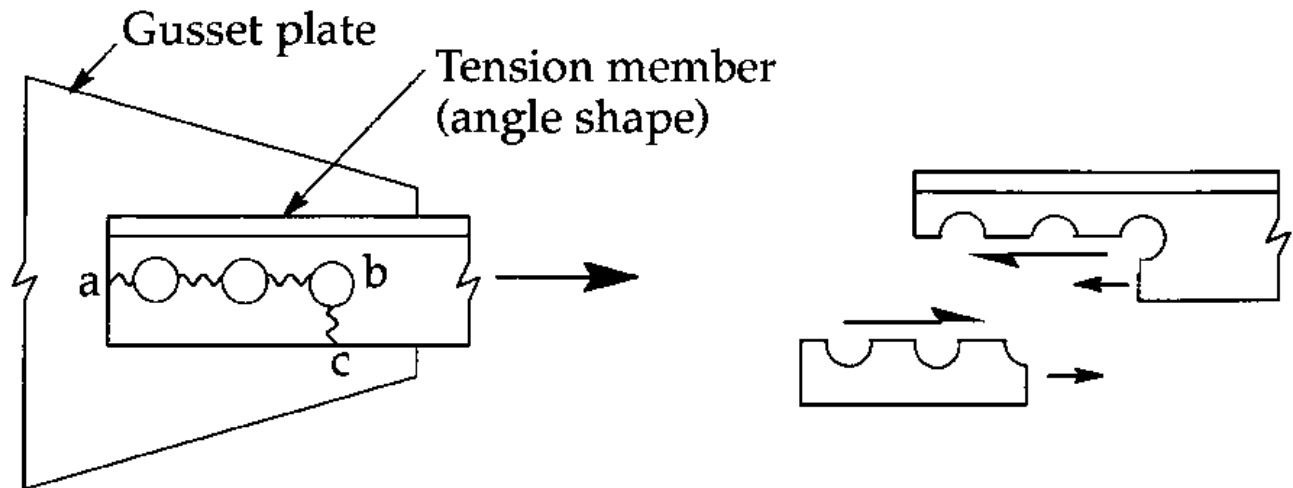
where  $A_g$  is the gross cross-sectional area and  $A_e$  is the *effective* net area. Because of a phenomenon called **shear lag**, an effective net area must be used in those cases where some of the cross-sectional elements are unconnected. The effective net area is given by

$$A_e = U A_n \quad (29.5)$$

where  $A_n$  is the actual computed net area and  $U$  is a reduction factor.

**Block shear** is another potential limit state for tension members. This failure mode must be investigated when the member is connected in such a way that a block of the material could tear out as in Fig. 29.2. Two loading conditions are involved: shear and tension. In the case illustrated, the shear is along line  $ab$  and the tension is along line  $bc$ . The strength, or resistance, is the sum of two contributions—either shear yielding plus tension fracture or shear fracture plus tension yielding. The governing case will be the one that has the larger fracture component.

**Figure 29.2** Block shear in a tension member.



## Compression Members

Compression members are found in trusses and as vertical supports in buildings and bridges, where they are usually referred to as columns. This discussion will be limited to *axially loaded* compression members. For a slender axially loaded compression member with pinned ends, the nominal strength is given by the Euler formula as

$$P_n = \frac{\pi^2 EI}{L^2} \quad (29.6)$$

where  $E$  is the modulus of elasticity,  $I$  is the moment of inertia about the minor principal axis of the member cross-sectional area, and  $L$  is the length. This equation can also be expressed in the following form:

$$P_n = \frac{\pi^2 E}{(L/r)^2} \quad (29.7)$$

where  $r$  is the radius of gyration and  $L/r$  is the **slenderness ratio**. For other end conditions,  $L$  can be replaced by an effective length  $KL$ , where  $K$  is the effective length factor.

AISC uses a modified form of the Euler formula for slender compression members and an empirical equation for nonslender members. The axial compressive design strength is  $\phi_c P_n$ , where  $\phi_c = 0.85$ .

The type of buckling just discussed—that is, buckling about one of the principal axes—is called **flexural buckling**. Other modes of failure include the following:

- **Torsional buckling.** Twisting without bending. This can occur in doubly symmetrical cross sections with slender elements (none of the standard hot-rolled shapes are subject to this failure mode).
- **Flexural-torsional buckling.** A combination of bending and twisting. Unsymmetrical cross sections are susceptible to this type of failure.
- **Local buckling.** Localized buckling of a cross-sectional element such as a web or projecting flange.

## Beams

The flexural design strength of a beam is  $\phi_b M_n$ , where  $\phi_b = 0.90$  and  $M_n$  is the nominal flexural strength, which is the bending moment at failure. The following discussion will be limited to beams bent about one principal axis. The nominal flexural strength is based on one of the following limit states:

- *A fully yielded cross section.* If a beam is prevented from becoming unstable in any way, the resisting moment will be equal to the internal couple corresponding to a uniform compressive stress of  $F_y$  on one side of the neutral axis and a uniform tensile stress of  $F_y$  on the other side.

This is the **plastic moment**,  $M_p$

- *Lateral-torsional buckling*. If a beam bent about its major principal axis is not adequately supported laterally (that is, in the direction perpendicular to the plane of bending), it can buckle outward, simultaneously bending about its minor principal axis and twisting about its longitudinal axis.
- *Flange local buckling*. This can occur in the *compression* flange if it is too slender.
- *Web local buckling*. This can occur in the compressed part of the web if it is too slender.

Cross sections can be categorized as compact, noncompact, or slender, depending upon the width-to-thickness ratios of their cross-sectional elements. Most of the standard hot-rolled shapes are compact, and only those will be considered here. Neither flange local buckling nor web local buckling are potential limit states for compact shapes. Furthermore, if a beam has adequate lateral support, lateral-torsional buckling will not occur, and the nominal strength is equal to the plastic moment; that is,

$$M_n = M_p \quad (29.8)$$

This will be the case when the distance between points of lateral support, called the *unbraced length*, is less than a prescribed value. If the unbraced length is too large, failure will be by either elastic lateral-torsional buckling or inelastic lateral-torsional buckling, depending upon whether yielding has begun when the buckling takes place.

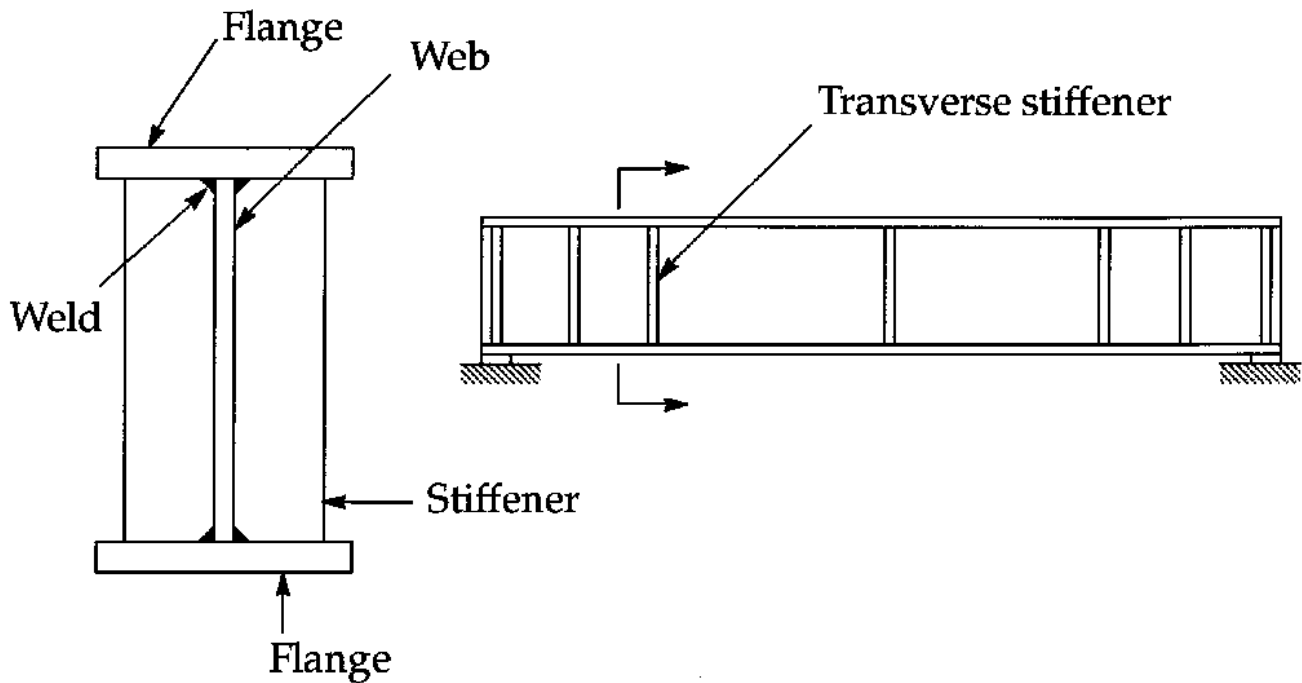
In addition to flexure, beams must be checked for shear strength (which usually does not control) and deflections. Deflections should be computed with service, and not factored, loads.

Flexural members with slender webs are classified by AISC as plate girders; otherwise, they are classified as beams. Plate girders (Fig. 29.3) are built up from plate elements that are welded together, and their flexural strength is based on one of the following limit states:

- Tension flange yielding
- Compression flange yielding
- Compression flange buckling, which can take the form of either lateral-torsional buckling or flange local buckling

Because plate girders have slender webs, shear strength is often critical. Part of the shear resistance can come from **tension-field action**, which relies on the post-buckling strength of the web. This component of the shear strength is a function of the spacing of transverse stiffeners, illustrated in Fig. 29.3.

**Figure 29.3** Plate girder details



## Beam-Columns

Many structural components, particularly those in continuous frames, are subjected to significant amounts of both bending and axial load—either tension or compression. If the axial load is compressive, the member is referred to as a *beam-column*. Because compression adds stability problems, bending plus compression is usually more serious than bending plus tension.

Combined loading can be accounted for by the use of interaction equations of the form

$$\frac{P_u}{\phi_c P_n} + \frac{M_u}{\phi_b M_n} \leq 1.0 \quad (29.9)$$

where each term on the left side is a ratio of a factored load effect to the corresponding design strength. The AISC specification uses two interaction equations: one for small axial loads and one for large axial loads. Furthermore, each equation has two bending terms: one for major axis bending and one for minor axis bending.

In addition to bending moment caused by transverse loads and end moments, beam-columns are subjected to secondary moments resulting from the eccentricity of the load with respect to the deflected member axis. This can be accounted for in an approximate, but very accurate, way through the use of **moment amplification** factors as follows:

$$M_u = B_1 M_{nt} + B_2 M_{lt} \quad (29.10)$$

where  $B_1$  and  $B_2$  are amplification factors,  $M_{nt}$  is the factored load moment corresponding to no



joint translation, and  $M_{lt}$  is the factored load moment corresponding to lateral joint translation. If the member is part of a braced frame, then no joint translation is possible and  $M_{lt} = 0$ . If the member is part of an unbraced frame,  $M_{nt}$  is computed as if the member were braced against joint translation, and  $M_{lt}$  is computed as the result of only the joint translation.

## 29.2 Connections

---

Modern steel structures are connected with bolts, welds, or both. Although hot-driven rivets can be found in many existing structures, they are no longer used in new construction.

### Bolts

Two types of bolts are used: common and high-strength bolts. Common bolts conform to ASTM A307 and are used in light applications. High-strength bolts are usually ASTM A325 or A490. Bolts can be loaded in tension, shear, or both. In addition, bearing stresses act between the bolts and the connected elements of shear connections. Although bearing is a problem for the connected part and not the bolt itself, the bearing load is a function of the bolt diameter and is therefore associated with the bolt strength. Thus, we may speak of the design strength of a bolted connection being based on shear, bearing, or tension. The shear or tension design strength of a single bolt is  $\phi R_n$ , where  $\phi = 0.75$  and

$$R_n = F_n A_b \quad (29.11)$$

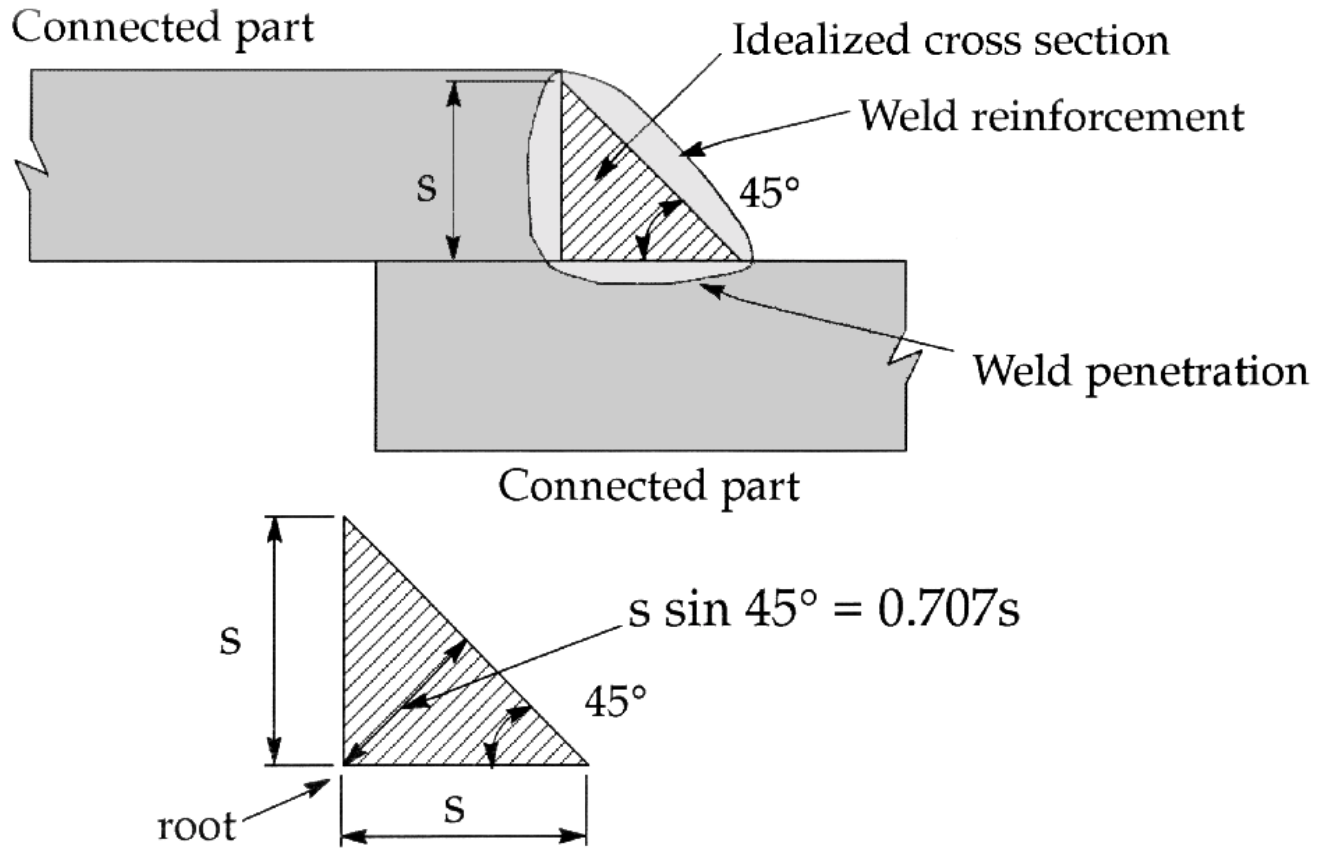
where  $F_n$  is the nominal ultimate shearing or tensile stress and  $A_b$  is the cross-sectional area of the bolt. Bearing strength depends upon such things as bolt spacing and edge distance and is a function of the *projected* area of contact between the bolt and the connected part.

High-strength bolts can be snug tight, where the bolts are installed by a single worker with an ordinary wrench, or fully tensioned, where they are installed to a prescribed minimum tension of about 70% of the ultimate strength. Fully tensioned bolts are used in slip-critical connections, which are those in which slippage is not permitted. Slip-critical connections are required for members subjected to impact or fatigue loading. A reduced value of  $F_n$ , treated as a shearing stress, is used in Eq. (29.11) to compute the slip-critical strength.

### Welds

Weld material is available in several grades, but E70 or E80 series electrodes are usually used. E70 electrodes have a tensile strength of 70 ksi and are designed to be used with steel having a yield stress between 36 ksi and 60 ksi. E80 electrodes have a tensile strength of 80 ksi and are used with steels having a yield stress of 65 ksi. Welding can be performed on the job site (field welds) or in the fabricating shop (shop welds). Because of automation, shop welding is generally more economical and of higher quality. The most common types of welds are the *groove weld*, in which the weld metal is deposited into a gap, or groove, and the *fillet weld* (Fig. 29.4), which is deposited into a corner. The fillet weld is the type most often used for field welding.

**Figure 29.4** Fillet weld properties.



The strength of a fillet weld is based on the premise that failure will occur by shear on a plane through the throat of the weld. The throat is the perpendicular distance from the root to the hypotenuse on the theoretical cross section of the weld, which is treated as an isosceles right triangle (Fig. 29.4). The design strength of the weld is  $\phi R_n$ , where  $\phi = 0.75$  and

$$R_n = \text{area} \times \text{ultimate shearing stress} = 0.707sLF_W \quad (29.12)$$

In Eq. (29.12),  $s$  is the weld size,  $L$  is the length, and  $F_W$  is the ultimate shearing stress, equal to 60% of the ultimate tensile stress of the electrode.

In many connections it is advantageous to use both welds and bolts, with all welding done in the shop and all bolting done in the field. This will usually be the most economical arrangement, since bolting requires less skilled labor.

## 29.3 Composite Members

Composite members are structural steel shapes acting in concert with attached reinforced concrete. The most common application is a set of parallel steel beams connected to and supporting a reinforced concrete floor slab. The connection is made by attachments that are welded to the top

flange of the beam and embedded in the slab. These attachments, called **shear connectors**, are usually in the form of headed studs. A portion of the slab is considered to act with each steel beam as a supplementary compression flange.

The flexural strength of a composite beam is computed in one of two ways:

1. If the web is compact, the strength is based on the plastic condition. This is when the steel has fully yielded and the concrete has reached its maximum compressive stress of  $0.85f_c'$ , where  $f_c'$  is the 28-day compressive strength.
2. If the web of the steel shape is noncompact, the strength is based on the limit state corresponding to the onset of yielding of the steel.

Deflections of composite beams are computed by the usual elastic methods but are computed for the transformed section, in which a consideration of strain compatibility is used to transform the concrete into an appropriate amount of steel. The resulting cross section can then be treated as a homogeneous steel shape.

The composite column is a combination of materials in which a structural steel shape is encased in concrete and supplemented by vertical reinforcing bars. Structural tubes or pipes filled with concrete are also used. The axial compressive strength of a composite column is computed in essentially the same way as for an ordinary steel shape but with modified values of the yield stress, modulus of elasticity, and radius of gyration.

## 29.4 Computer Applications

---

Many commercial computer programs are available to the structural steel designer. These include standard structural analysis programs for statically indeterminate structures as well as those containing an AISC "code-checking" feature. These are available for both ASD and LRFD versions of the AISC specification. The American Institute of Steel Construction markets computer software including connection design programs and a database of standard steel shapes for use with spreadsheet software or for those wishing to write their own computer programs.

### Defining Terms

**Block shear:** A limit state in which a block of material is torn from a member or connecting element, such as a gusset plate. When this occurs, one or more surfaces of the block fail in shear, either by fracture or yielding, and another surface fails in tension, either by fracture or yielding.

**Design strength:** The nominal strength, or resistance, multiplied by a resistance factor. This is a reduced strength that accounts for uncertainties in such things as theory, material properties, and workmanship.

**Flexural buckling:** A mode of failure of an axially loaded compression member where buckling occurs by bending about one of the principal axes of the cross section.

**Flexural-torsional buckling:** A mode of failure of an axially loaded compression member in which it simultaneously buckles about one of the principal axes of its cross section and twists about its longitudinal axis.

**Lateral-torsional buckling:** A limit state in which a beam buckles by deflecting laterally and twisting. This is prevented by providing lateral bracing at sufficiently close spacing.

**Limit state:** A failure condition upon which the strength of a member is based. Yielding of a tension member is an example of a limit state, and the axial tensile force causing the yielding is the corresponding strength.

**Load factor:** A multiplier of a load effect (force or moment) to bring it to a failure level. A load factor is usually greater than unity, although it can be equal to unity. It is a safety factor that is applied to loads.

**Local buckling:** A localized buckling, or wrinkling, of a cross-sectional element. This form of instability is in contrast to overall buckling, as when a compression member buckles by bending.

**Moment amplification:** An approximate technique used to account for the secondary bending moment in beam-columns. The total moment is obtained by multiplying the primary moment by a moment amplification factor.

**Nominal strength:** The theoretical strength of a member before reduction by a resistance factor.

**Plastic moment:** The bending moment necessary to cause yielding throughout the depth of a given cross section. There will be a uniform compressive stress equal to the yield stress on one side of the neutral axis and tension yielding on the other side. The plastic moment can be attained if there is no local buckling of any cross-sectional element.

**Resistance factor:** A reduction factor applied to the nominal, or theoretical, resistance (strength). Although it can be equal to unity, it is usually less than unity.

**Shear connectors:** Devices that are welded to the top flanges of beams and embedded in the concrete slab supported by the beams. The most common type of shear connector is the headed stud.

**Shear lag:** A reduction in the strength of a tension member caused by not connecting some of the cross-sectional elements. It is accounted for by reducing the actual net area of the member to an effective net area.

**Slenderness ratio:** The ratio of the effective length of a member to the radius of gyration about one of the principal axes of the cross section.

**Tension field:** A condition existing in the buckled web of a plate girder in which the web cannot resist compression but is capable of resisting the diagonal tension within a panel defined by transverse web stiffeners.

**Torsional buckling:** A limit state for axially loaded compression members in which the member twists about its longitudinal axis.

## References

- American Institute of Steel Construction. 1994. *Manual of Steel Construction: Load and Resistance Factor Design*, 2nd ed. American Institute of Steel Construction, Chicago.
- American Institute of Steel Construction. 1993. *Load and Resistance Factor Design Specification for Steel Buildings*. American Institute of Steel Construction, Chicago.
- American Society of Civil Engineers. 1994. *Minimum Design Loads for Buildings and Other Structures*, ASCE 7-93 (formerly ANSI A58.1). American Society of Civil Engineers, New

York.

American Society for Testing and Materials. 1994. *1994 Annual Book of ASTM Standards*.

American Society for Testing and Materials, Philadelphia.

## Further Information

The American Institute of Steel Construction is a source of much useful and up-to-date information on structural steel design. A monthly magazine, *Modern Steel Construction*, provides information on structural steel construction projects, technical issues, and AISC activities. The *Engineering Journal* is a quarterly refereed journal containing technical papers with a practical orientation. Approximately every two years, AISC conducts a national lecture series on structural steel design topics of current interest, and the annual AISC National Steel Construction Conference provides a forum for designers, fabricators, and producers.

A valuable source of background material on the AISC specification is the commentary that accompanies it. This document, which along with the specification is contained in the *Manual of Steel Construction*, explains and elaborates on the specification provisions. The *Manual* also contains many illustrative examples and discussions.

Several useful textbooks on structural steel design are available. Comprehensive works covering both allowable stress design and LRFD include *Design of Steel Structures*, by E. H. Gaylord, C. N. Gaylord, and J. E. Stallmeyer, and *Steel Structures, Design and Behavior*, by C. G. Salmon and J. E. Johnson. *Structural Steel Design: LRFD Method*, by J. C. McCormac, and *LRFD Steel Design*, by W. T. Segui, treat load and resistance factor design exclusively.

Edward G. Nawy. "Concrete"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 30

## Concrete

---

### 30.1 Structural Concrete

Admixtures in Concrete • Properties of Hardened Concrete

### 30.2 Flexural Design of Reinforced Concrete Members

General Principles • Singly Reinforced Beam Design • Doubly Reinforced Sections • Columns • Walls and Footings

### 30.3 Shear and Torsion Design of Reinforced Concrete Members

Shear • Torsion

### 30.4 Prestressed Concrete

### 30.5 Serviceability Checks

### 30.6 Computer Applications for Concrete Structures

**Edward G. Nawy**

*Rutgers University*

---

## 30.1 Structural Concrete

---

Structural concrete is a product composed of properly designed mixture comprising portland cement, coarse aggregate (stone), fine aggregate (sand), water, air, and chemical admixtures. The cement acts as the binding matrix to the aggregate and achieves its strength as a result of a process of hydration. This chemical process results in recrystallization in the form of interlocking crystals producing the cement gel, which has high compressive strength when it hardens. The aggregate could be either natural, producing normal concrete, weighing  $150 \text{ lb/ft}^3$  ( $2400 \text{ kg/m}^3$ ), or artificial aggregate, such as pumice, producing lightweight concrete weighing  $\sim 110 \text{ lb/ft}^3$  ( $1750 \text{ kg/m}^3$ ).

Structural concrete should have a cylinder compressive strength of 3000 psi (20 MPa) at least, but often exceeding 4000–5000 psi (34.5 MPa). As the strength exceeds 6000 psi (42 MPa), such concrete is presently considered high strength concrete. Concrete mix designed to produce 6000 to 12 000 psi is easily obtainable today with the use of silica fume or plasticizers to lower the water/cement ratio and hence achieve higher strength due to the lower water content in the mix. A low water/cement or water/cementitious ratio of 30 to 25% can be achieved with these admixtures, with good workability and high slump fluidity for placing the concrete in the framework.

Concretes of compressive strengths reaching 20 000 psi (140 MPa) have been used in some concrete.

# Admixtures in Concrete

Admixtures in concrete can be summarized as follows:

1. *Accelerating admixtures*. They hasten the chemical hydration process.
2. *Air-entraining admixtures*. They form minute bubbles 1 mm in diameter and smaller evenly distributed in the mix to protect the concrete from freeze and thaw cycles.
3. *Water-reducing and set-controlling admixtures*. They increase the strength of the concrete through reducing the water content but maintaining the slump (fluidity) of the concrete.
4. *Polymers*. They replace a major portion of the needed water content and can produce concretes of strength in excess of 15 000 psi.
5. *Superplasticizers*. These are high-range water-reducing chemical admixtures. A dosage of 1 to 2% by weight of cement is recommended.
6. *Silica fume admixtures*. They are new pozzolanic materials as a by-product of high-quality quartz with coal in the electric arc furnace that produces silicon and ferrosilicon alloys. They are used to attain very high-strength concrete in three to seven days with relatively less increase in strength than normal concrete after 28 days. A dosage of 5 to 30% by weight of the cement can be used, depending on the strength needed. A compressive strength of 15 000 psi (105 MPa) or more can be readily achieved with good control.

## Properties of Hardened Concrete

The mechanical properties of hardened concrete can be classified as (1) short-term or instantaneous properties, and (2) long-term properties.

The short-term properties can be enumerated as (1) strength in compression, tension, and shear, and (2) stiffness measured by the modulus of elasticity. The long-term properties can be classified in terms of creep and shrinkage.

1. *Compressive strength,  $f'_c$* . It is based on crushing 6 in. diameter by 12 in. height standard concrete cylinders at a specified loading rate in a compression testing machine.
2. *Tensile strength,  $f'_t$* . Tensile strength of concrete is relatively low. A good approximation of tensile strength is  $f'_t$  ranging between 10 and 15% of  $f'_c$ .
3. *Shear strength,  $v_c$* . It is more difficult to determine experimentally. It can vary from about  $2\sqrt{f'_c}$  for normal weight reinforced concrete beams to about 80%  $f'_c$  in direct shear combined with compression.
4. *Modulus of elasticity  $E_c$  for stiffness or ductility determination*. The **ACI 318-95 Code** [ACI Committee 318, 1996] specifies using a secant modulus, given in psi or MPa:

$$E_c = 33W_c^{1.5} \sqrt{f'_c} \text{ psi} \quad (30.1)$$

$$E_c = 0.043W_c^{1.5} \sqrt{f'_c} \text{ MPa} \quad (30.2)$$

5. *Shrinkage*. There are two types of shrinkage: plastic shrinkage and drying shrinkage. *Plastic shrinkage* occurs during the first few hours after placing fresh concrete in the forms, resulting in a random map of cracks. *Drying shrinkage* occurs after the concrete has already attained its final set and a good portion of the chemical hydration process in the cement gel has been accomplished. Drying shrinkage results in well-defined linear cracks of larger width than plastic shrinkage cracks.



6. *Creep*. It is the lateral flow of the material under external load. The member sustains an increase in lateral strains with time due to the sustained load, hence increased stresses in the member and sometimes an almost 100% increase in deflection with time.

Details of all these effects and the ACI 318 Code provisions to control them in the design of reinforced and prestressed concrete structures are given in Nawy [1996a, b].



---

#### THE TUNNEL LINING

The Eurotunnel system connects Britain and France via a massive underground transportation system. The Eurotunnel system is not one but actually three transportation tunnels which extend 24 miles (38 km) under the English Channel.

To create a safe mass-transportation tunnel, the walls are lined with rings made of concrete segments. These segments were brought into the tunnel by a supply train before being installed. Each concrete ring was then locked into place with a smaller, wedge-shaped "key" segment. The tiny 3/4 in. (20 mm) gap between the ring and the rock face was then sealed with cement grout. It took close to 20 minutes to fix each segment in place.

In both Britain and France, special factories were set up to make the tunnel linings. The French factory was in Sangatte; the British factory was at the Isle of Grain in Kent, some 60 miles (100 km) away. There was not enough room on site for these factories. The British tunnel linings each consisted of eight segments plus a "key" segment. The French tunnel linings consisted of six segments plus the key. Cast-iron linings were used instead of concrete in areas of weaker rock and at cross-passage junctions. (Courtesy of ©Eurotunnel 1994. Photo by QA Photos, Hythe. Used with permission.)

## 30.2 Flexural Design of Reinforced Concrete Members

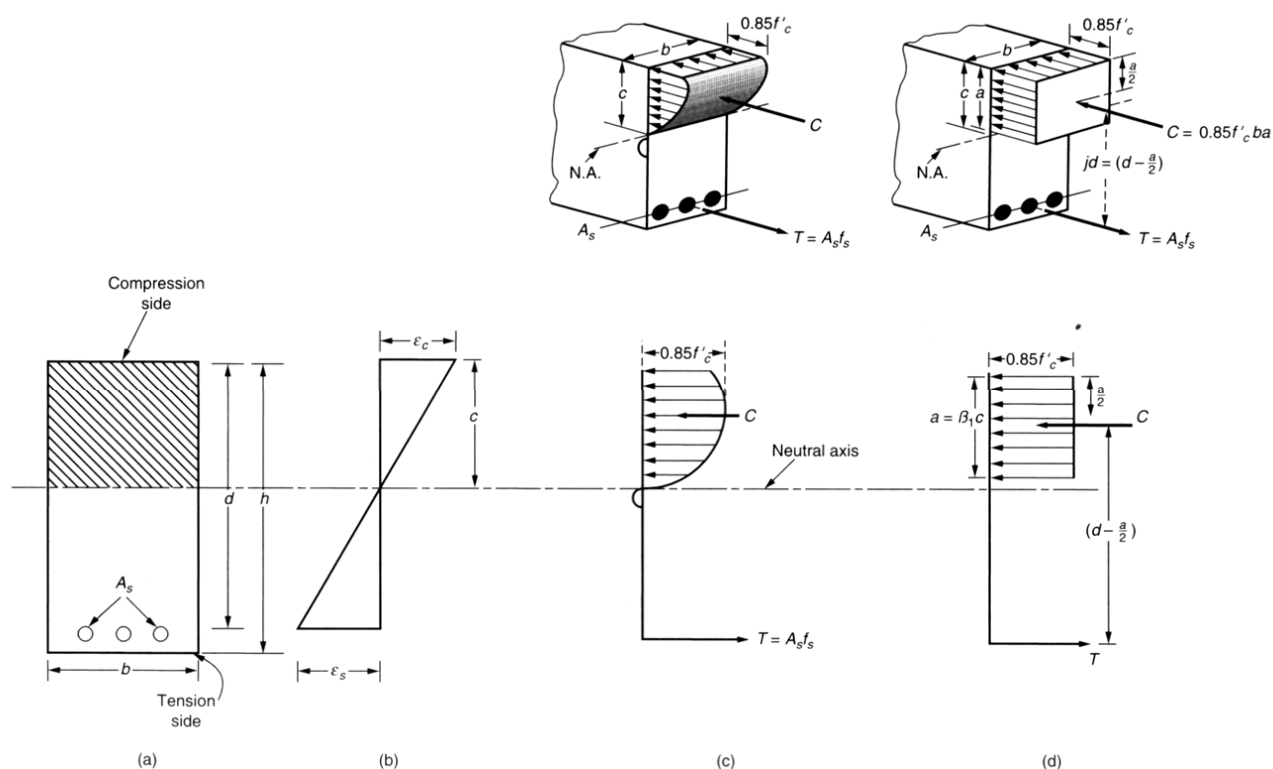
### General Principles

Concrete structural systems are generally composed of floor slabs, **beams**, **columns**, walls, and foundations. Present codes of practice (ACI, PCI, UBC, CEB) all require ultimate strength procedures in proportioning the structural elements, what is termed *strength design* by the code of the American Concrete Institute (ACI 318-95).

The strength of a particular structural unit is termed *nominal strength*. For example, in the case of a beam, the resisting moment capacity of the section calculated using the equations of equilibrium and the properties of the concrete and the steel reinforcement is called *nominal moment strength*  $M_n$ . This nominal strength is reduced using a strength reduction factor,  $\phi$ , to account for inaccuracies in construction such as in the geometrical dimensions or position of reinforcing bars or variation in concrete properties.

The design principles are based on equilibrium of forces and moments in any section. Since concrete is weak in tension, the design assumes that the concrete in the tensile part of a beam cross section does *not* carry any load or stress, hence it is disregarded. By doing so, the tensile equilibrium force is wholly taken by the tension bars  $A_s$  in Fig. 30.1(a). The actual distribution of compressive stress is parabolic as seen in Fig. 30.1(c). However, the ACI Code adopted the use of an equivalent rectangular block, as in Fig. 30.1(d), in order to simplify the computations.

**Figure 30.1** Stress and strain distribution across reinforced concrete beam depth: (a) beam cross section, (b) strain distribution, (c) actual stress block, (d) assumed equivalent block. (Source: Nawy, E. G. 1996. *Reinforced Concrete—A Fundamental Approach*, 3rd ed. Prentice Hall, Englewood Cliffs, NJ. With permission.)



## Singly Reinforced Beam Design

Such a beam would have reinforcement only on the tension side. If one considers the equilibrium forces in Fig. 30.1(d):

C = Volume of the equivalent rectangular block =  $0.85 f'_c b a$

T = Tensile force in the reinforcement =  $A_s f_y$  where  $f_y$  = yield strength of the reinforcement

then  $C = T$  or  $0.85 f'_c b a = A_s f_y$ . Therefore, the depth of the equivalent rectangular block is

$$a = \frac{A_s f_y}{0.85 f'_c b} \quad (30.3)$$

Since the center of gravity of the compressive force  $C$  is at a distance  $a/2$  from the top compression fibers, the arm of the moment couple is  $[d - (a/2)]$ . Hence, the nominal moment strength of the section is

$$M_n = A_s f_y \left( d - \frac{a}{2} \right) \quad (30.4)$$

Note that if the total thickness of the section is  $h$ , Eq. (30.4) considers  $d$  the effective depth to the *centroid* of the reinforcement, thereby disregarding the concrete cover and assuming it to be only for fire and corrosion protection. The percentage of reinforcement is  $\rho = A_s / bd$ , and the reinforcement index is  $\omega = (A_s / bd) \times (f_y / f'_c)$ . Hence, Eq. (30.4) can also be written as

$$M_n = [\omega f'_c (1 - 0.59\omega)] b d^2 \quad (30.5)$$

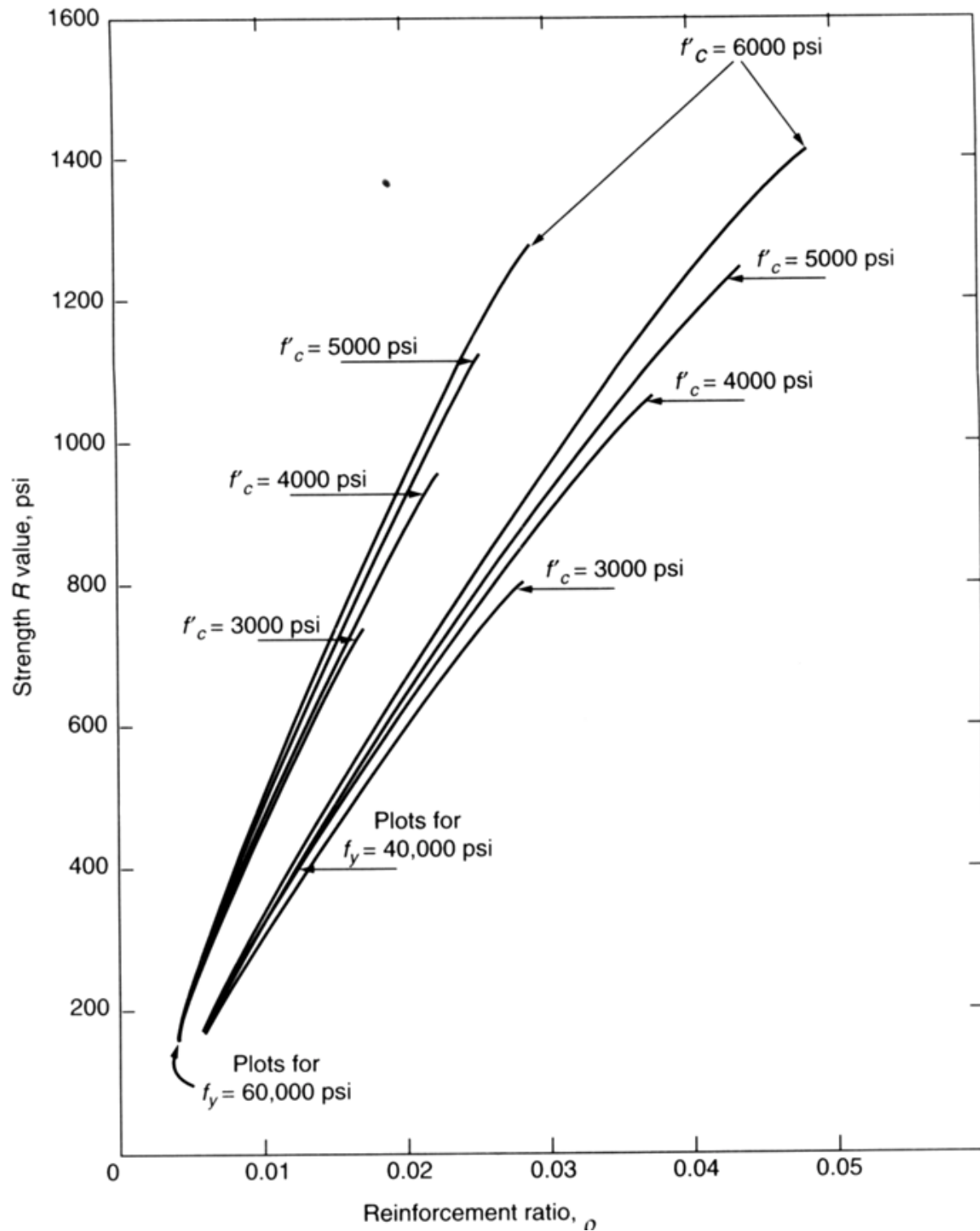
If one uses

$$R = \omega f'_c (1 - 0.59\omega) \quad (30.6)$$

then  $M_n = R b d^2$ . Figure 30.2, for singly reinforced beams, can give a rapid choice of the width and depth of a beam section, as one can usually use  $b = \frac{1}{2} d$ .

In order to ensure ductility of the structural member and corresponding warning of failure, the ACI Code limits the reinforcement percentage  $\rho$  to a maximum 75% of the balanced percentage  $\rho_b$ , namely that the steel would have to yield before the concrete crushes. Balanced failure

**Figure 30.2** Strength  $R$  curves for singly reinforced concrete beams. (Source: Nawy, E. G. 1996. *Reinforced Concrete—A Fundamental Approach*, 3rd ed. Prentice Hall, Englewood Cliffs, NJ. With permission.)



denoted by  $\rho_b$  occurs by simultaneous crushing of the concrete on the compression side and yielding of the steel at the tension side. To prevent congestion of bars, it is advisable not to use  $\rho$  more than  $0.50 \rho_b$ .

The code also requires a minimum reinforcement area so that the beam can behave as a reinforced concrete section. The minimum  $A_s = (3 - \sqrt{f'_c/f_y})b_w d$ , where  $b_w$  = width of the beam web in case of a flanged section.

**Design Example.** A simply supported singly reinforced concrete beam is subjected to a total factored moment including its self weight,  $M_u = 4.1 \cdot 10^6$  in.-lb. Given values are  $f'_c = 4000$  psi (34.5 MPa),  $f_y = 60\,000$  psi (414 MPa), and  $\rho_b = 0.0285$ .

**Solution.** Required nominal moment strength  $M_n$  equals  $M_u/\phi$ , where  $\phi = 0.90$  for flexure:  $M_n = 4.1 \cdot 10^6/0.90 = 4.5 \cdot 10^6$  in.-lb. Assume  $\rho = 0.50\rho_b = 0.0285 \times 0.5 = 0.0143$ .

$$\omega = \frac{\rho f_y}{f'_c} = \frac{0.0143 \times 60\,000}{4000} = 0.215$$

Equation (30.5) yields

$$R = \omega f'_c (1 - 0.59\omega) = 0.215 \times 4000 (1 - 0.59 \times 0.215) \simeq 750.$$

Alternatively, entering the chart in [Fig. 30.2](#) also gives  $R \simeq 750$ . Assuming  $b = \frac{1}{2}d$ , Eq. (30.6) yields

$$d = \sqrt[3]{\frac{M_n}{0.5R}} = \sqrt[3]{\frac{4.5 \times 10^6}{0.5 \times 750}} = 22.8 \text{ in.}$$

Based on practical considerations, use  $b = 12$  in. (305 mm),  $d = 23$  in. (585 mm), and total depth  $h = 26$  in. (660 mm). Then,  $A_s = 0.0143bd = 0.0143 \times 12 \times 23 = 3.95 \text{ in}^2$ . Using three #10 bars (32.2 mm diameter),  $A_s = 3.81 \text{ in}^2$  (2460 mm<sup>2</sup>).

To check nominal moment strength, use Eq. (30.3) to find

$$a = \frac{A_s f_y}{0.85 f'_c b} = \frac{3.81 \times 60\,000}{0.85 \times 4000 \times 12} = 5.60 \text{ in.}$$

Available  $M_n = 3.81 \times 60\,000 [23.0 - (5.6/2)] = 4.6 \cdot 10^6$  in.-lb (5.2 · 10<sup>5</sup> kN-M) > Required  $M_n = 4.5 \cdot 10^6$  in.-lb; adopt design.

A check for minimum reinforcement and shear capacity also has to be performed as must deflection, as detailed in [\[Nawy, 1996a\]](#).

$$\text{Overdesign} = \frac{4.6 - 4.5}{4.6} = 2.2\%$$

Overdesign should not be in excess of 4 to 5%.

## Doubly Reinforced Sections

These are beam sections where compression reinforcement  $A'_s$  is used about 2 in. from the compression fibers. The reinforcement  $A'_s$  contributes to reducing the required depth of section where there are clearance limitations. An extra nominal moment  $M' = A'_s f_y (d - d')$  is added to the section,  $d'$  being the depth from the extreme compression fibers to the centroid of the compression reinforcement  $A'_s$ . Consequently, Eq. (30.4) becomes

$$M_n = (A_s - A'_s) f_y \left( d - \frac{a}{2} \right) + A'_s f_y (d - d') \quad (30.7)$$

A similar expression to Eq. (30.7) can be derived for flanged sections, as in Nawy [1996a].

## Columns

Columns are compression members that can fail either by material failure if they are nonslender or by buckling if they are slender. Columns designed using the material failure criteria should have a slenderness ratio  $kl/r$  not to exceed 22 for nonbraced columns. The value  $k$  is the stiffness factor at column ends,  $l$  is the effective length, and  $r$  is the radius of gyration  $= 0.3h$  for rectangular sections.

A column is essentially a doubly reinforced flexural section that is also subjected to an axial force  $P_n$  in addition to the forces  $C$  and  $T$  and the  $A'_s$  force  $C'$  shown in Fig. 30.1. Therefore, from equilibrium of forces,

$$P_n = 0.85 f'_c b a + A'_s f'_s + A_s f_s \quad (30.8)$$

$$M_n = P_n e = 0.85 f'_c b a \left( \bar{y} - \frac{a}{2} \right) + A'_s f'_s (\bar{y} - d') + A_s f_s (d - \bar{y}) \quad (30.9)$$

$$e = \frac{M_n}{P_n} = \text{eccentricity}$$

where  $\bar{y}$  = distance to center of gravity of section  $= h/2$  for rectangular section. Notice the similarities between Eqs. (30.9) and (30.7).

Although initial failure in beams is always by yielding of the reinforcement through limiting  $\rho$ , this is not possible in columns, as the mode of failure depends on the magnitude of eccentricity,  $e$ . If  $e_b$  is the balanced condition eccentricity, then  $e < e_b$  = compression failure by concrete crushing, and  $e > e_b$  = tensile failure by yielding of the reinforcement at the tension side.

This subject is very extensive, particularly if buckling is also to be considered. The reader is advised to consult textbooks such as [Nawy, 1996a] and handbooks.

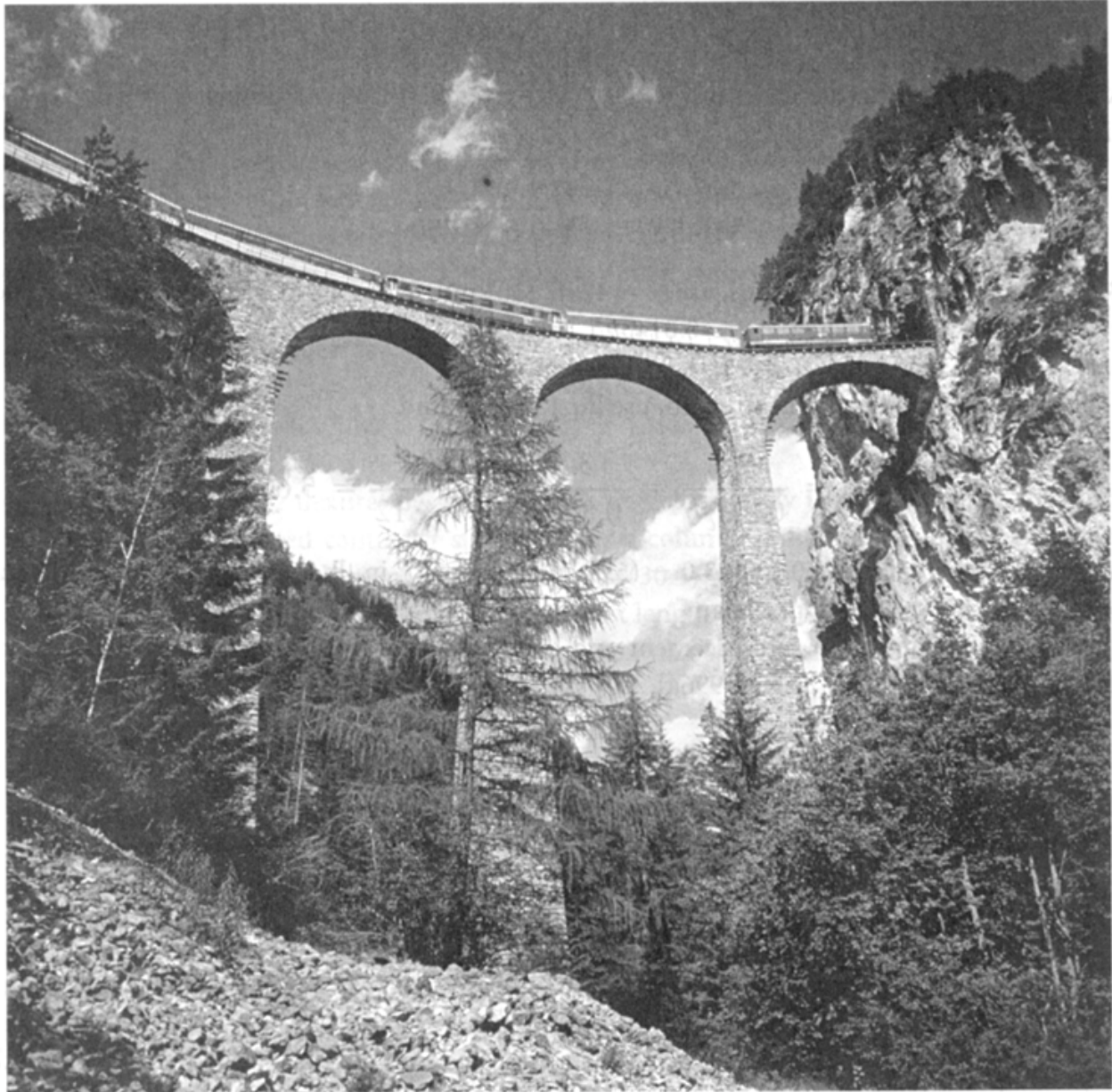
## Walls and Footings

The same principles for the flexural design of beams and slabs apply to walls and footings. An



isolated footing is an inverted cantilever supported by a column and subjected to uniform soil pressure. It is treated as a singly reinforced beam [Eq. (30.4)] subjected to a factored moment  $M_u = w_u l^2 / 2$ , where  $w_u$  is the intensity of load per unit length and  $l$  is the arm of the cantilever. In the same manner, one can design retaining walls and similar structural systems.

---



THE LANDWASSER BRIDGE

B. van Gelder, Purdue University

In the canton of Graubünden, near the city of Filisur, the Landwasser bridge, with a height of

183 feet and a span of 390 feet, provides a means for the Rhaetian railroad to cross the Albula river. This part of the railroad connects the cities of Chur and St. Moritz.

Chur is the oldest city of Switzerland, archaeologically dated back to 3000 B.C. The Romans built an important traffic center in Chur around 2000 years ago. St. Moritz is a ski resort for the international jet set and received fame after the Winter Olympics were staged there twice this century (1928 and 1948).

The viaduct shown is one of 19 bridges and no fewer than 23 tunnels and galleries on a 20 km railroad segment between Filisur and the Albula tunnel entrance at Preda. The Albula river alone must be crossed four times before the 5865 m tunnel is entered.

Close by is the Langwieser viaduct, which is on the Rhaetian line between Chur and Arosa. It was the longest steel-reinforced concrete bridge of its time, with a length of 932 feet. (Photo courtesy of the Swiss National Tourist Office.)

## 30.3 Shear and Torsion Design of Reinforced Concrete Members

---

### Shear

External transverse load is resisted by internal **shear** in order to maintain section equilibrium. As concrete is weak in tension, the principal tensile stress in a beam cannot exceed the tensile strength of the concrete. The principal tensile stress is composed of two components—shear stress,  $v$ , and tensile stress,  $f_t$ —causing diagonal tension cracks at a distance  $d$  from the face of the support in beams and at a distance  $d/2$  from the face of the support in two-way slabs.

Consequently, it is important that the beam web be reinforced with diagonal tension steel, called *stirrups*, in order to prevent diagonal shear cracks from opening. The resistance of the plain concrete in the web sustains part of the shear stress and the balance has to be borne by the diagonal tension reinforcement. The shear resistance of the plain concrete in the web is termed *nominal shear strength*,  $V_c$ . A conservative general expression for  $V_c$  from the ACI 318 code is

$$V_c (\text{lb}) = 2.0\lambda\sqrt{f'_c}b_wd \quad (30.10)$$

$$V_c (\text{newton}) = \lambda(\sqrt{f'_c}/6)b_wd \quad (30.11)$$

where  $f'_c$  is in MPa in Eq. (30.11) and  $\lambda$  equals

- 1 for stone aggregate concrete
- 0.85 for sand-lightweight concrete
- 0.75 for all lightweight concrete

If  $V_n = V_u/\phi$  = the required nominal shear, the stirrups should be designed to take the difference



between  $V_u$  and  $V_c$ —namely  $V_s = [(V_u/\phi) - V_c]$  —where  $V_u$  is the factored external shear and  $\phi = 0.85$  in shear and torsion. The spacing of the transverse web stirrups is hence

$$s = \frac{A_v f_y d}{(V_u/\phi - V_c)} \quad (30.12)$$

where  $A_v$  = cross-sectional area of the web steel (2 stirrup legs). Maximum spacing of stirrups is  $d/2$  or 12 in., whichever is smaller. A concrete section designed for flexure as described earlier has to be enlarged if  $V_s = (V_n - V_c) > 8\sqrt{f'_c} b_w d$ .

## Torsion

If a beam is also subjected to **torsion** combined with shear, diagonal cracks described in the previous section have to be prevented from opening. This is accomplished by use of *both* vertical closed stirrups and additional longitudinal bars evenly divided among the four faces of the beam. The longitudinal reinforcement is required since torsion causes a three-dimensional warped surface.

The ACI 318-95 code disallows utilization of the nominal torsional strength  $T_c$  of the plain concrete in the web and requires that all the torsional moment  $T_n$  be borne by the transverse closed stirrups and the longitudinal bars. It assumes that the volume of the transverse stirrups is equal to the volume of the longitudinal bars. The same equations for torsion in reinforced concrete elements are used with adjusting modifiers when applied to prestressed concrete.

## 30.4 Prestressed Concrete

Reinforced concrete is weak in tension but strong in compression. To maximize utilization of its material properties, an internal compressive force is induced in the structural element through the use of highly stressed prestressing tendons to precompress the member prior to application of the external gravity live load and superimposed dead load. Typical effect of the prestressing action is shown in Fig. 30.3, using a straight tendon as is usually the case in precast elements. For in situ-cast elements, the tendon can be either harped or usually draped in a parabolic form. As can be seen from Fig. 30.3, the prestressing force  $P$  alone induced a compressive stress at the bottom fibers  $f_b$  and a tensile stress at the top fibers  $f_t$  such that

$$f_b = -\frac{P_i}{A} - \frac{P_i e c}{I} \quad (30.13a)$$

$$f_t = -\frac{P_i}{A} + \frac{P_i e c}{I} \quad (30.13b)$$

where  $P_i$  is the initial prestressing force prior to losses.

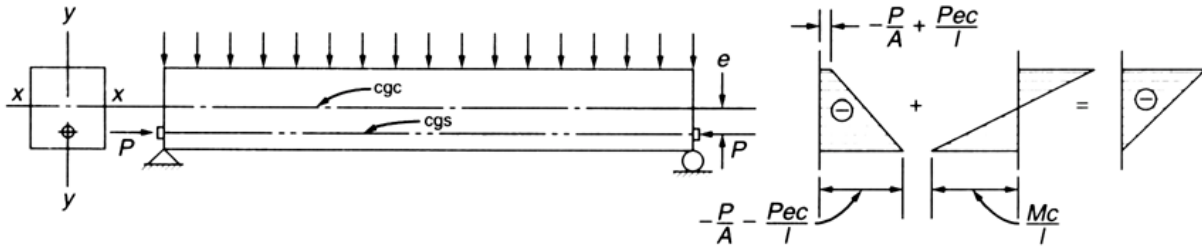
With addition of stress due to self-weight of the concrete beam and external live and superimposed dead load—moment  $M_T$ —the stresses become:

$$f_b = -\frac{P_e}{A} - \frac{P_e c}{I} + \frac{M_T c}{I} \quad (30.14a)$$

$$f_t = -\frac{P_e}{A} + \frac{P_e c}{I} - \frac{M_T c}{I} \quad (30.14b)$$

where  $P_e$  is the effective prestressing force after losses in prestress.

**Figure 30.3** Stress distribution at service load in prestressed concrete beams with constant tendons.  
(Source: Nawy, E. G. 1996. *Prestressed Concrete—A Fundamental Approach*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ. With permission.)



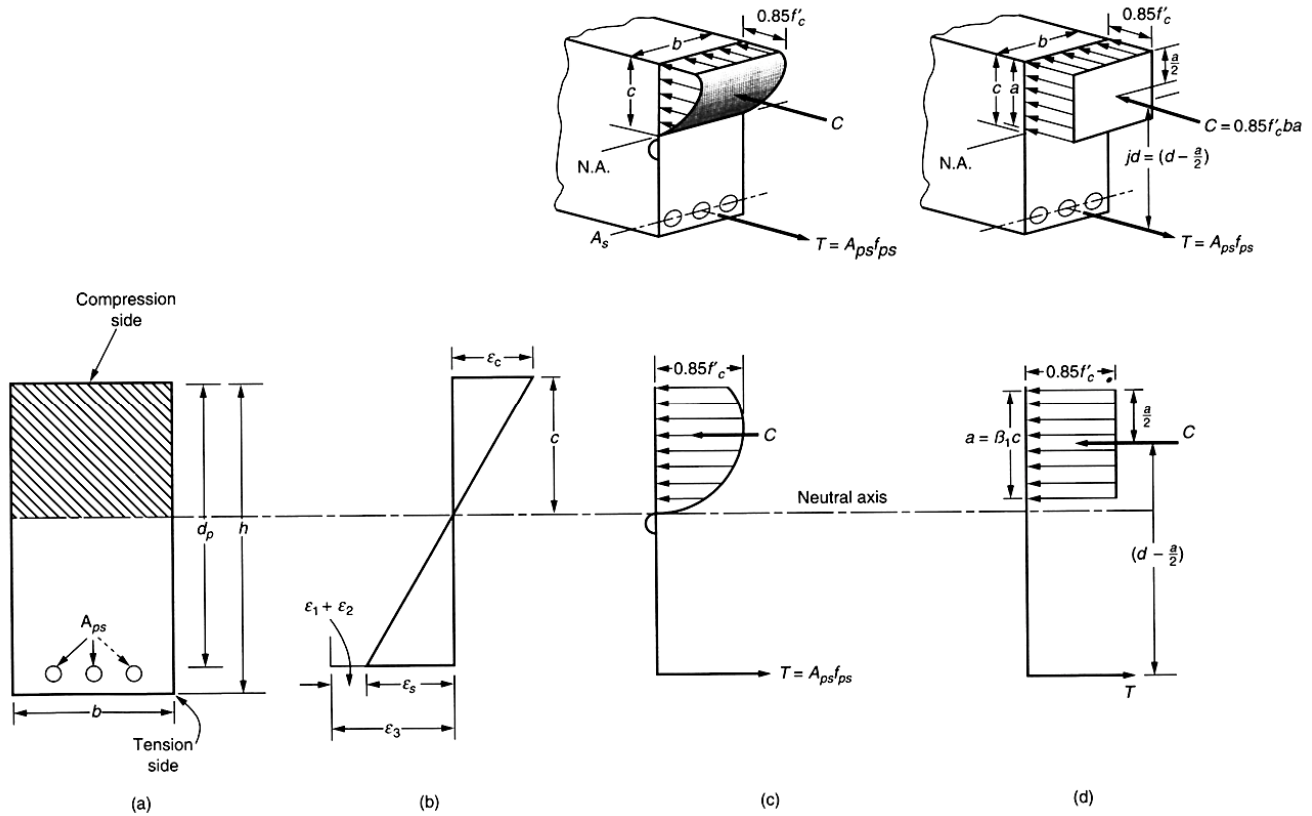
The stress diagram at the extreme right in Fig. 30.3 indicates zero tension at the bottom fibers and maximum compression at the top fibers. In this manner, prestressing the beam has resulted in full utilization of the properties of the concrete, eliminating tension cracking at the bottom fibers.

Whereas reinforced concrete members are designed only for ultimate load, prestressed concrete members are first designed for service load moments as in Eq. (30.14) and then analyzed for ultimate load capacity, namely the nominal moment strength  $M_n$ . This is necessary for determining the reserve strength available in the member between the service load level and collapse, as prestressed beams can be underreinforced (tension steel yielding) or overreinforced (compression side concrete crushing). Figure 30.4 gives the sets of equilibrium forces acting on the concrete section. Notice their similarity to those of reinforced concrete sections in Fig. 30.1. If both prestressing tendons and mild steel are used in the prestressed beam, the nominal moment strength is

$$M_n = A_{ps} f_{ps} \left( d_p - \frac{a}{2} \right) + A_s f_y \left( d - \frac{a}{2} \right) + A'_s f_y (d - d') \quad (30.15)$$

It should be noted that the prestressing force  $P$  can either be the initial prestressing force,  $P_i$ , in Eq. (30.14) or the *effective* service load,  $P_e$ , after losses due to concrete shrinkage, concrete creep, relaxation of the prestressing steel, frictional losses in posttensioned beams, and anchorage loss. A loss in the initial prestress of 20 to 25% is not unreasonable, so that  $P_e$  is quite often 0.80 to 0.75  $P_i$ .

**Figure 30.4** Stress and strain distribution across prestressed concrete beam depth: (a) beam cross section, (b) strain distribution, (c) actual stress block, (d) assumed equivalent block. (Source: Nawy, E. G. 1996. *Prestressed Concrete—A Fundamental Approach*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ. With permission.)



The ACI code prescribes the following allowable stresses at service load. For concrete stresses

$$\begin{aligned}
 f'_{ci} &\simeq 0.75 f'_c \text{ psi} \\
 f_{ci} &\simeq 0.60 f'_c \text{ psi} \\
 f_{ti} &= 3\sqrt{f'_{ci}} \text{ psi} && \text{on span } (\sqrt{f'_c}/4 \text{ MPa}) \\
 &= 6\sqrt{f'_{ci}} \text{ psi} && \text{at support } (\sqrt{f'_{ci}}/2 \text{ MPa}) \\
 f_c &= 0.45 f'_c \text{ to } 0.60 f'_c \\
 f_t &= 6\sqrt{f'_c} \text{ psi} && (\sqrt{f'_c}/2 \text{ MPa}) \\
 &= 12\sqrt{f'_c} \text{ psi} && \text{if deflection verified } (\sqrt{f'_c} \text{ MPa})
 \end{aligned}$$

and for reinforcing tendon stresses

Tendon jacking:  $f_{ps} = 0.94f_{py} \leq 0.80f_{pu}$

Immediately after stress transfer:  $f_{ps} = 0.82f_{py} \leq 0.74f_{pu}$

Posttensioned members  
at anchorage immediately  
after anchorage:  $f_{ps} = 0.70f_{pu}$

where

$f_{ps}$  = Ultimate design stress allowed in tendon

$f_{py}$  = Yield strength of tendon

$f_{pu}$  = Ultimate strength of tendon

$f_{ti}$  = Initial tensile stress in concrete

$f_c$  = Service load concrete compressive strength

$f_{ci}$  = Initial compressive stress in concrete

$f_t$  = Service load concrete tensile strength

View of the oldest concrete street in the U.S. paved in 1893 in Bellefontaine, Ohio. The street was designated a National Historic Civil Engineering Landmark in 1976 by the American Society of Civil Engineers. (Photo courtesy of ASCE.)



#### FIRST CONCRETE PAVEMENT IN THE U.S.

The first concrete pavement in the U.S. was laid in 1893 in Bellefontaine, Ohio. This pavement, located on Court Avenue in Bellefontaine, represents the first engineering use of portland cement in the nation and was the forerunner of the many thousands of miles of such roads in the United States.

Through experiments with limestone and clay marl, George W. Bartholomew, founder of Buckeye Portland Cement Company, turned out the first known cement samples in the U.S. He poured the first test section of concrete pavement in Bellefontaine in 1891. Two years later he built the first concrete-paved street in the U.S. A section of the Bellefontaine pavement was taken to the Chicago's World Fair in 1893, where it won first prize for engineering technology advancement. East Court Avenue has stood the test of time and traffic and has been virtually maintenance-free for more than 100 years.

The use of concrete to withstand steel-rimmed wagon wheels and pounding horses' hooves marked the transition in highway construction as automobiles appeared on the scene. Perhaps no single factor has contributed more to the rapid commercial and industrial progress of the U.S. than paved roads. During the early 1900s delegations from various cities throughout the country visited Bellefontaine to see the concrete streets. Henceforth, thousands upon thousands of miles of concrete highways link the U.S. (Courtesy of ASCE.)

**The Invention of Portland Cement** On December 15, 1824, Joseph Aspdin, a bricklayer in England, obtained a patent for the manufacture of a new and improved cement. He called his product portland cement because it resembled a natural limestone quarried in England on the Isle of Portland. Aspdin promoted the idea to his son William who carried out its manufacture.

Portland cement began to be manufactured in earnest in Europe about 1850. Its first extensive use was in constructing London's sewer system in 1859–1867, a project that greatly boosted portland cement's popularity.

Manufacturing of portland cement in the U.S. began in the 1870s. The first U.S. plant manufacturing portland cement was located in Coplay, Pennsylvania in 1872. The first manufacturing plant in Canada was in Hull, Quebec in 1889. (Courtesy of ASCE.)

## 30.5 Serviceability Checks

**Serviceability** of structural components is a major factor in designing structures to sustain acceptable long-term behavior. It is controlled by limiting deflection and cracking.

For deflection computation and control, an effective moment of inertia is used. Details of design for deflection in reinforced concrete beams and slabs and for deflection and camber in prestressed concrete with design examples are given in ACI Committee 435 [1995]. Table 30.1 (from ACI) gives the allowable deflections in terms of span for reinforced concrete beams.

**Table 30.1** Minimum Thickness,  $h$ , of Nonprestressed Beams or One-way Slabs

Member*	Simply Supported	One End Continuous	Both Ends Continuous	Cantilever
Solid one-way slabs	$l/20$	$l/24$	$l/28$	$l/10$
Beams or ribbed one-way slabs	$l/16$	$l/18.5$	$l/21$	$l/8$

Note: Span length  $l$  is in inches.

\*Members not supporting or attached to partitions or other construction are likely to be damaged by large deflections.

For crack control in beams and two-way slab floor systems, it is more effective to use smaller diameter bars at smaller spacing for the same area of reinforcement. ACI Committee 224 [1990] gives a detailed treatment of the subject of crack control in concrete structures. Table 30.2 gives the tolerable crack widths in concrete elements.

**Table 30.2** Tolerable Crack Widths

Exposure Condition	Tolerable Crack Width	
	in.	mm <sup>2</sup>
Dry air or protective membrane	0.016	0.40
Humidity, moist air, soil	0.012	0.30
Deicing chemicals	0.007	0.18
Seawater and seawater spray; wetting and drying	0.006	0.15
Water-retaining structures (excluding nonpressure pipes)	0.004	0.10

## 30.6 Computer Applications for Concrete Structures

In the design of concrete structures, several canned computer programs are available both for reinforced concrete and prestressed concrete systems. They can be either analysis or design programs. These programs are available for personal computers using MSDOS or MS Windows operating systems. Typical general purpose programs are STRUDEL, ANSYS, SAP 90, and ETABS. The SAP 90 program can handle in excess of 5000 nodes and requires a larger memory than the others. Except for lack of space, a long list of available programs could be compiled here. The structural engineer can without difficulty get access to all present and forthcoming programs in the market.

The specialized concrete programs are numerous. The following programs are widely used. PCA programs include

1. *ADOSS*. For two-way reinforced concrete slabs and plates.
2. *PCA Columns*. For nonslender and slender regular and irregular columns.
3. *PCA Frame*. For analysis of reinforced concrete frames, including second-order analysis for slender columns taking into account the  $P - \Delta$  effect.



4. *PCA MATS*. For the design of flexible mat foundations.

Other programs include

1. *ADAPT*. A comprehensive program for the design of reinforced and prestressed concrete two-way action slabs and plates with the capability of drafting many AUTOCAD version 12.0.
2. Miscellaneous design or analysis programs for proportioning sections:
  - NRCPCLI*. For reinforced concrete sections [Nawy 1996a]
  - NRCPCLII*. For prestressed concrete sections [Nawy 1996b]
  - RCPCDH*. For reinforced concrete beams, columns, and isolated footings.
  - RISA2D* and *RISA3D*. General purpose programs for both steel and concrete and frame analysis.

Among drafting programs, AUTOCAD is a comprehensive general purpose program very widely used for drafting working drawings including reinforcing details. Presently, AUTOCAD version 13.0 is available. To make efficient use of the program, it is advisable to have a personal computer with 486–66 MHz speed capacity or higher, 16 MB of internal memory, and a 200 MB hard disk capacity or higher in order to accommodate also the structural program of choice for the design of the concrete system.

In summary, it should be emphasized that the computer programs discussed in this section are only representative. Other good software is available and being developed with time. Users should always strive to utilize the program that meets their particular needs, preferences, and engineering backgrounds.

## Defining Terms

**Admixtures:** Chemical additives to the concrete mix in order to change the mechanical and performance characteristics of the hardened concrete.

**ACI:** American Concrete Institute.

**Beams:** Supporting elements to floors in structural systems.

**Codes:** Standards governing the design and performance of constructed systems to ensure the safety and well-being of the users.

**Columns:** Vertical compression supports.

**Ductility:** Ability of the member to absorb energy and deform in response to external load.

**Flexure:** Bending of a structural element due to applied load.

**Footings:** Foundation elements within the soil supporting the superstructure.

**PCI:** Prestressed Concrete Institute.

**Prestressed concrete:** Concrete elements such as beams, columns, or piles subjected to internal (or external) compression prior to the application of external loads.

**Serviceability:** Cracking and deflection performance of a structural member or system.

**Shear:** Force due to external load acting perpendicular to the beam span to shear the section.

**Torsion:** Twisting moment on a section.

**U.B.C.:** Uniform Building Code.

## References

- ACI Committee 224. 1990. *Control of Cracking in Concrete Structures*. Committee Report, Publ. American Concrete Institute, Detroit.
- ACI Committee 435. 1995. *Control of Deflection in Concrete Structures*. Committee Report, E. G. Nawy, Chairman, Publ. American Concrete Institute, Detroit.
- ACI Committee 318. 1996. *Building Code Requirements for Reinforced Concrete: ACI 318-95 and Commentary ACI R-95*. Institute Standard, American Concrete Institute, Detroit.
- Hsu, T. C. 1993. *Unified Theory of Reinforced Concrete*. CRC Press, Boca Raton, FL.
- Nawy, E. G. 1996a. *Prestressed Concrete—A Fundamental Approach*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.
- Nawy, E. G. 1996b. *Reinforced Concrete—A Fundamental Approach*, 3rd ed. Prentice Hall, Englewood Cliffs, NJ.

## Further Information

A comprehensive treatment of all aspects of concrete proportioning, design, construction, and long-term performance can be found in the five volumes *Manual of Concrete Practice* published by the American Concrete Institute, Detroit.

The proceedings of the *Structural Journal*, *Materials Journal*, and *Concrete International Journal*, all published by the American Concrete Institute, are an additional source for the latest research and development in this area.

The proceedings of the *PCI Journal* published bimonthly by the Prestressed Concrete Institute, Chicago, deals with all aspects of fabrication, design, and construction of precast and prestressed concrete beams for residential building as well as bridges.

A text book by E. G. Nawy on *High Strength High Performance Concrete*—dealing with the fundamentals of developing high strength concrete and detailed discussion of cementitious based materials and ultra high strength concretes—was published in 1995 by Longman Higher Education Series.

The *Design Handbook*, in three volumes, SP-17, published annually by the American Concrete Institute, Detroit, MI, is an all-encompassing publication containing numerous charts, monograms, and tables, as well as examples pertaining to all design aspects of reinforced concrete structural members.



Breyer, D. E. "Timber"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 31

## Timber

---

- 31.1 Durability of Wood
- 31.2 Wood Products
- 31.3 Member Design
- 31.4 Connections
- 31.5 Lateral Force Design

### Donald E. Breyer

*California State Polytechnic University, Pomona*

The four primary building materials used in the construction of civil engineering structures are reinforced concrete, reinforced masonry, structural steel, and *timber*. This chapter will give a brief introduction to *engineered wood structures*.

Wood is often used as the framing material in low-rise buildings (one to four stories), but timber has been used in taller structures. Bridges are also constructed from wood and are generally limited to relatively short-span bridges on rural and forest service roads, but glued-laminated timber framing has been used in the construction of some highway bridges. Wood is also used in foundation systems such as timber piles; utility poles and towers are other examples of wood structures. In addition to these more permanent structures, wood is commonly used for such temporary structures as concrete formwork and falsework and for shoring of trenches during construction. Although many residential structures are engineered to some extent, wood is also used as a structural material in many nonengineered homes that fall into a category known as conventional construction.

As a biological product wood is a unique structural material. It is a renewable resource that can be obtained by growing and harvesting new trees. Proper forest management is necessary to provide a sustainable supply of wood products and to ensure that this is accomplished in an environmentally responsible way.

Wood can be used to create permanent structures. However, the proper use of wood as a structural material requires that the designer be familiar with more than stress calculations. Pound for pound, wood is stronger than many materials, but wood also has characteristics that, if used improperly, can lead to premature failure. Understanding the unique characteristics of wood is the key to its proper use.

### 31.1 Durability of Wood

---

In addition to being overstressed by some type of loading, a wood structure can be destroyed by

several environmental causes, including decay, insect attack, and fire.

The moisture content of wood is defined as the weight of water in the wood expressed as a percentage of the oven dry weight of the wood. In a living tree the moisture content can be as high as 200%. In a structure the moisture content of a wood member will be much less, and for a typical enclosed building the moisture content will range between 7 and 14%, depending on climate. However, if a structure houses a swimming pool or if there are high humidity conditions as in certain manufacturing plants, higher moisture contents will occur.

The best recommendation for preventing *decay* is to keep wood continuously dry. Special detailing may be required to accomplish this. However, the low moisture content of framing lumber in most enclosed buildings generally does not lead to decay. High moisture content or exposure to the weather (alternate wetting and drying) will cause decay in untreated wood products.

Wood that will be exposed to the weather or subject to other high-moisture conditions can be pressure impregnated with an approved chemical treatment to protect against decay. Pressure-treated lumber is obtained from a processing plant that specializes in treating wood products by forcing the appropriate chemicals under pressure into the wood cells. Paint-on chemicals are generally not effective.

Wood can also be destroyed by *insect attack*. Termites are the most common pest, but marine borers are found in ocean water. Termite protection may be obtained by providing a physical barrier, by maintaining a minimum clearance between the wood and soil, or by using pressure-treated wood. The same pressure-treated lumber is effective for both decay and termite protection. Marine borers require a different chemical treatment.

*Fire* is a threat to any structure, whether it is wood, steel, concrete, or masonry. However, wood is a combustible material, and building codes place restrictions on the height, area, and occupancy classification of a building that uses wood framing. Wood becomes harder to ignite as the cross-sectional dimensions of the framing members increase. Consequently, building codes recognize both *light-frame wood construction* (using relatively small size framing members) and *heavy-timber (HT) construction*. In a fire, a large wood member performs much better because a protective coating of char develops that helps to insulate the member.

## 31.2 Wood Products

---

The following wood products are used in structural design:

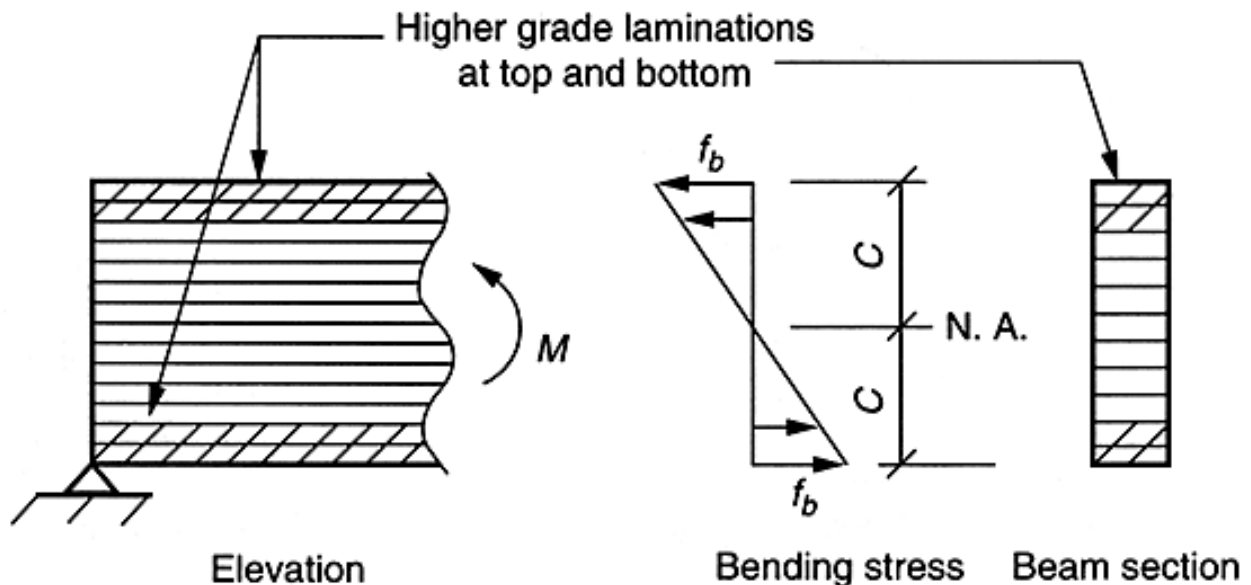
1. *Solid sawn lumber* (sawn lumber). Lumber of rectangular cross section cut from trees. A variety of species of commercial lumber are available, and a variety of stress grades are available for each species group. Tabulated design stresses depend on the size category of a member in addition to grade. Size categories include dimension lumber, beams and stringers, posts and timbers, among others. The nominal size of a member (such as  $4 \times 12$ ) is used for call-out purposes, but actual cross-sectional dimensions are less. Most grading of lumber is done by visual inspection, but some material is machine stress rated. The maximum size of sawn lumber is limited by tree size.
2. *Structural glued-laminated timber* (glu-lam). Lumber formed by gluing together small pieces

(usually 2 in. nominal thickness) of wood to form virtually any size structural member. Laminating stock is usually from a western species group or southern pine. Tabulated stresses for glu-lam are generally larger than for sawn lumber because higher-quality wood can be optimized. For example, bending combinations of laminations have higher-quality laminating stock placed in the areas of higher bending stress (i.e., at the top and bottom of a beam) and lower-quality laminating stock near the neutral axis (i.e., at the center). See Fig. 31.1.

3. *Structural composite lumber (SCL)*. Lumber that is a reconstituted wood product. Because of its manufacturing process, SCL has even higher stress values than glulam.
  - (a) *Laminated veneer lumber (LVL)*. Lumber formed by gluing together thin sheets of wood known as *veneers*.
  - (b) *Parallel strand lumber (PSL)*. Lumber formed by gluing together thin, narrow pieces of wood known as *strands*.
4. *Round timber poles and piles*.
5. *Structural-use panels*. Usually 4 ft × 8 ft panels of wood with directional properties used for sheathing and other structural applications.
  - (a) *Plywood*.
  - (b) *Oriented strand board (OSB)*.

These products may be used individually as building components, or they may be combined in a manufacturing plant to form composite structural products such as prefabricated wood trusses, wood I joists, or factory-built roof or wall panels.

**Figure 31.1** Layup of laminating stock in a glu-lam beam. Stronger wood is located at points of higher stress.



The basic reference for structural design in wood is the *National Design Specification* [AF&PA, 1991]. The basic reference for plywood is the *Plywood Design Specification* [APA, 1986]. Detailed examples and additional references are given in Breyer [1993] and the *NDS Commentary* [AF&PA, 1993]. Basic building code criteria are available from such sources as the *UBC* [ICBO, 1991], which is one of the three model building codes used in the U.S.

### 31.3 Member Design

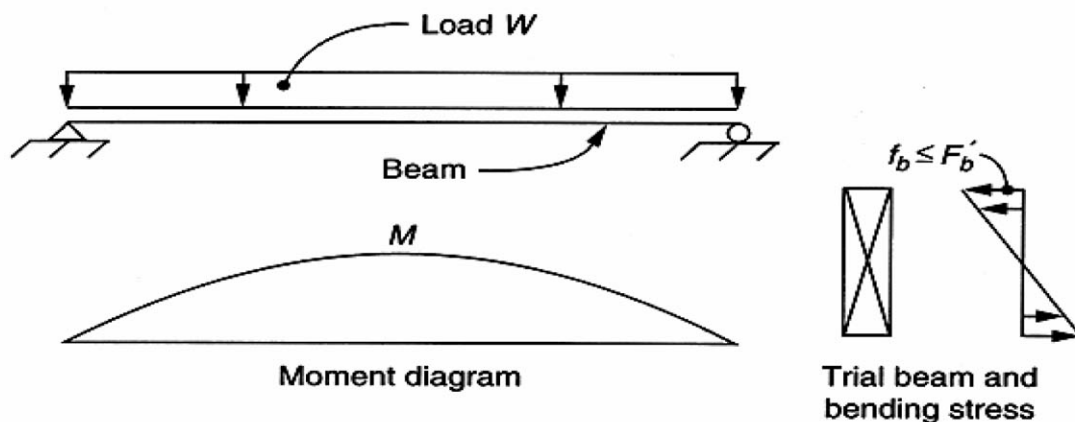
Wood structures are designed to resist vertical (gravity) loads and lateral (wind and earthquake) forces. Generally, a series of beams and columns is used to carry gravity loads. Thus a path is formed for the progressive transfer of gravity loads from the top of the structure down into the foundation. Design includes consideration of dead load, roof live load or snow load, floor live load, and other possible loads.

The 1991 edition of the *National Design Specification* [AF&PA, 1991] introduced sweeping changes to the design equations for engineered wood structures. Revised stresses for lumber were also introduced as a result of a 12-year study known as the *in-grade testing program*. Brief introductions to the design of a wood beam and a wood column are given in the remainder of this section.

*Wood beams* are designed using familiar formulas from engineering mechanics for bending, shear, deflection, and bearing. Although the basic concepts are simple, wood design can appear to be complicated because of the nature of the material. The variability of mechanical properties for different wood products is one factor. Another is the presence of natural growth characteristics such as knots, density, slope of grain, and others. The natural growth characteristics in wood have led to the development of a relatively involved lumber-grading system. However, the basic design procedure is straightforward. Basic formulas from strength of materials are used to evaluate the *actual stress* in a member. The actual stress is then checked to be less than or equal to an *allowable stress* for the species, grade, and size category. The design size is accepted or revised based on this comparison.

The allowable stress is where the unique nature of the material is taken into account. Determination of allowable stress begins by first finding the tabulated stress for a given species, stress grade, and size category. The tabulated stress applies directly to a set of base conditions (e.g., normal duration of load, dry service conditions, standard size, normal temperature, and so on). The tabulated stress is then subjected to a series of adjustment factors, which converts the base conditions for the table to the conditions for a particular design. See Fig. 31.2.

**Figure 31.2** Bending stress in a wood beam. A trial member size is structurally safe if the actual stress  $f_b$  is less than or equal to the allowable stress  $F'_b$ .



For example, the bending stress in a wood beam is checked as follows:

$$f_b = \frac{Mc}{I} = \frac{M}{S} \leq F'_b$$

where

$f_b$  = actual bending stress, psi

$M$  = moment in beam, in.-lb

$c$  = distance from neutral axis to extreme fiber, in.

$I$  = moment of inertia, in.<sup>4</sup>

$S$  = section modulus, in.<sup>3</sup>

$F'_b$  = allowable bending stress, psi

$= F_b \times (\text{series of adjustment factors})$

$= F_b \times (C_D \times C_M \times C_L \times C_F \times C_t \times x \dots)$

$F_b$  = tabulated bending stress, psi

$C_D$  = adjustment factor for duration of load (**load duration factor**)

$C_M$  = adjustment factor for high moisture conditions (**wet service factor**)

$C_L$  = adjustment factor for lateral torsional buckling (beam stability)

$C_F$  = adjustment factor for size effect (**size factor**)

$C_t$  = adjustment factor for high temperature applications (**temperature factor**)

$x \dots$  = any other adjustment factor that may apply

For the common case of a continuously braced beam in an enclosed building (dry service at normal temperatures), a number of the adjustment factors default to unity. Thus, the complicated nature of the problem is often simplified for frequently encountered design conditions.

Wood columns are checked in a similar manner (See [Fig. 31.3](#)) using the following formula for axial compressive stress:

$$f_c = \frac{P}{A} \leq F'_c$$

where

$f_c$  = actual column stress, psi

$P$  = axial column load, lb

$A$  = cross-sectional area of column, in.<sup>2</sup>

$F'_c$  = allowable column stress, psi

$= F_c \times (\text{series of adjustment factors})$

$= F_c \times (C_D \times C_M \times C_P \times C_F \times C_t \times x \dots)$

$C_D$  = adjustment factor for duration of load

$C_M$  = adjustment factor for high-moisture conditions (wet service)

$C_P$  = adjustment factor for column buckling (stability)

$C_F$  = adjustment factor for size effect

$C_t$  = adjustment factor for high-temperature use

$x \dots$  = any other adjustment factor that may apply

The column stability factor  $C_P$  is a coefficient that measures the tendency of the column to buckle between points of lateral support.  $C_P$  is based on the slenderness ratio

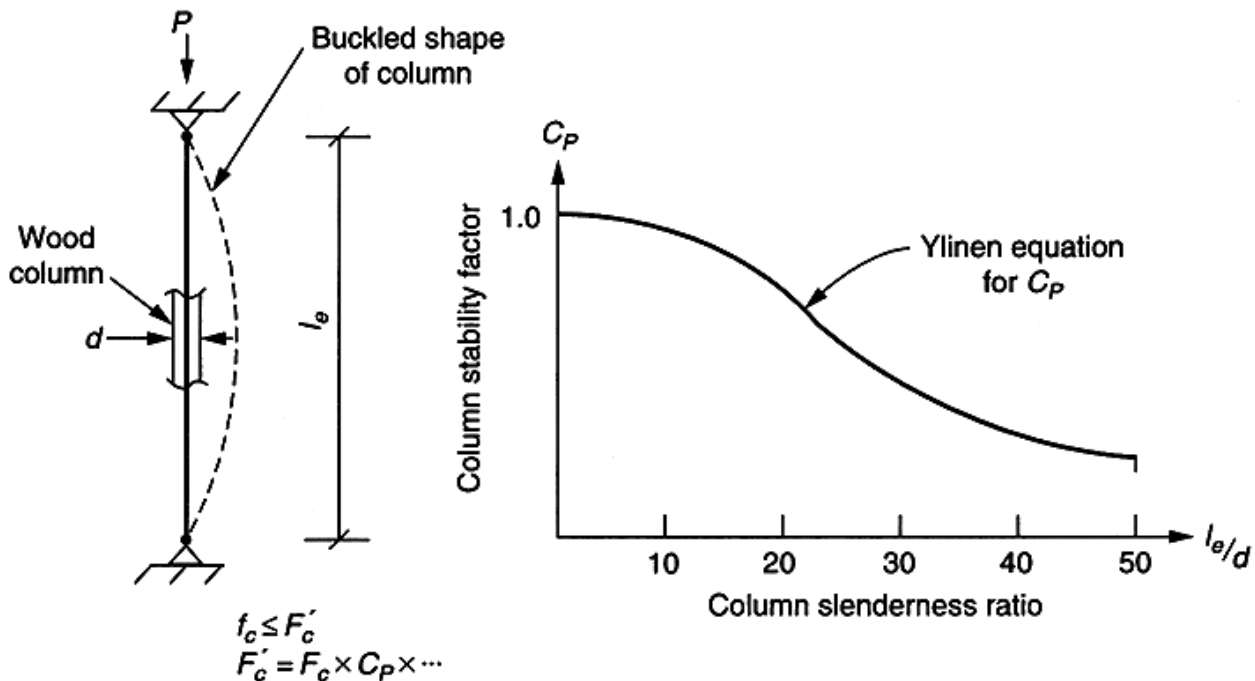
$$l_e/r = \text{column slenderness ratio}$$

where

$l_e$  = effective column length, in.

$r$  = radius of gyration associated with the axis of column buckling, in.

**Figure 31.3** Compressive stress in a wood column. A trial column size is judged to be safe if the actual stress  $f_c$  is less than or equal to the allowable stress  $F'_c$ . Column buckling is taken into account by the Ylinen equation in the column stability factor  $C_P$ .



For a column with a rectangular cross section the radius of gyration is directly proportional to the cross-sectional dimension of the member, and the slenderness ratio becomes

$$l_e/d = \text{slenderness ratio for a rectangular column}$$

where  $d$  = cross-sectional dimension of column associated with the axis of column buckling, in inches.

In the most recent *NDS* [AF&PA, 1991] the column stability factor is defined by the Ylinen column equation, which is graphed in Fig. 31.3, showing the allowable compressive stress versus column slenderness ratio. However, space does not permit a detailed review of the Ylinen equation.

Wood members that are subject to combined stress (e.g., a beam-column has both a bending moment and an axial compressive force) are handled with an interaction formula. The most recent *NDS* has a new interaction formula based on work done at the U.S. Forest Products Laboratory by Zahn.

## 31.4 Connections

Connections in engineered wood structures may be made with a variety of fasteners and other materials. These include nails, staples, bolts, lag bolts (lag screws), wood screws, split ring connectors, shear plate connectors, nail plates, and prefabricated metal connection hardware. Fasteners may connect one wood member to another (wood-to-wood connection) or they may connect a wood member to a piece of steel connection hardware (wood-to-metal connection). The

most common type of loading on a fastener is perpendicular to the axis of the fastener. This may be described as a *shear-type connection*. Connections are usually either single shear or double shear. In the past the allowable design load on a shear connection was determined from a table. These tables for nails, bolts, and other fasteners were based on a limited series of tests conducted many years ago and were empirically determined.

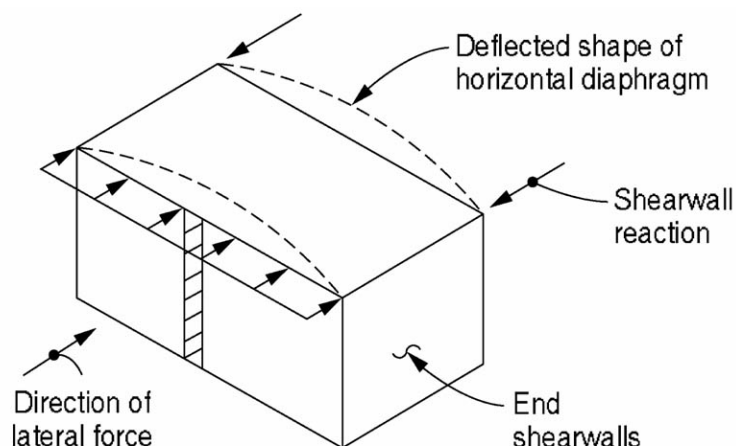
The 1991 *NDS* [AF&PA, 1991] introduced a new method for obtaining the nominal design value for these types of connections. The new method is based on engineering mechanics and is referred to as the *yield limit theory for dowel-type fasteners*. In this approach an equation is evaluated for each possible mode of failure for a given type of connection. The nominal design value for the connection is defined as the smallest load capacity from the yield equations. *NDS* tables cover commonly encountered connection problems so that it is not necessary to apply the rather complicated yield limit equations for every connection design.

## 31.5 Lateral Force Design

Lateral forces include wind and seismic forces. Although both wind and seismic forces involve vertical components, the emphasis is on the effect of horizontal forces. The common lateral force resisting systems (*LFRSs*) are moment-resisting frames, braced frames (horizontal and vertical trusses), and horizontal diaphragms and shearwalls. Most wood frame buildings use a combination of *horizontal diaphragms* and *shearwalls*. Economy is obtained in this approach because the usual sheathing materials on roofs, floors, and walls can be designed to carry lateral forces. In order to make this happen, additional nailing for the sheathing may be required, and additional connection hardware may be necessary to tie the various elements together.

In a shearwall-type building, walls perpendicular to the lateral force are assumed to span vertically between story levels. Thus, in a one-story building, the wall spans between the foundation and the roof. See Fig. 31.4. The reaction at the roof becomes a force on the horizontal diaphragm. The diaphragm acts as a large horizontal beam that is loaded in its plane. The beam spans between shearwalls. The diaphragm is composed of the roof sheathing, nailing, boundary members, and anchorage connections. The reactions on the horizontal diaphragm in turn are the forces on the shearwalls. A shearwall is designed as a beam that cantilevers vertically from the foundation. A shearwall includes the sheathing, nailing, boundary members, and connections to the horizontal diaphragm and to the foundation.

**Figure 31.4** Lateral forces are distributed from the perpendicular walls to the horizontal diaphragm. The diaphragm in turn transfers the lateral force to the shearwalls.





## Defining Terms

**Load duration factor,  $C_D$ :** A multiplying factor used to adjust the allowable stress in a wood member or connection based on the total accumulated length of time that a load is applied to a structure.  $C_D$  ranges from 0.9 for long-term (dead) loads to 2.0 for very short-term (impact) loads.

**Size factor,  $C_F$ :** A multiplying factor used to adjust the allowable stress in a wood member based on the dimensions of the cross section. Depending on the size, type of stress, and grade of lumber, the size factor may be less than, equal to, or greater than unity.

**Temperature factor,  $C_t$ :** A multiplying factor used to adjust the allowable stress in a wood member or connection based on the temperature conditions.  $C_t$  is 1.0 for normal temperatures and less than 1.0 for high temperatures.

**Wet service factor,  $C_M$ :** A multiplying factor used to adjust the allowable stress in a wood member or connection based on the moisture content of the wood.  $C_M$  is 1.0 for dry service applications and less than 1.0 for high-moisture conditions. The value of  $C_M$  depends on the type of stress and material type and may depend on other factors.

## References

- AF&PA. 1991. *National Design Specification for Wood Construction*. American Forest and Paper Association (formerly the National Forest Products Association), Washington, DC.
- AF&PA. 1993. *Commentary on the 1991 Edition of the National Design Specification for Wood Construction*. American Forest and Paper Association (formerly the National Forest Products Association), Washington, DC.
- APA. 1986. *Plywood Design Specification*. American Plywood Association, Tacoma, WA.
- Breyer, D. E. 1993. *Design of Wood Structures*, 3rd ed. McGraw-Hill, New York.
- ICBO. 1991. *Uniform Building Code*. International Conference of Building Officials, Whittier, CA.

## Further Information

Information on engineered wood structures may be obtained from a number of sources. Several governmental and industrial organizations are listed below with areas of expertise.

General information on forest products, structural engineering, and wood research:

U.S. Forest Products Laboratory (FPL)  
One Gifford Pinchot Drive  
Madison, WI 53705

Sawn lumber and connection design :

American Forest and Paper Association (AF&PA)  
1111 19th Street N.W., Suite 800  
Washington, D.C. 20036

Glu-lam:

American Institute of Timber Construction (AITC)  
7012 South Revere Parkway, Suite 140  
Englewood, CO 80112

Structural-use panels and glu-lam:

American Plywood Association (APA)  
P.O. Box 11700  
Tacoma, WA 98411-0700

Amrhein, J. E. "Masonry Design"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

**32.1 Basis of Design****32.2 Masonry Materials****32.3 Masonry Units**

Clay Masonry • Solid Clay Units • Hollow Clay Units

**32.4 Concrete Masonry**

Hollow Load Bearing Concrete Masonry Units

**32.5 Mortar**

Types of Mortar

**32.6 Grout****32.7 Unreinforced Masonry****32.8 Strength of Masonry**

Modulus of Elasticity • Specified Compressive Strength • Reinforcing Steel

**32.9 Design of Reinforced Masonry Members**

Working Stress Design • Flexural Design • Moment Capacity of a Section • Shear • Columns

**32.10 Design of Structural Members—Strength Design**

General • Strength Design Procedure • Strength Design for Sections with Tension Steel Only

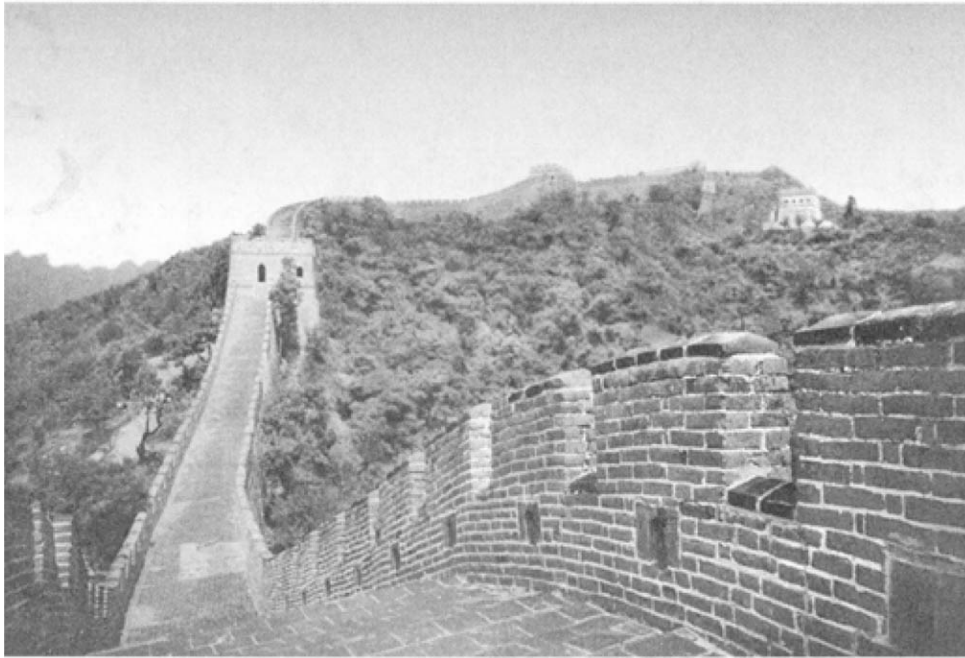
**James E. Amrhein***Masonry Institute of America*

Masonry structures have been constructed since the earliest days of mankind, not only for homes but also for works of beauty and grandeur. Stone was the first masonry unit and was used for primitive but breathtaking structures, such as the 4000-year-old Stonehenge ring on England's Salisbury Plains. Stone was also used around 2500 B.C. to build the Egyptian pyramids in Giza. The 1500-mile (2400-km) [Great Wall of China](#) was constructed of brick and stone between 202 B.C. and 220 A.D.

Masonry has been used worldwide to construct impressive structures, such as St. Basil's Cathedral in Moscow, the Taj Mahal in Agra, India, as well as homes, churches, bridges, and roads. In the U.S., masonry was used from Boston to Los Angeles and has been the primary material for building construction from the 18th to the 20th centuries.

Currently, the tallest reinforced masonry structure is the 28-story [Excalibur Hotel](#) in Las Vegas, Nevada. This large high-rise complex consists of four buildings, each containing 1008 sleeping rooms. The load-bearing walls for the complex required masonry with a specified compressive strength of 4000 psi (28 MPa).

### **The Great Wall of China.**



### **Tallest concrete masonry building in the world, Excalibur Hotel, Las Vegas, Nevada**



## **32.1 Basis of Design**

---

This chapter is based on the specification of materials, construction methods, and testing as given in the ASTM standards. The design parameters are in accordance and reprinted with permission from *Building Code Requirements and Commentary for Masonry Structures* (ACI 530, ASCE 5, TMS 402).

In addition, the Uniform Building Code, published by the International Conference of Building Officials, provides requirements and recommendations for the design and construction of masonry systems both unreinforced (plain) and reinforced. (See [Table 32.1.](#))

**Table 32.1** Notation

---

$a_b$	= Depth of stress block for balanced strength design conditions.
$A$	= Area of compression area for walls or columns.
$A_n$	= Net cross-sectional area of masonry, in. <sup>2</sup>
$A_s$	= Area of tension steel.
$A_{se}$	= Equivalent area of tension steel considering effect of vertical load.
$A_v$	= Cross-sectional area of shear reinforcement, in. <sup>2</sup>
$b$	= Width of section, in. <sup>2</sup>
$b_w$	= Width of wall beam.
$c_b$	= Depth to neutral axis for balanced strength design conditions.
$C$	= Total compression force.
$d$	= Distance from extreme compression fiber to centroid of tension reinforcement, in.
$D$	= Dead load or related internal moments and forces.
$e$	= Eccentricity of axial load, in.
$E_m$	= Modulus of elasticity of masonry in compression, psi.
$E_s$	= Modulus of elasticity of steel, psi.
$f$	= Calculated stress on section.
$f_a$	= Calculated compressive stress in masonry due to axial load only, psi.
$f_b$	= Calculated compressive stress in masonry due to flexure only; psi.
$f'_m$	= Specified compressive strength of masonry, psi.
$f_s$	= Calculated tensile or compressive stress in reinforcement, psi.
$f_t$	= Calculated tension stress on masonry, psi.
$f_v$	= Calculated shear stress in masonry, psi.
$f_y$	= Specified yield stress of steel for reinforcement and anchors, psi.
$F_a$	= Allowable compressive stress due to axial load only, psi.
$F_b$	= Allowable compressive stress due to flexure only, psi.
$F_s$	= Allowable tensile or compressive stress in reinforcement, psi.
$F_v$	= Allowable shear stress in masonry, psi.
$h$	= Effective height of column, wall, or pilaster, in.
$I$	= Moment of inertia of masonry, in. <sup>4</sup>
$j$	= Ratio of distance between centroid of flexural compressive forces and centroid of tensile forces to depth, $d$ .
$k$	= Ratio of depth of stress block to depth of section.
$l$	= Clear span between supports.
$L$	= Live load or related internal moments and forces.
$M$	= Maximum moment occurring simultaneously with design shear force $V$ at the section under consideration, in.-lb.
$n$	= Ratio of the modulus of elasticity of steel to the modulus of elasticity of masonry.
$p$	= Ratio of tensile steel area to total area of section, $bd$ .
$P$	= Design axial load, lb.
$P_u$	= Factored load on section, strength design.
$r$	= Radius of gyration, in.
$s$	= Spacing of reinforcement, in.
$S$	= Section modulus.
$t$	= Nominal thickness of wall.
$T$	= Total tension force on section.
$v$	= Shear stress, psi.
$V$	= Design shear force.
$w$	= Load or weight per unit length or area.
$W$	= Wind load or related internal moments and forces or total uniform load.

---

## 32.2 Masonry Materials

---

The principal materials used in plain masonry are the masonry units, mortar plus grout, and reinforcing steel for reinforced masonry. These materials are assembled into homogeneous structural systems.

## 32.3 Masonry Units

---

The masonry units considered are clay brick, concrete brick, hollow clay bricks, and hollow concrete blocks. Masonry units are available in a variety of sizes, shapes, colors, and textures.

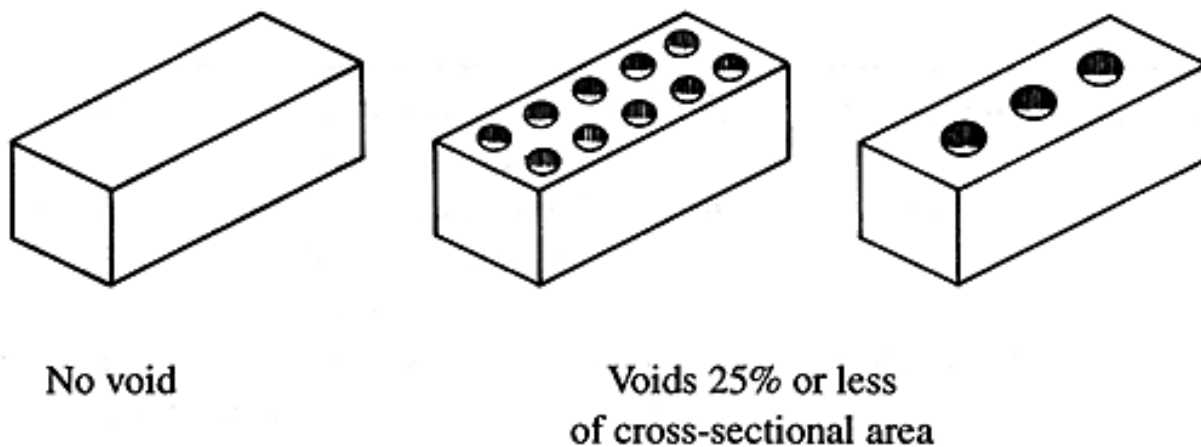
### Clay Masonry

Clay masonry is manufactured to comply with the ASTM C 62, *Specification for Building Brick (Solid Masonry Units Made From Clay or Shale)*; C 216, *Specification for Facing Brick (Solid Masonry Units Made From Clay or Shale)*; and C 652, *Specification for Hollow Brick (Hollow Masonry Units Made From Clay or Shale)*. It is made by firing clay in a kiln for 40 to 150 hours, depending upon the type of kiln, size and volume of the units, and other variables. For building brick and face brick the temperature is controlled between 1600°F (870°C) and 2200°F (1200°C), whereas the temperature ranges between 2400°F (1315°C) and 2700°F (1500°C) for fire brick.

### Solid Clay Units

A solid clay masonry unit, as specified in ASTM C 62 and C 216, is a unit whose net cross-sectional area, on every plane parallel to the bearing surface, is 75% or more of its gross cross-sectional area measured in the same plane. A solid brick may have a maximum coring of 25%. See Fig. 32.1.

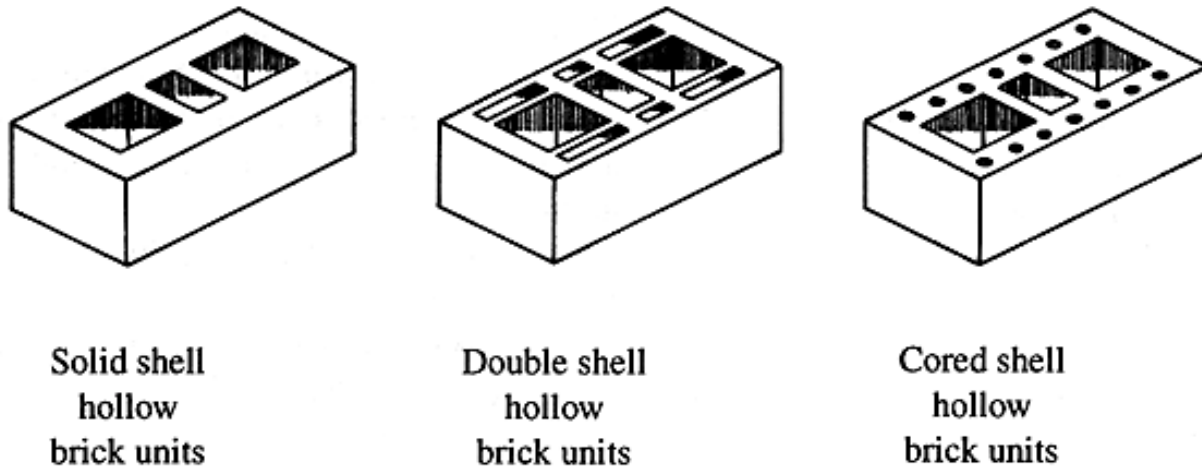
**Figure 32.1** Solid clay brick.



### Hollow Clay Units

A hollow clay masonry unit, as specified in ASTM C 652, is a unit whose net cross-sectional area in every plane parallel to the bearing surface is less than 75% of its gross cross-sectional area measured in the same plane. See Fig. 32.2.

**Figure 32.2** Hollow clay brick.



## 32.4 Concrete Masonry

---

Concrete masonry units for load-bearing systems can be either concrete brick as specified by ASTM C 55, *Specification for Concrete Building Brick*, or hollow load-bearing concrete masonry units as specified by ASTM C 90, *Specification for Hollow Load-Bearing Concrete Masonry Units*.

Concrete brick and hollow units are primarily made from portland cement, water, and suitable aggregates with or without the inclusion of other materials and may be made from lightweight or normal weight aggregates or both.

### Hollow Load Bearing Concrete Masonry Units

ASTM C 90-90, *Specification for Load Bearing Concrete Masonry Units*, requires all load-bearing concrete masonry units to meet the requirements of Grade N designation. The types of hollow concrete units are

Type I. For moisture-controlled concrete brick.

Type II. Non-moisture-controlled units need not meet water absorption requirements.

## 32.5 Mortar

---

Mortar is a plastic mixture of materials used to bind masonry units into a structural mass. It is used for the following purposes:

1. It serves as a bedding or seating material for the masonry units.
2. It allows the units to be leveled and properly placed.
3. It bonds the units together.
4. It provides compressive strength.
5. It provides shear strength, particularly parallel to the wall.
6. It allows some movement and elasticity between units.

7. It seals irregularities of the masonry units.
8. It can provide color to the wall by using color additives.
9. It can provide an architectural appearance by using various types of joints.

## Types of Mortar

The requirements for mortar are provided in ASTM C 270, *Mortar for Unit Masonry*. There are four types of mortar, which are designated M, S, N, and O. The types are identified by every other letter of the word MaSoNwOrk.

*Proportion specifications* limit the amount of the constituent parts by volume. Water content, however, may be adjusted by the mason to provide proper workability under various field conditions. The most common cement-lime mortar proportions by volume are:

*Type M mortar.* 1 portland cement; 1/4 lime; 3 1/2 sand

*Type S mortar.* 1 portland cement; 1/2 lime; 4 1/2 sand

*Type N mortar.* 1 portland cement; 1 lime; 6 sand

*Type O mortar.* 1 portland cement; 2 lime; 9 sand

## 32.6 Grout

---

Grout is a mixture of portland cement, sand pea gravel, and water mixed to fluid consistency so that it will have a slump of 8 to 10 inches (200 to 250 mm). Requirements for grout are given in ASTM C 476, *Grout for Masonry*.

Grout is placed in the cores of hollow masonry units or between wythes of solid units to bind the reinforcing steel and the masonry into a structural system. Additionally, grout provides:

1. More cross-sectional area, allowing a grouted wall to support greater vertical and lateral shear forces than a nongrouted wall
2. Added sound transmission resistance, thus reducing the sound passing through the wall
3. Increased fire resistance and an improved fire rating of the wall
4. Improved energy storage capabilities of a wall
5. Greater weight, thus improving the overturning resistance of retaining walls

## 32.7 Unreinforced Masonry

---

Unreinforced masonry considers the tensile resistance of masonry for the design of structures. The effects of stresses in reinforcement, if present, are neglected and all forces and moments are resisted by the weight of the masonry and the tension and compression capabilities of the system.

The stress due to flexural moment is  $f_t = OAM/S$ , where  $M$  is the moment on the wall and  $S$  is



the section modulus. The condition is generally limited to the allowable flexural tension stress shown in [Table 32.2](#).

**Table 32.2** Allowable Flexural Tension for Clay and Concrete Masonry, psi\* (kPa)

Masonry Type	Mortar Types							
	Portland Cement/Lime				Masonry Cement and Air-Entrained Portland Cement Lime			
	M or S	kPa	N	kPa	M or S	kPa	N	kPa
Normal to bed joints								
Solid units	40	276	30	207	24	166	15	103
Hollow units*								
UngROUTed	25	172	19	131	15	103	9	62
Fully grouted	68	470	58	400	41	283	26	180
Parallel to bed joints in running bond								
Solid units	80	550	60	415	48	330	30	207
Hollow units								
UngROUTed and partially grouted	50	345	38	262	30	207	19	131
Fully grouted	80	550	60	415	48	330	30	207

\*For partially grouted masonry allowable stresses shall be determined on the basis of linear interpolation between hollow units, which are fully grouted or ungrouted, and hollow units based on amount of grouting.

**Example 32.1—Thickness of Unreinforced Masonry Wall.** An unreinforced building wall is 10 ft (3 m) high and spans between footing and roof ledger. It could be subjected to a wind force of 15 psf (103 kPa) should an 8" (200 mm) CMU wall or a 12" (300 mm) CMU wall be used.

**Solution:**

Moment on wall

Assumed pinned top and bottom

$$M = \frac{wh^2}{8} = \frac{15 \cdot 10^2}{8} = 187.5 \text{ ft} \cdot \text{lb} / \text{ft} \quad (834 \text{ N} \cdot \text{m} / \text{m})$$

Allowable tension stress type M or S mortar = 25 psi (172 kPa)

Stress may be increased 1/3 due to temporary wind force

$$S = \frac{M}{1.33f_t} = \frac{187.5 \times 12}{1.33 \times 25} = 67.5 \text{ in.}^3 (1.1 \cdot 10^6 \text{ mm}^3)$$

For section modulus at 8" (200 mm) CMU (see [Fig. 32.3](#)):

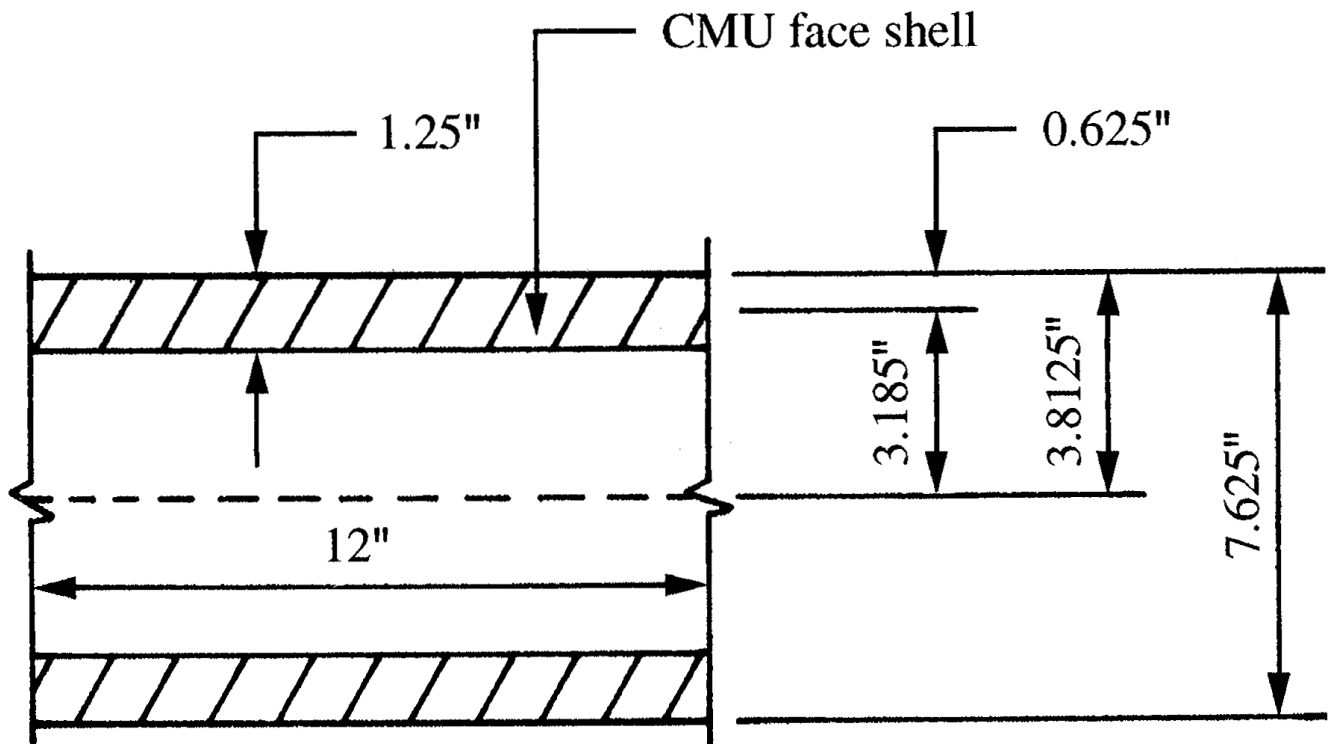
$$\begin{aligned}
 I &= 2 \left[ \frac{bd^3}{12} + bdx^2 \right] = 2 \left[ \frac{12 \times 1.25^3}{12} + 12 \times 1.25 \times 3.1875^2 \right] \\
 &= 2 [1.95 + 152.40] = 308.7 \text{ in.}^4 (128.5 \cdot 10^6 \text{ mm}^4) \\
 S &= \frac{308.7 \times 2}{7.625} = 81 \text{ in.}^3 (1.3 \cdot 10^6 \text{ mm}^3)
 \end{aligned}$$

At 12'' (300 mm) CMU (see Fig. 32.4):

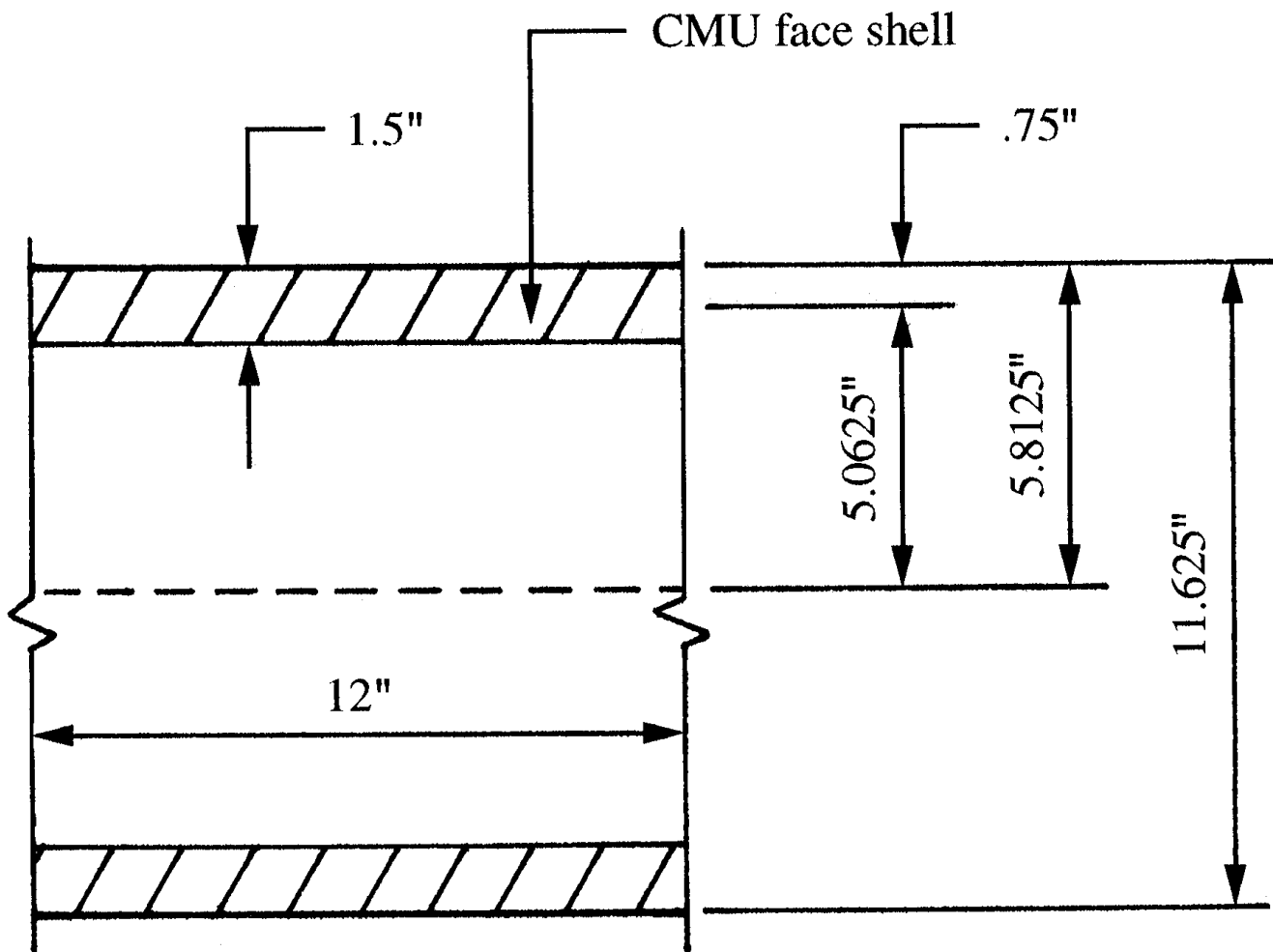
$$\begin{aligned}
 I &= 2 \left[ \frac{bd^3}{12} + bdx^2 \right] \\
 &= 2 \left[ \frac{12 \times 1.5^3}{12} + 12 \times 1.5 \times 5.0625^2 \right] \\
 &= 2 [3.375 + 461.3] = 929.4 \text{ in.}^4 (386.8 \cdot 10^6 \text{ mm}^4) \\
 S &= \frac{I}{t/2} = \frac{2 \times 929.4}{11.625} = 159.9 \text{ in.}^3 (2.6 \cdot 10^6 \text{ mm}^3)
 \end{aligned}$$

Thus, 8'' CMU = 81.0 in.<sup>3</sup> (1.3 · 10<sup>6</sup> mm<sup>3</sup>); 12'' CMU = 159.9 in.<sup>3</sup> (2.6 · 10<sup>6</sup> mm<sup>3</sup>). Use 8'' (200 mm) CMU.

**Figure 32.3** Plan cross section of 8'' CMU wall, face shells only.



**Figure 32.4** Plan cross section of 12" CMU wall, face shells only.



**Example 32.2 — Vertical and Lateral Load on Unreinforced Masonry Wall.** If a wall is subjected to a 20 psf (958 kPa) wind, and is 15 ft (4.6 m) high, and carries 2000 plf (96 kPa), what thickness concrete masonry unit should be used?

**Solution:**

$$M = \frac{wh^2}{8} = \frac{20 \times 15^2}{8} = 563 \text{ ft} \cdot \text{lb/ft} \text{ (2500 N} \cdot \text{m/m)}$$

Try 8" (200 mm) CMU.

$$\begin{aligned}
S &= 81 \text{ in.}^3 (1.3 \cdot 10^6 \text{ mm}^3) \\
\frac{P}{A} \pm \frac{M}{S} &= \frac{2000}{2 \times 12 \times 1.25} \pm \frac{563 \times 12}{81} \\
&= 66.7 \pm 83 \\
&= 150 \text{ psi compression } (1.0 \text{ MPa}) \\
&= 16.3 \text{ psi tension } (115 \text{ kPa})
\end{aligned}$$

Tension is less than  $25 \times \frac{4}{3} = 33.3$  psi (230 kPa) allowable tension, so 8" (200 mm) CMU is satisfactory.

## 32.8 Strength of Masonry

The ultimate compressive strength of the masonry assembly is given the symbol  $f'_{mu}$  to distinguish it from the specified compressive strength  $f'_m$ .

## Modulus of Elasticity

For steel reinforcement,  $E_s = 29\,000\,000$  psi (199 955 MPa). For concrete masonry, see [Table 32.3](#).

**Table 32.3** Modulus of Elasticity for Concrete Masonry\*

Net Area Compressive Strength of Units		$E_m$ for Type N Mortar		$E_m$ for Type M or S Mortar	
psi	MPa	psi $\times 10^6$ *	MPa $\times 10^3$	psi $\times 10^6$ *	MPa $\times 10^3$
6000 and greater	41.4			3.5	24.5
5000	34.5	2.8	19.6	3.2	22.4
4000	27.6	2.6	18.2	2.9	20.3
3000	20.7	2.3	16.1	2.5	17.5
2000	13.8	1.8	12.6	2.2	15.4
1500	10.3	1.5	10.5	1.6	11.2

\*Linear interpolation permitted.

## Specified Compressive Strength

For specified compressive strength of concrete masonry, see [Table 32.4](#).

**Table 32.4** Compressive Strength of Concrete Masonry

Net Area Compressive Strength for Concrete Masonry Units				Net Area Compressive Strength of Masonry, $f'_m$ (specified)	
Type M or S mortar		Type N mortar		psi*	MPa
psi*	MPa	psi*	MPa		
1250*	(8.5)**	1300	9.0	1000	6.9
1900	(13.0)	2150	14.6	1500	10.3
2800	(19.3)	3050	20.6	2000	13.8
3750	(26.0)	4050	27.6	2500	17.2
4800	(33.1)	5250	36.5	3000	20.7

\*For units of less than 4 in. (100 mm) height, 85% of the values listed.

Note: Compressive strength based on the compressive strength of concrete masonry units and type of mortar used in construction.

## Reinforcing Steel

Reinforcing steel in masonry has been used extensively in the West since the 1930s, revitalizing the masonry industry in earthquake-prone areas. Reinforcing steel extends the characteristics of ductility, toughness, and energy absorption that are so necessary in structures subjected to the dynamic forces of earthquakes.

Reinforcing steel may be either Grade 40, with a minimum yield strength of 40 000 psi (276 MPa), or Grade 60, minimum yield strength of 60 000 psi (414 MPa).

Allowable stresses for reinforcing steel are as follows. Tension stress in reinforcement shall not exceed the following:

Grade 40 or Grade 50 reinforcement	20 000 psi (138 MPa)
Grade 60 reinforcement	24 000 psi (165MPa)
Wire joint reinforcement	30 000 psi (207MPa)

Compression stress has these restrictions:

1. The compressive resistance of steel reinforcement is neglected unless lateral reinforcement is provided to tie the steel in position.
2. Compressive stress in reinforcement may not exceed the lesser of  $0.4f_y$  or 24 000 psi (165 MPa).

## 32.9 Design of Reinforced Masonry Members

### Working Stress Design

Reinforced masonry members are designed by elastic analysis using service loads and permissible stresses, which considers that the reinforcing steel resists tension forces and the masonry and grout resists compression forces.

The design and analysis of reinforced masonry structural systems have been by the straight line, elastic working stress method. In working stress design (WSD), the limits of allowable stress for

the materials are established based on the properties of each material.

The procedure presented is based on the working stress or straight line assumptions where all stresses are in the elastic range and:

1. Plane sections before bending remain plane during and after bending.
2. Stress is proportional to strain, which is proportional to distance from the neutral axis.
3. Modulus of elasticity is constant throughout the member.
4. Masonry carries no tensile stresses.
5. Span of the member is large compared to the depth.
6. Masonry elements combine to form a homogeneous and isotropic member.
7. External and internal moments and forces are in equilibrium.
8. Steel is stressed about the center of gravity of the bars equally.
9. The member is straight and of uniform cross section.

## Flexural Design

The basis of the flexural equations for working stress design, WSD, is the concept of the modular ratio. The modular ratio,  $n$ , is the ratio of the modulus of elasticity of steel,  $E_s$ , to the modulus of elasticity of masonry,  $E_m$ .

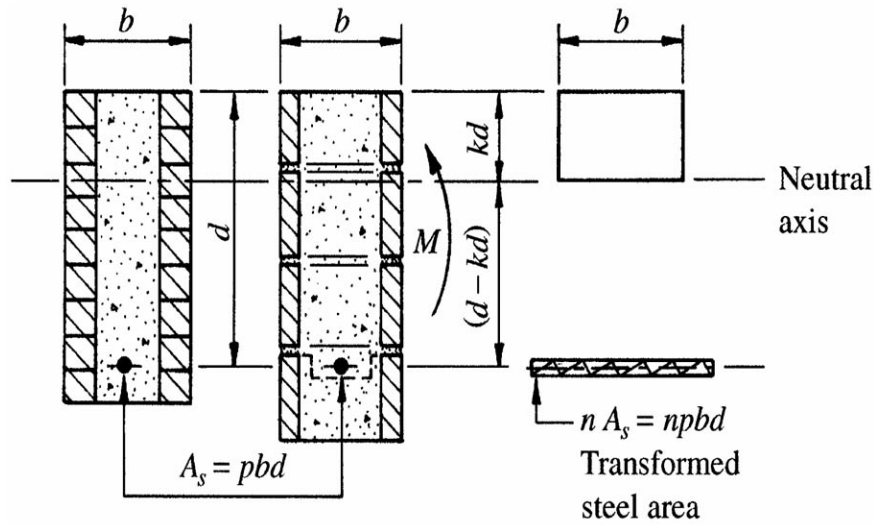
$$n = \frac{E_s}{E_m}$$

By use of the modular ratio,  $n$ , the steel area can be transformed into an equivalent masonry area. The strain is in proportion to the distance from the neutral axis and therefore the strain of the steel can be converted to stress in the steel. In order to establish the ratio of stresses and strains between the materials, it is necessary to locate the neutral axis.

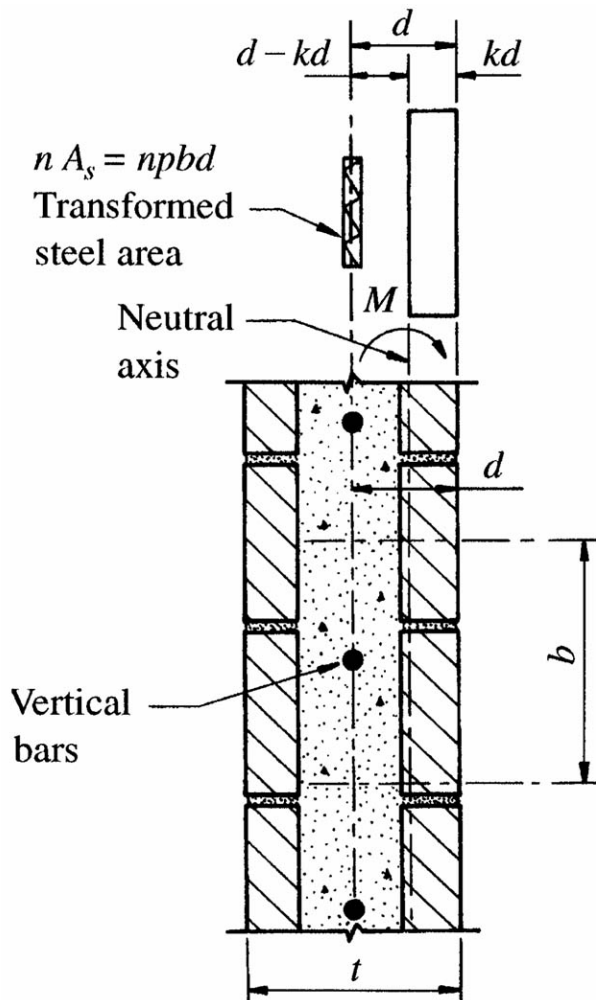
The location of the neutral axis is defined by the dimensions,  $kd$ , which is dependent on the modular ratio,  $n$ , and the reinforcing steel ratio,  $p = A_s/bd$ . For a given modular ratio,  $n$ , the neutral axis can be raised by decreasing the amount of steel (reducing  $p$ ) or lowered by increasing the amount of steel (increasing  $p$ ). See [Figs. 32.5](#) and [32.6](#). Solving for  $k$ ,

$$k = \sqrt{(np)^2 + 2np} - np$$

**Figure 32.5** Location of neutral axis for a beam.



**Figure 32.6** Location of neutral axis for a wall, in plane of wall.



## Moment Capacity of a Section

The moment capacity of a reinforced structural masonry wall or beam can be limited by the allowable masonry stress (overreinforced), allowable steel stress (underreinforced), or both, in which case it would be a balanced design condition.

When a member is designed for the specified loads and the masonry and reinforcing steel are stressed to their maximum allowable stresses, the design is said to be a "balanced" design. This balanced design is different from the balanced design for the strength design method. For working stresses, balanced design occurs when the masonry is stressed to its maximum allowable compressive stress and the steel is stressed to its maximum allowable tensile stress.

However, in many cases, the "balanced" design does not satisfy the conditions for the materials available or for the predetermined member size or the economy of the project. It may be advantageous to understress (underreinforce) the masonry or understress (overreinforce) the steel so that the size of the member can be maintained.

The moment capability of a section based on the steel stress is defined as

$M_s = \text{force} \times \text{moment arm}$ , where:

$$\text{Force in the steel, } T = A_s f_s = p b d f_s$$

$$\text{Moment arm} = j d$$

$$M_s = T \times j d = A_s f_s j d$$

$$M_s = p b d f_s j d = f_s p j b d^2$$

The moment capability of a section based on the masonry stress is defined as

$M_m = \text{force} \times \text{moment arm}$ , where:

$$\text{Force in the masonry, } C = \frac{1}{2} f_b (k d) b = \frac{1}{2} f_b k b d$$

$$\text{Moment arm} = j d$$

$$M_m = C \times j d = \left( \frac{1}{2} f_b k b d \right) \times (j d)$$

$$M_m = \frac{1}{2} f_b k j b d^2$$

**Example 32.3—Determination of Moment Capacity of a Wall.** A partially grouted 8" (200 mm) concrete masonry wall with type S mortar is reinforced with #5 bars at 32" (813 mm) o.c. The steel is 5.3" (135 mm) from the compression face and is Grade 60. If  $f'_m = 2500$  psi (17.2 MPa), what is the moment capacity of the wall?

**Solution:** For  $f'_m = 2500$  psi (17.2 MPa),

$$F_b = 0.33 f'_m = 833 \text{ psi (5.6 MPa)} \quad \text{max. allowable compression}$$

$$E_m = 2\,400\,000 \text{ psi (16\,550 MPa)} \quad (\text{see Table 32.3})$$

Also, for  $f_y = 60\,000$  psi (414 MPa),

$$F_s = 24\,000 \text{ psi (165 MPa)} \quad \text{max. allowable tension}$$

$$E_s = 29\,000\,000 \text{ psi (199\,955 MPa)}$$

For steel ratio,



$$p = \frac{A_s}{bd}$$

$$= \frac{0.31}{32 \times 5.3} = 0.0018$$

For modular ratio,

$$n = \frac{E_s}{E_m}$$

$$= \frac{29\,000\,000}{2\,400\,000} = 12.1$$

Furthermore,

$$np = 12.1 \times 0.0018 = 0.022$$

$$k = \sqrt{(np)^2 + 2n} - np$$

$$= \sqrt{(0.022)^2 + 2 \times 0.0022} - 0.022$$

$$= 0.189$$

$$kd = 0.189 \times 5.3 = 1.00 \text{ in. (25 mm)}$$

The neutral axis falls on the shell of CMU.

$$j = 1 - \frac{k}{3} = 1 - \frac{0.189}{3} = 0.937$$

$$M_m = \frac{1}{2} f_b k j b d^2 = \frac{1}{2} (833) (0.189) (0.937) (12) (5.3)^2$$

$$= 24\,863 \text{ in.-lb /ft (9027 N} \cdot \text{m/m)}$$

$$= 2.07 \text{ ft k/ft (8940 N} \cdot \text{m/m)}$$

$$M_s = f_s p j b d^2 = 24\,000 (0.0021) (0.937) (12) (5.3)^2$$

$$= 15\,918 \text{ in.-lb /ft (5800 N} \cdot \text{m/m)}$$

$$= 1.33 \text{ ft k/ft (5800 N} \cdot \text{m/m)} \leftarrow \text{Controls}$$

## Shear

Structural elements such as walls, piers, and beams are subjected to shear forces as well as flexural stresses. The unit shear stress is computed based on the formula

$$f_v = \frac{V}{bjd}$$

### Beam Shear

When masonry flexural members are designed to resist shear forces without the use of shear reinforcing steel, the calculated shear stress is limited to  $1.0(f'_m)^{1/2}$ , 50 psi max. If the unit shear stress exceeds the allowable masonry shear stress, all the shear stress must be resisted by reinforcing steel.

For flexural members with reinforcing steel resisting all the shear forces, the maximum allowable shear stress is  $3.0(f'_m)^{1/2}$  psi with 150 psi as a maximum. The steel resists the shear by tension and it must be anchored in the compression zone of the beam or the wall.

The unit shear,  $f_v$ , is used to determine the shear steel spacing based on the formula:

$$\text{Spacing, } s = \frac{A_v F_s}{f_v b}$$

$$\text{Unit shear stress, } f_v = \frac{A_v F_s}{b_s}$$

For continuous or fixed beams, the shear value used to determine the shear steel spacing may be taken at a distance  $d/2$  from the face of the support. The shear value at the face of the support should be used to calculate the shear steel spacing in simple beams.

The maximum spacing of shear steel should not exceed  $d/2$ . The first shear-reinforcing bar should be located at half the calculated spacing but no more than  $d/4$  from the face of support.

## Columns

Columns are vertical members that basically support vertical loads. They may be plain masonry or reinforced masonry. Reduction in the load-carrying capacity is based on the  $h/r$  ratio, where  $h$  is the unbraced height and  $r$  is the minimum radius of gyration for the unbraced height.

The reduction factor for members having an  $h/r$  ratio not exceeding 99 is  $[1 - (h/140r)^2]$ . For members with an  $h/r$  greater than 99 the factor is  $(70r/h)^2$ . The maximum allowable axial stress for walls or plain columns is  $F_a = \frac{1}{4}f'_m$  times the reduction factor.

The maximum allowable axial load on a reinforced masonry column is:

$$P_a = (0.25f'_m A_e + 0.65A_s F_{sc}) (\text{reduction factor})$$

and is limited to  $P_a \leq 1/4P_e$ , where

$$P_e = \frac{\pi^2 E_m I}{h^2} \left( 1 - 0.577 \frac{e}{r} \right)^3$$

The maximum allowable unit axial stress is  $F_a = P_a/A_e$ .

The reduction factor based on the  $h/r$  ratio is the same for reinforced columns and for walls. The same consideration is made for the determination of the effective height,  $h$ .

The effective thickness,  $t$ , is the specified thickness in the direction considered. For nonrectangular columns the effective thickness is the thickness of a square column.

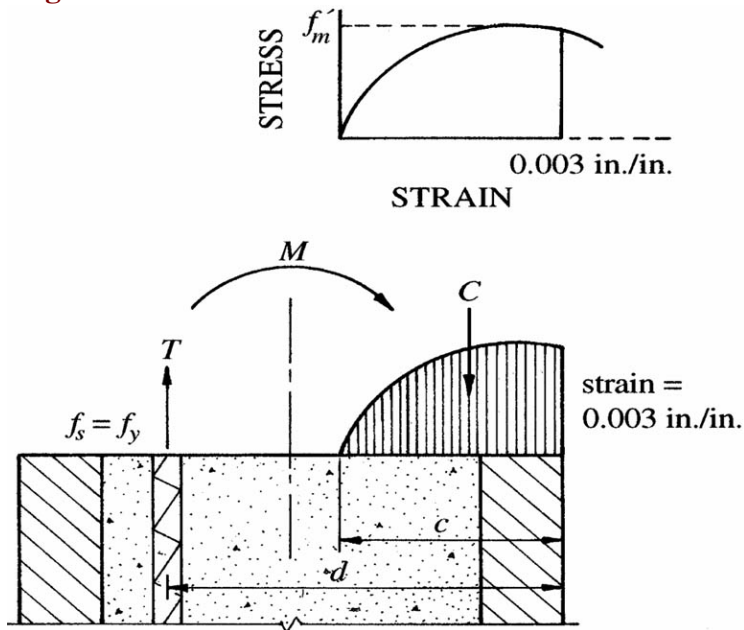
## 32.10 Design of Structural Members—Strength Design

### General

The structural design of reinforced masonry is changing from the elastic working stress method to strength design procedures.

The concept of strength design states that, when a reinforced masonry section is subjected to high flexural moments, the masonry stress from the neutral axis to the extreme compression fibers conforms to the stress-strain curve of the materials as if it were being tested in compression. See Fig. 32.7.

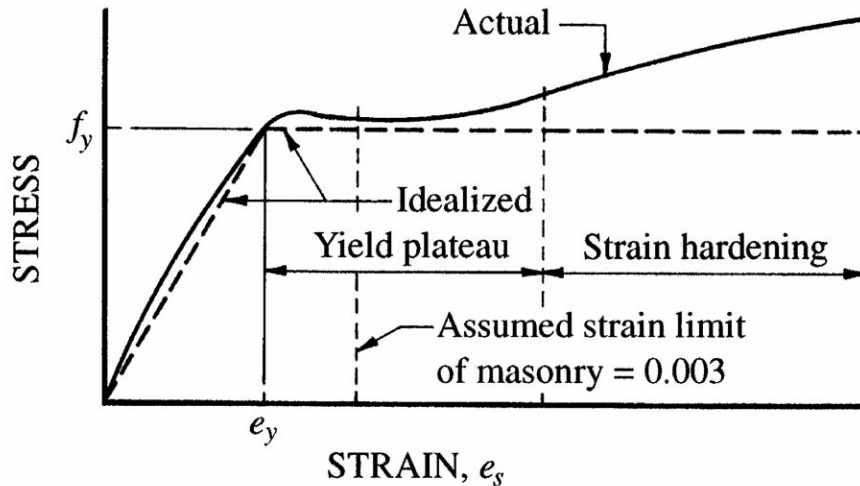
**Figure 32.7** Stress due to flexural moment for balanced condition.



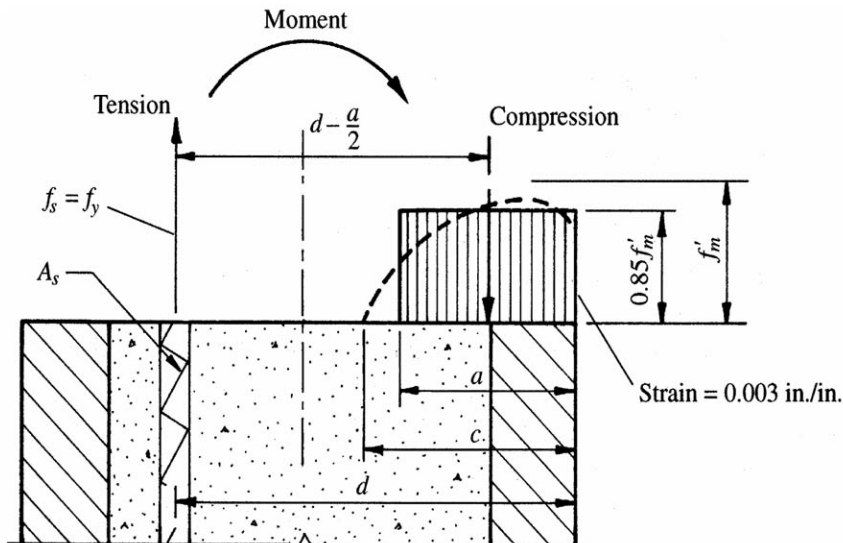
It also states that when the tension reinforcing reaches its yield stress, it will continue to elongate without an increase in moment or forces. This condition occurs at the yield plateau of the steel as shown on the stress-strain curve in Fig. 32.8.

The compressive stress block of the masonry, as shown in Fig. 32.9, is simplified from the curved or parabolic shape to a rectangular configuration. This rectangular stress block, which is called *Whitney's stress block*, is approximated as having a length of  $a$  and a height of  $0.85 f'_m$ .

**Figure 32.8** Idealized stress-strain diagram for reinforcing steel.



**Figure 32.9** Assumed stress block at yield condition.



Masonry systems have compression stress-strain curves similar to those of concrete, in that the curves are curved or parabola-shaped and that they reach a strain of at least 0.003. Accordingly, the parameters of reinforced concrete strength design are being adopted with minor changes for masonry design.

## Strength Design Procedure

There are two conditions included in strength design: load parameters and design parameters.

### Load Parameters

Service or actual loads are generally used for working stress design procedures. For strength design procedures, the actual or specified loads are increased by appropriate load factors. These load factors consider live and dead load, wind, earthquake, temperature, settlement, and earth pressure.

In addition to load factors, a capacity reduction factor,  $\phi$ , is used to adjust for the lack of perfect materials, strength, and size. The phi factor also varies for the stress considered, whether flexural or shear.

## Design Parameters

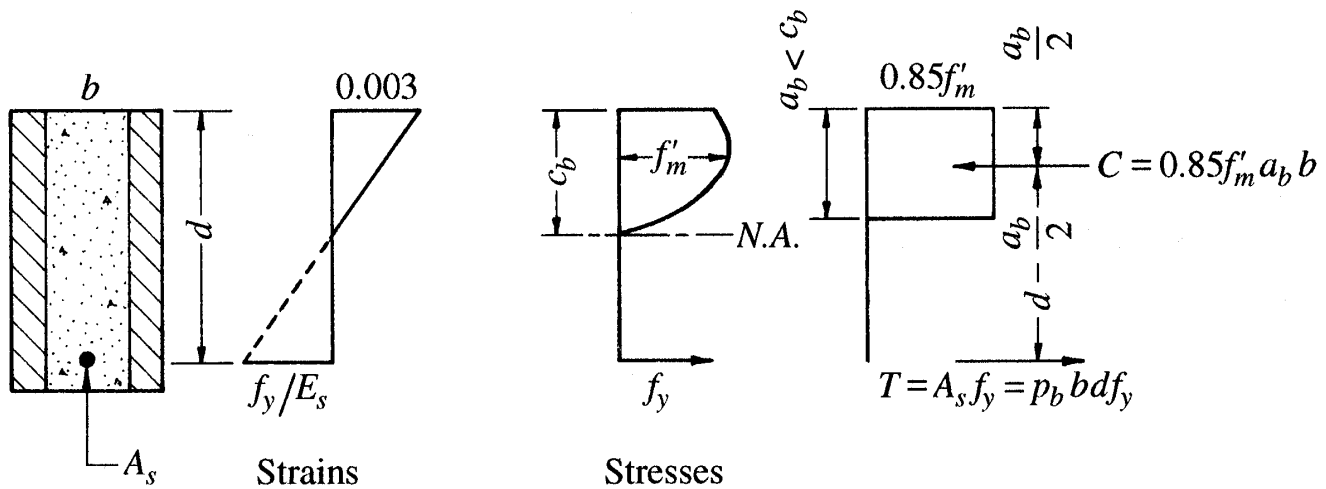
The parameters for strength design are as follows:

1. The steel is at yield stress.
2. The masonry stress block is rectangular.
3. The masonry strain is limited to 0.003 in./in.
4. The steel ratio,  $p$ , is limited to 50% of the balanced reinforcing ratio,  $p_b$ , to ensure that a ductile mechanism forms prior to brittle, crushing behavior.

## Strength Design for Sections with Tension Steel Only

As stated above, the limits for flexural design using strength methods are that the stress in the steel is at yield strength and that the strain in the masonry is at 0.003. When these conditions occur at the same moment, the section is considered to be a balanced design. See [Figure 32.10](#).

**Figure 32.10** Strain and stress distribution on a flexural member, balanced design.



The depth to the neutral axis,  $c_b$ , for a balanced design is

$$c_b = \frac{0.003}{0.003 + f_y/E_s} d = \frac{87\,000}{87\,000 + f_y} d$$

## Defining Terms

**Allowable work stress design or elastic design:** A technique based on and limiting the stress in the structural element to a value that is always in the elastic range.

**Brick:** Solid unit  $\leq 25\%$  void; hollow unit  $> 25\% < 75\%$  void.

**Grout:** Material to tie reinforcing steel and masonry units together to form a structural system

**Modular ratio:** Ratio between the modulus of elasticity of steel to the modulus of elasticity of masonry.

**Mortar:** Plastic material between units in bed and head joints.

**Steel ratio:** Area of steel to area of masonry.

**Strength design:** A technique based on capacity of structural section considering the maximum strain in masonry, yield strength of steel, load factors for various loads considered, and phi

factors for materials and workmanship.

## References

- Amrhein, J. E. 1994. *Reinforced Masonry Engineering Handbook*, 5th ed. Masonry Institute of America, Los Angeles, CA.
- Beall, C. 1984. *Masonry Design and Detailing*. Prentice Hall, Englewood Cliffs, NJ.
- Drysdale, R. G., Hamid, A. A, Baker, L. R. 1993. *Masonry Structures, Behavior and Design*. Prentice Hall, Englewood Cliffs, NJ.
- Schneider, R. A., Dickey, W. L. 1987. *Reinforced Masonry Design*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.

## Further Information

- ACI. 1992. *Building Code Requirements and Commentary for Masonry Structures; Specifications for Masonry Structures; Commentaries*. American Concrete Institute, American Society of Civil Engineers, and The Masonry Society, Detroit, MI.
- Matthys, J. H. (Ed.) 1993. *Masonry Designers Guide 1993*. ACI, ASCE, and TMS.
- Uniform Building Code*. 1994. International Conference of Building Officials, Whittier, CA.

Kreith, F. "Fluid Mechanics"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



The air flow from the wing of this agricultural airplane is made visible by a technique that uses colored smoke rising from the ground. The swirl at the wing tip traces the airplane's wake vortex, a source of problems in aerial application of spray material. The vortex, which can be seen here, exerts a powerful influence on the flow field behind the airplane. (Photo courtesy of NASA, Langley.)



# VI

## Fluid Mechanics

---

**Frank Kreith**

*University of Colorado*

**33 Incompressible Fluids** *A. T. Mc Donald*

Fundamentals of Incompressible Fluid Flow • Fluids without Relative Motion • Basic Equations in Integral Form for Control Volumes • Differential Analysis of Fluid Motion • Incompressible Inviscid Flow • Dimensional Analysis and Similitude • Internal Incompressible Viscous Flow • External Incompressible Viscous Flow

**34 Compressible Fluids** *J. D. Hoffman*

General Features • Basic Equations • Steady Quasi-One-Dimensional Flow • Equations of State and Thermodynamics • Stagnation Properties • Isentropic Flow • Nozzles • Shock Waves • Friction and Heat Transfer

**35 The Rheology of Non-Newtonian Fluids** *D. Doraiswamy*

Kinematics, Flow Classification, and Material Functions • Fluids • Constitutive Equations • Some Useful Correlations for Material Functions

**36 Airfoils/Wings** *B. R. Munson and D. J. Cronin*

Nomenclature • Airfoil Shapes • Lift and Drag Characteristics for Airfoils • Lift and Drag of Wings

**37 Boundary Layers** *E. R. Braun and P.-l. Wang*

Theoretical Boundary Layers • Reynolds Similarity in Test Data • Friction in Pipes • Noncircular Channel • Example Solutions

**38 Valves** *J. P. Tullis*

Control Valves • Air Valves • Check Valves

**39 Pumps and Fans** *R. F. Boehm*

Pumps • Fans

**40 Two-Phase Flow** *R. T. Lahey, Jr.*

Notation • Conservation Equations • Closure • Two-Phase Instabilities • Conclusion

**41 Basic Mixing Principles for Various Types of Fluid Mixing Applications** *J. Y. Oldshue*

Scaleup/Scaledown • Effect of the Circulation Time Spectrum and the Spectrum of Shear Rates on Ten Different Mixing Technologies • Computational Fluid Dynamics

**42 Fluid Measurements** *S. A. Sherif*

Fundamental Principles • Basic Equations

FLUID MECHANICS IS ONE OF THE UNDERPINNINGS for several fields of engineering. It has applications in design, energy conversion, chemical engineering, civil engineering, and manufacturing, among other areas. The treatment of fluid mechanics is based on basic concepts of thermodynamics and mechanics, but its applications are manifold and complex. The field is generally divided into compressible and incompressible flow, and this subdivision is followed here.

This section covers the basic elements of the field in sufficient detail to allow an engineer to design pipelines and ducting systems and to select the pumps and fans required to move the fluid through them, as well as the valves to control the flow. The section also includes flow over airfoils, boundary layers, two-phase flow, mixing, and measurements. These are the most important topics in fluid mechanics, but the field is vast and there are handbooks devoted entirely to it. Consequently, the presentations in this section are somewhat abbreviated and emphasize the basic elements. Given the space limitations, the authors have done an excellent job of providing an overview to the field, but for more detailed design the reader is referred to the references at the end of each chapter.

McDonald, A. T. "Incompressible Fluids"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Incompressible Fluids

---

- 33.1 Fundamentals of Incompressible Fluid Flow
- 33.2 Fluids without Relative Motion
- 33.3 Basic Equations in Integral Form for Control Volumes
- 33.4 Differential Analysis of Fluid Motion
- 33.5 Incompressible Inviscid Flow
- 33.6 Dimensional Analysis and Similitude
- 33.7 Internal Incompressible Viscous Flow
- 33.8 External Incompressible Viscous Flow

**Alan T. McDonald**

*Purdue University*

A fluid is a substance that cannot sustain shear stress while at rest; even a small shear stress causes a continuous rate of angular deformation within a fluid. Under typical conditions liquids and gases behave as fluids. Under extreme conditions, solids may exhibit fluid characteristics as in the "flow" of ice in a glacier.

Fluids are characterized by the relationship between applied shear stress and rate of angular deformation (*shear rate*). *Newtonian fluids* obey a simple linear relationship. For parallel flow this may be expressed as  $\tau_{yx} = \mu du/dy$ , where  $\tau_{yx}$  is shear stress applied in the direction of the velocity,  $y$  is distance perpendicular to velocity,  $\mu$  is dynamic viscosity or simply *viscosity*, and  $du/dy$  is rate of angular deformation. Most gases and many liquids—such as water, gasoline, and other pure substances—are closely approximated by the Newtonian fluid model.

Viscosity depends primarily on temperature at moderate pressures. Viscosity decreases sharply with increasing temperature for liquids and increases slightly for gases. At extremely high pressures, viscosities of liquids may increase significantly.

Fluid systems are commonly encountered in engineering practice. Transportation vehicles of all types—whether immersed or floating—experience viscous and pressure drag forces caused by fluid flow around the vehicle. Pipeline transportation and human circulation are fluid systems, as are convection heating and ventilating systems.

### 33.1 Fundamentals of Incompressible Fluid Flow

---

This section covers incompressible fluid flow. *Density* is defined as mass per unit volume and denoted by  $\rho$ ; fluids with constant density are *incompressible*. Liquids are nearly incompressible.

Gases are compressible, but gas flow may be treated as incompressible when the maximum speed is less than one-third of the speed of sound.

The objective of fluid flow analysis is to predict the pressure drop and pumping power for internal flow through conduits and the forces and moments on bodies in external flow. In principle this may be accomplished if the three components of velocity and the pressure are known. In practice it is often impossible to solve problems analytically; in these cases it is necessary to rely upon experimental data from tests of models and model systems.

This section covers fundamentals of incompressible fluid mechanics. First, fluids without relative motion are considered. Details of flow fields are then considered. Dimensional analysis to help simplify design of experiments and presentation of experimental data is then treated, followed by applications of the results.

The basic equations used to analyze fluid flows are conservation of mass, Newton's second law of motion (linear momentum), and the first law of thermodynamics. These equations are derived in mechanics and physics for application to fixed masses. Special forms of the equations are required to analyze moving fluids.

A *system* is defined as a fixed mass of fluid. A *control volume* is an arbitrary boundary defined in the flow field to identify a region in space. Fluid may flow through the control volume, and exchanges of heat and work with the surroundings may occur.

The best system or control volume size for analysis depends on the information sought. Integral control volumes are used to obtain overall information such as thrust of a jet engine or force exerted by a liquid jet on a surface. Differential systems and control volumes are used to obtain detailed information about flow fields, such as point-by-point variation of velocity.

For analysis, the fluid is assumed to be a *continuum*; individual molecules are not considered. Fluid velocity is a vector that varies continuously throughout the flow; the velocity field is a vector field. It is possible to resolve the velocity into scalar components. Thus  $\vec{V} = u\hat{i} + v\hat{j} + w\hat{k}$ , where  $\vec{V}$  is the velocity vector of the *fluid particle* (small volume of fluid surrounding point  $xyz$ );  $u$ ,  $v$ , and  $w$  are scalar components of velocity; and  $\hat{i}$ ,  $\hat{j}$ , and  $\hat{k}$  are unit vectors in the  $x$ ,  $y$ , and  $z$  directions, respectively.

In the most general case, velocity is a function of three space coordinates and time,  $\vec{V} = \vec{V}(x, y, z, t)$ . For steady flow there is no time dependence. The number of space coordinates defines the dimensions of the flow field;  $\vec{V} = \vec{V}(x)$  is a steady, one-dimensional flow field.

*Stress* is defined as the limiting value of force per unit area as the area is reduced to differential size. The simplest description of stress uses area elements having normals in the three coordinate directions; the infinitesimal force on each area element also may have three components. The notation  $\tau_{yx}$  signifies a shear stress acting in the  $x$  direction on an area element with normal in the  $y$  direction. The *stress field* is the continuous distribution of stresses throughout the fluid; the stress field behaves as a second-order tensor.

## 33.2 Fluids without Relative Motion

---

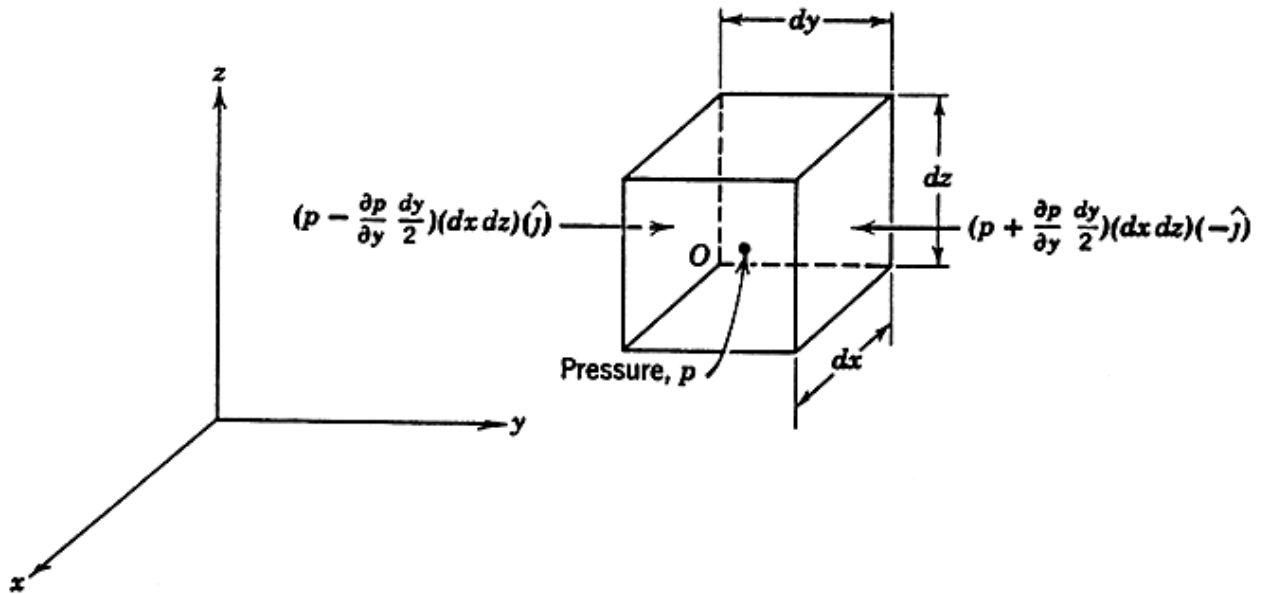
In the absence of relative motion, there can be no viscous stresses within a fluid. The only surface stress is pressure, which acts against the surface of a fluid element. Pressure must vary

continuously throughout the fluid, so it may be expanded in a Taylor series. Summing forces on an infinitesimal fluid element (Fig. 33.1) leads to the expression

$$-\nabla p + \rho \vec{g} = \rho \vec{a} \quad (33.1)$$

This result shows that the *pressure gradient*  $\nabla p$  is the negative of the net body force per unit volume. (When  $\vec{a} = 0$  this is the *basic equation of fluid statics*.)

**Figure 33.1** Differential fluid element showing pressure forces.



When the fluid is static or the acceleration field is known, Eq. (33.1) may be solved for pressure distribution. This is done most easily by expanding the pressure gradient into components (most advanced mathematics books give the *del operator*  $\nabla$  in rectangular, cylindrical, and spherical coordinates).

### 33.3 Basic Equations in Integral Form for Control Volumes

To formulate the basic equations for control volume application requires a limiting process [Fox and McDonald, 1992] or derivation of the Reynolds transport theorem [Fay, 1994]. The resulting relation between the system expression and control volume variables is

$$\left. \frac{dN}{dt} \right|_{\text{system}} = \frac{\partial}{\partial t} \int_{CV} \eta \rho dV + \int_{CS} \eta \rho \vec{V} \cdot d\vec{A} \quad (33.2)$$

To apply Eq. (33.2) the system equation is formulated in terms of the rate of change of any extensive property  $N$  of the system; the corresponding intensive property is represented by  $\eta$ . The first integral in Eq. (33.2) represents the quantity of  $N$  stored within the control volume; the second integral represents the net flux of  $N$  carried outward through the control surface.

Conservation of mass is obtained by substituting, for a system of constant mass,  $dM/dt = 0$ ; the corresponding intensive property is "mass per unit mass," so  $\eta = 1$ . Thus,

$$0 = \frac{\partial}{\partial t} \int_{CV} \rho d\forall + \int_{CS} \rho \vec{V} \cdot d\vec{A} \quad (33.3)$$

For incompressible flow, density cannot vary with time, so it is tempting to factor  $\rho$  from under the volume integral. However, parts of the control volume could be occupied by fluids having different densities at different times.

The momentum equation for control volumes is obtained by substituting the system form of Newton's second law into the left side of Eq. (33.2) and setting  $\eta = \vec{V}$  on the right side:

$$\vec{F}_S + \vec{F}_B = \frac{\partial}{\partial t} \int_{CV} \vec{V} \rho d\forall + \int_{CS} \vec{V} \rho \vec{V} \cdot d\vec{A} \quad (33.4)$$

The left side of Eq. (33.4) represents external surface and body forces acting *on* the control volume. The first integral represents the rate of change of linear momentum contained within the control volume. The second integral accounts for the net flux of linear momentum from the control surface. Equation (33.4) is a vector equation; each of its three components must be satisfied.

The first law of thermodynamics is obtained by substituting the rate form of the system equation into Eq. (33.2). The result is the scalar equation

$$\dot{Q} - \dot{W}_{\text{shaft}} = \frac{\partial}{\partial t} \int_{CV} e \rho d\forall + \int_{CS} \left( e + \frac{p}{\rho} \right) \rho \vec{V} \cdot d\vec{A} \quad (33.5)$$

In Eq. (33.5) the intensive property stored energy  $e = u + (V^2/2) + gz$  includes internal thermal energy  $u$ , kinetic energy  $V^2/2$ , and gravitational potential energy,  $gz$  (all per unit mass). The rate of heat transfer  $\dot{Q}$  is positive when into the control volume; the rate of shaft work  $\dot{W}_{\text{shaft}}$  represents work extracted from the control volume. The first integral accounts for energy stored within the control volume; the second integral accounts for the flux of stored energy and flow work  $p/\rho$  done by pressure forces acting on the control surface.

### 33.4 Differential Analysis of Fluid Motion

Conservation of mass, Newton's second law of motion, and the first law of thermodynamics are independent physical principles that must be satisfied by any real flow. In principle it is possible to solve for three components of velocity and the pressure (four unknowns) using conservation of mass and the three components of the momentum equation. This usually is done using differential formulations to obtain detailed information about the flow field. The differential formulations may

be developed using a differential system or control volume. Figure 33.2 shows a differential CV with velocity vectors; to set up the analysis the velocity vectors are chosen in positive coordinate directions.

**Figure 33.2** Differential control volume showing velocity vectors.

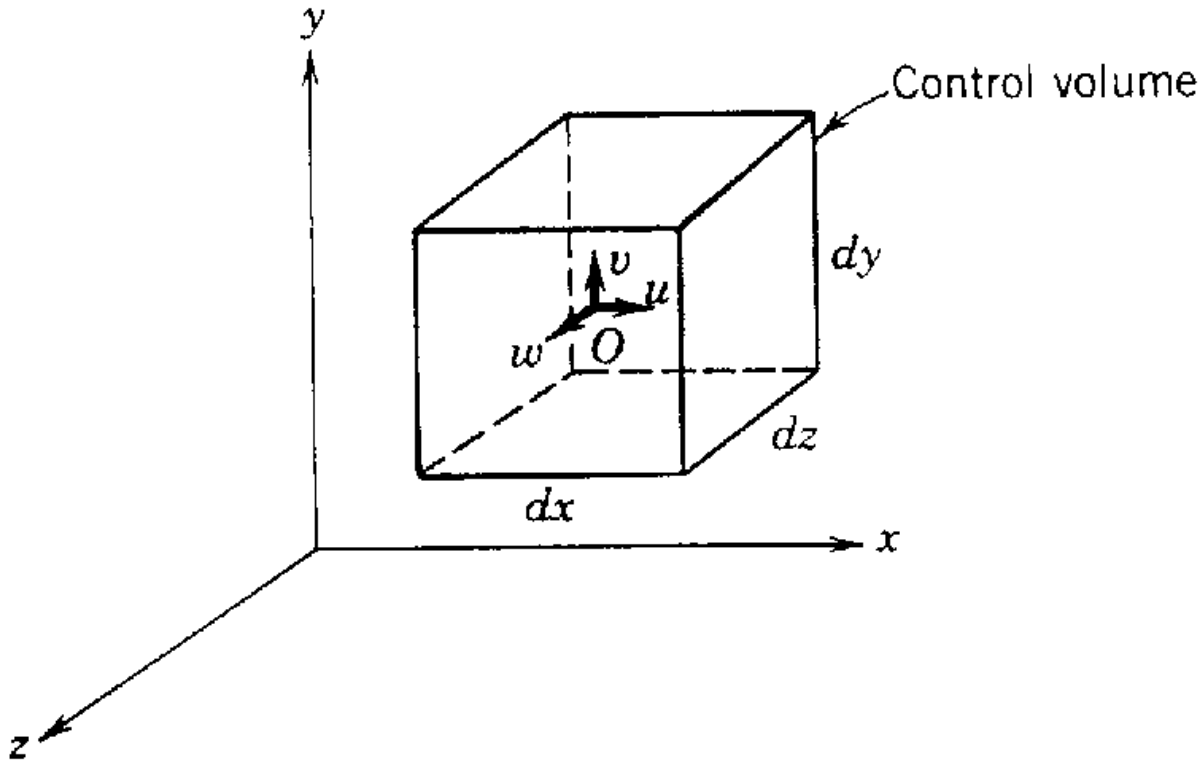


Figure 33.2 shows the first term in the Taylor series expansion of each velocity component in the  $x$  direction. Similar expansions for velocity components in the other coordinate directions are summed to obtain the total flux of mass from the control volume (no mass storage term is needed since the fluid is incompressible):

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0 \quad \text{or} \quad \nabla \cdot \vec{V} = 0 \quad (33.6)$$

Equation (33.6) expresses conservation of mass in differential form. The equation was derived using an infinitesimal control volume but is valid at any point in the flow. The velocity field for incompressible flow must satisfy Eq. (33.6).

Since the velocity varies from point to point, the acceleration of a fluid particle in a velocity field must be calculated using a special derivative called the *substantial derivative*. The acceleration of a fluid particle is given a special symbol  $D/Dt$  and written:

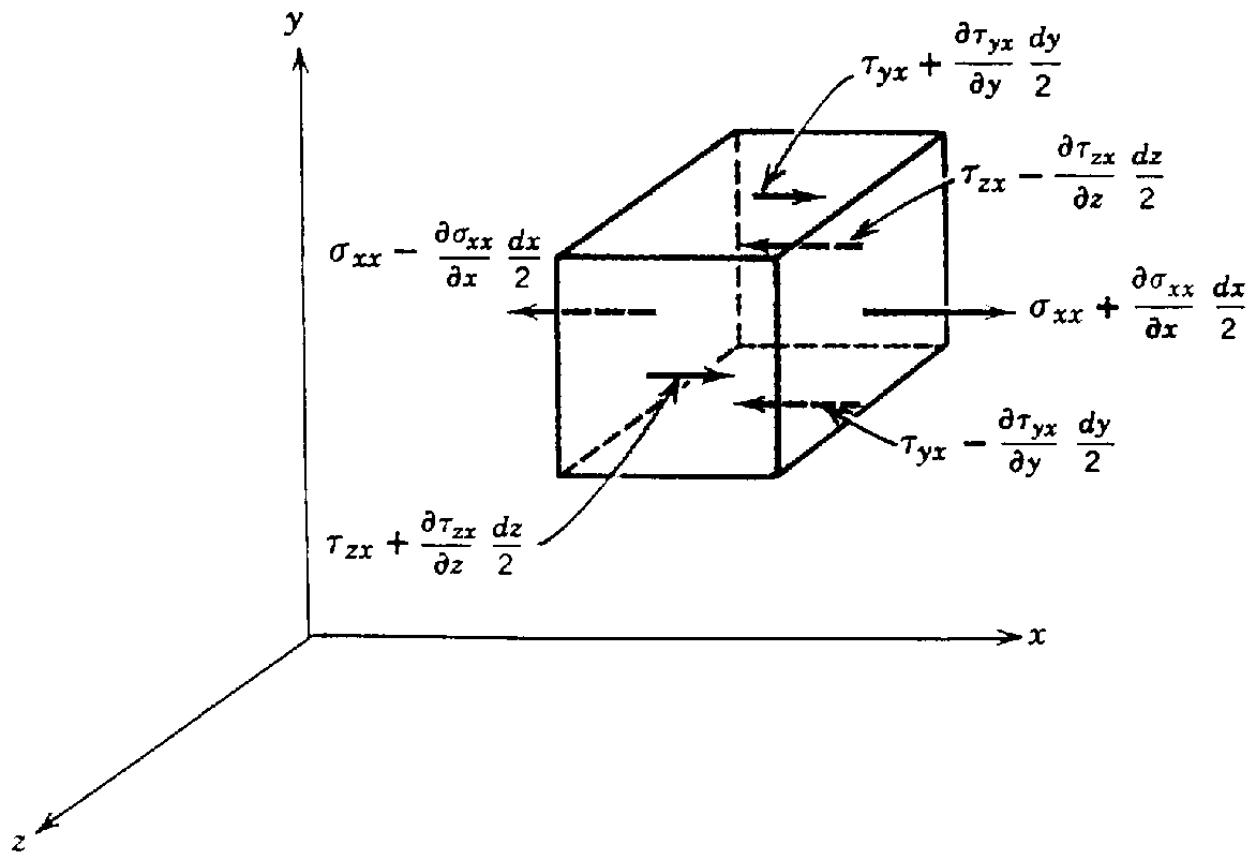


$$\frac{D\vec{V}}{Dt} = \underbrace{u \frac{\partial \vec{V}}{\partial x} + v \frac{\partial \vec{V}}{\partial y} + w \frac{\partial \vec{V}}{\partial z}}_{\substack{\uparrow \\ \text{Convective} \\ \text{acceleration}}} + \underbrace{\frac{\partial \vec{V}}{\partial t}}_{\substack{\uparrow \\ \text{Local} \\ \text{acceleration}}} \quad \text{or} \quad \frac{D\vec{V}}{Dt} = \vec{V} \cdot \nabla \vec{V} + \frac{\partial \vec{V}}{\partial t} \quad (33.7)$$

*Convective acceleration* occurs when fluid particles are convected into regions of differing velocity; it may be nonzero even in a steady flow, such as steady flow through a nozzle. *Local acceleration* is caused by velocity variations with time; it is nonzero only for unsteady flow.

Forces acting on a fluid particle also may be obtained using the Taylor series expansion procedure. The results of expanding the stresses acting in the  $x$  direction on an infinitesimal fluid particle are shown in Fig. 33.3.

**Figure 33.3** Differential fluid element showing stresses acting in the  $x$  direction.



Performing the same expansion in the other directions and collecting terms gives the *stress equations of motion*:

$$\rho g_x + \frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \tau_{yx}}{\partial y} + \frac{\partial \tau_{zx}}{\partial z} = \rho \left( u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + w \frac{\partial u}{\partial z} + \frac{\partial u}{\partial t} \right) \quad (33.8a)$$

$$\rho g_y + \frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \sigma_{yy}}{\partial y} + \frac{\partial \tau_{zy}}{\partial z} = \rho \left( u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + w \frac{\partial v}{\partial z} + \frac{\partial v}{\partial t} \right) \quad (33.8b)$$

$$\rho g_z + \frac{\partial \tau_{xz}}{\partial x} + \frac{\partial \tau_{yz}}{\partial y} + \frac{\partial \sigma_{zz}}{\partial z} = \rho \left( u \frac{\partial w}{\partial x} + v \frac{\partial w}{\partial y} + w \frac{\partial w}{\partial z} + \frac{\partial w}{\partial t} \right) \quad (33.8c)$$

Before Eqs. (33.8a) through (33.8c) may be used to solve for velocity, the stress field must be related to the velocity field. Details of this development are beyond the scope of this article, but are well covered by Sherman [1990]. For incompressible flow the stress components in rectangular coordinates are

$$\tau_{xy} = \tau_{yx} = \mu \left( \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \right) \quad (33.9a)$$

$$\tau_{yz} = \tau_{zy} = \mu \left( \frac{\partial w}{\partial y} + \frac{\partial v}{\partial z} \right) \quad (33.9b)$$

$$\tau_{zx} = \tau_{xz} = \mu \left( \frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \right) \quad (33.9c)$$

$$\sigma_{xx} = -p - \frac{2}{3} \mu \nabla \cdot \vec{V} + 2\mu \frac{\partial u}{\partial x} \quad (33.9d)$$

$$\sigma_{yy} = -p - \frac{2}{3} \mu \nabla \cdot \vec{V} + 2\mu \frac{\partial v}{\partial y} \quad (33.9e)$$

$$\sigma_{zz} = -p - \frac{2}{3} \mu \nabla \cdot \vec{V} + 2\mu \frac{\partial w}{\partial z} \quad (33.9f)$$

Note that shear stresses on adjacent faces of a fluid element are equal but directed oppositely.

The final form of the momentum equation is obtained by substituting stresses from Eq. (33.9) into Eq. (33.8). The result is the *Navier-Stokes equations*:

$$\rho g_x - \frac{\partial p}{\partial x} + \mu \left( \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) = \rho \left( u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + w \frac{\partial u}{\partial z} + \frac{\partial u}{\partial t} \right) \quad (33.10a)$$

$$\rho g_y - \frac{\partial p}{\partial y} + \mu \left( \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2} \right) = \rho \left( u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + w \frac{\partial v}{\partial z} + \frac{\partial v}{\partial t} \right) \quad (33.10b)$$

$$\rho g_z - \frac{\partial p}{\partial z} + \mu \left( \frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} + \frac{\partial^2 w}{\partial z^2} \right) = \rho \left( u \frac{\partial w}{\partial x} + v \frac{\partial w}{\partial y} + w \frac{\partial w}{\partial z} + \frac{\partial w}{\partial t} \right) \quad (33.10c)$$

These second-order, nonlinear partial differential equations are the fundamental equations of motion for viscous incompressible fluids. The Navier-Stokes equations are extremely difficult to solve analytically; only a handful of exact solutions are known [Sherman, 1990]. Some simplified cases can be solved numerically using today's advanced computers.

The Navier-Stokes equations also provide the starting point for stability analyses that predict the breakdown of laminar flow and the onset of turbulence.

## 33.5 Incompressible Inviscid Flow

All real fluids are viscous. However, in many situations it is reasonable to neglect viscous effects. Thus it is useful to consider an incompressible *ideal fluid* with zero viscosity. When the fluid is inviscid there are no shear stresses; pressure is the only stress on a fluid particle.

The equations of motion for frictionless flow are called the *Euler equations*. They are obtained by substituting the acceleration of a fluid particle into Eq. (33.1):

$$-\nabla p + \rho \vec{g} = \rho \frac{D\vec{V}}{Dt} \quad (33.11)$$

Equation (33.11) can be integrated to relate pressure, elevation, and velocity in a flowing fluid.

The Euler equations may be written in components using rectangular coordinates or using *streamline coordinates* defined along and normal to flow streamlines. In streamline coordinates the components of the Euler equations are

$$\frac{\partial p}{\partial s} = -\rho V \frac{\partial V}{\partial s} \quad (\text{along a streamline}) \quad (33.12a)$$

$$\frac{\partial p}{\partial n} = \rho \frac{V^2}{R} \quad (\text{normal to a streamline}) \quad (33.12b)$$

Equation (33.12a) shows that, for frictionless flow, variations in pressure and velocity are opposite; pressure falls when velocity increases and vice versa. (Frictionless flow is an excellent model for accelerating flow; it must be used with caution for decelerating flow, in which viscous effects are likely to be important.)

Equation (33.12b) shows that pressure always increases in the direction outward from the center of curvature of streamlines. (The increasing pressure causes each fluid particle to follow a curved path along the curved streamline.) When streamlines are straight, the radius of curvature is infinite and pressure does not vary normal to the streamlines.

The *Bernoulli equation* is obtained when the Euler equation is integrated along a streamline for steady, incompressible flow without viscous effects:

$$\frac{p_1}{\rho} + \frac{V_1^2}{2} + gz_1 = \frac{p_2}{\rho} + \frac{V_2^2}{2} + gz_2 \quad (33.13)$$

The Bernoulli equation is one of the most useful equations in fluid mechanics, but it also is incorrectly applied frequently. The restrictions of steady, incompressible flow without friction must be justified carefully each time the Bernoulli equation is used.

The Bernoulli equation may be used to predict pressure variations in external flow over objects and to design instrumentation for measuring pressure and velocity. *Stagnation pressure* is sensed by a total-head tube where  $V = 0$ . For this situation the Bernoulli equation reduces to

$$p_0 = p + \frac{1}{2}\rho V^2 \quad (33.14)$$

Equation (33.14) defines stagnation pressure  $p_0$  as the sum of *static pressure*,  $p$ , and *dynamic pressure*,  $\frac{1}{2}\rho V^2$ . A detailed discussion of fluid measurements is beyond the scope of this chapter, but these pressures can be measured using probes and a suitable instrument to sense pressure.

## 33.6 Dimensional Analysis and Similitude

---

Dimensional analysis is the process of combining key parameters of a flow situation into dimensionless groups. Several methods may be used to obtain the dimensionless groups [Fox and McDonald, 1992], which reduce the number of variables needed to express the functional dependence of the results of an experiment or analysis. Thus, dimensionless groups simplify the presentation of data and permit analytical results to be generalized.

Each dimensionless group is a ratio of forces. Significant dimensionless groups in fluid mechanics include Reynolds number  $Re$  (ratio of inertia to viscous forces), pressure coefficient  $C_p$  (ratio of pressure force to inertia force), Froude number  $Fr$  (ratio of gravity to inertia forces), Weber number  $We$  (ratio of surface tension to inertia forces), and Mach number  $M$  (which may be interpreted as the ratio of inertia to compressibility forces).

*Dynamic similarity* occurs when ratios of all significant forces are the same between two flows. Dynamic similarity is required to scale model test results for use in prediction or design. Dynamic similarity is ensured for geometrically similar flows with corresponding flow patterns when all relevant dimensionless groups but one are duplicated between the two flows.

The basic differential equations also may be nondimensionalized to obtain dimensionless groups. Dynamic similarity is ensured when two flows are governed by the same differential equations with the same dimensionless coefficient values in the equations and boundary conditions. Strouhal number  $St$  is a frequency parameter that arises from boundary conditions for external flow with

vortex shedding.

### 33.7 Internal Incompressible Viscous Flow

---

*Laminar flow* occurs at low Reynolds number; as Reynolds number increases, transition occurs and flow becomes turbulent. The numerical value corresponding to "low" Reynolds number depends on flow geometry. For circular pipes the Reynolds number at transition is  $Re = \rho \bar{V} D / \mu \approx 2000$ , where  $\bar{V}$  is the average velocity at any cross section. Transition Reynolds numbers for other geometries differ significantly.

Fully developed laminar flow cases in simple geometries can be solved by (1) using a differential control volume and Taylor series expansion to obtain an equation for shear stress variation, or (2) reducing the Navier-Stokes equations to a simple form applicable to the flow field. Then the shear stress profile is integrated using appropriate boundary conditions to obtain the velocity profile. Once the velocity profile is obtained, volume flow rate, flow rate as a function of pressure drop, average velocity, and point(s) of maximum velocity can be found. Analyses of fully developed laminar flow cases are presented by Fox and McDonald [1992]; all known exact solutions of the Navier-Stokes equations are described in detail by Schlichting [1979] and White [1991].

*Turbulent flow* cannot be analyzed from first principles. Turbulence is characterized by velocity fluctuations that transport momentum across streamlines; there is no simple relationship between shear stress and strain rate in turbulent flow. Instantaneous properties cannot be predicted in a turbulent flow field; only average values can be calculated. For engineering analyses, turbulent flow is handled empirically using curve-fits to velocity profiles and experimentally determined loss coefficients.

Analysis of turbulent pipe flow is based on the first law of thermodynamics. Viscous friction causes irreversible conversion from mechanical energy to thermal energy. This conversion is regarded as a loss in mechanical energy called *head loss*:

$$\frac{p_1}{\rho} + \alpha_1 \frac{\bar{V}_1^2}{2} + gz_1 - \left( \frac{p_2}{\rho} + \alpha_2 \frac{\bar{V}_2^2}{2} + gz_2 \right) = h_{lT} \quad (33.15)$$

In Equation (33.15) total head loss  $h_{lT}$  is the difference between the mechanical energies at cross sections 1 and 2;  $\alpha \bar{V}^2 / 2$  is the kinetic energy flux ( $\alpha$  is the *kinetic energy coefficient*).

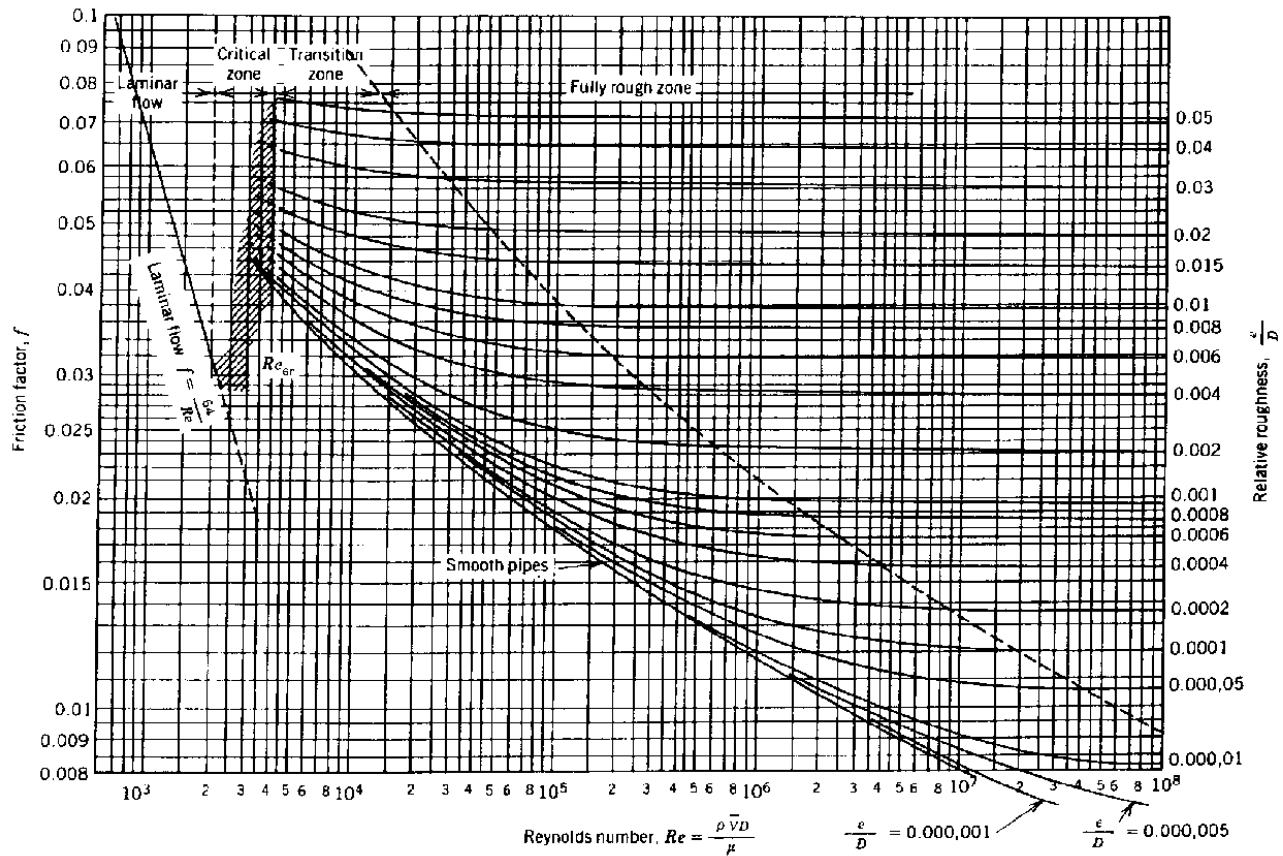
To make calculations, total head loss is subdivided into "major" losses that occur in sections of constant area where flow is fully developed and "minor" losses in transitions such as entrances, fittings, valves, and exits. Major losses,  $h_l$ , in sections with fully developed flow are expressed in terms of the experimentally determined friction factor  $f$ :

$$h_l = f \frac{L}{D} \frac{\bar{V}^2}{2} \quad (33.16)$$

Friction factor is a function of Reynolds number  $Re$  and relative roughness  $e/D$  (equivalent

roughness height  $e$  divided by tube diameter  $D$ ). Results from numerous experiments were compiled and smooth curves fitted by Moody; the results are shown on the *Moody diagram* (Fig. 33.4).

**Figure 33.4** Moody diagram giving friction factors for pipe flow. (Source: Moody, L. F. 1944. Friction factors for pipe flow. *Trans. ASME*. 66(8):671–684. With permission.)



Minor loss data also are measured experimentally; minor losses  $h_{lm}$  may be expressed as:

$$h_{lm} = f \frac{L_e}{D} \frac{\bar{V}^2}{2} = K \frac{\bar{V}^2}{2} \quad (33.17)$$

Equivalent length ratios  $L_e/D$  and minor loss coefficients  $K$  are available from numerous sources. More details on computation of minor losses are in **Chapter 38, "Valves."**

Computer programs that make calculations for pipe flow systems are commonly available. One such program accompanies the Fox and McDonald text [1992].

## 33.8 External Incompressible Viscous Flow

---

The *boundary layer* is the thin region near the surface of a body in which viscous effects are important. Boundary layers may be laminar or turbulent, depending on Reynolds number and factors such as pressure gradient, surface roughness, and heat transfer.

Basic characteristics of all laminar and turbulent boundary layers are present in developing flow over a flat plate in a semi-infinite fluid. The boundary layer is thin, so there is negligible disturbance of the inviscid flow outside the boundary layer; thus, the pressure gradient is close to zero for this flow field. *Transition* from laminar to turbulent boundary-layer flow on a flat plate occurs for Reynolds numbers above  $Re = \rho Ux/\mu \approx 5 \cdot 10^5$ ; this usually is considered the transition Reynolds number for flat-plate flow. Transition may occur earlier if the surface is rough, if pressure rises in the flow direction, or if separation occurs. Following transition the turbulent boundary layer thickens more rapidly than the laminar boundary layer as a result of the increased shear stress on the surface beneath the turbulent boundary layer.

Bodies immersed in flowing fluids experience forces due to the shear stresses and pressure differences caused by the fluid motion. *Drag* is the force parallel to the flow direction and *lift* is the force perpendicular to the flow direction. *Streamlining* is the art of shaping a body to reduce the fluid dynamic drag force. Airfoils (and hydrofoils) are designed to produce lift in air (or water); they are streamlined to reduce drag and attain high lift/drag ratios.

In general, lift and drag cannot be predicted analytically, although progress continues on computational fluid dynamics (CFD) computer programs. For most engineering purposes, drag and lift are calculated from experimentally derived coefficients. The defining equations for drag and lift coefficients  $C_D$  and  $C_L$  are

$$F_D = C_D A \frac{1}{2} \rho V^2 \quad \text{and} \quad F_L = C_L A \frac{1}{2} \rho V^2 \quad (33.18)$$

where  $\frac{1}{2} \rho V^2$  is the dynamic pressure and  $A$  is the area upon which each coefficient is based. Common practice is to base drag coefficients on projected frontal area and lift coefficients on projected *planform* area. See Fox and McDonald [1992] for more details.

### Defining Terms

**Boundary layer:** The thin layer of fluid adjacent to a surface where viscous effects are important; outside the boundary layer viscous effects may be neglected.

**Head loss:** The irreversible conversion from mechanical to thermal energy resulting from viscous friction in pipe flow (expressed as energy per unit mass).

**Newtonian fluid:** A fluid characterized by a linear relationship between shear rate (rate of angular deformation) and shear stress.

**Separation:** Phenomenon that occurs when fluid layers adjacent to a solid surface are brought to rest and the boundary-layer flow departs from the surface contour, forming a relatively low-pressure *wake* region. Separation can occur only in an *adverse pressure gradient*, in which pressure increases in the flow direction.

**Viscosity:** The coefficient that relates rate of shearing strain to shear stress for a Newtonian fluid (also called *dynamic viscosity*).

## References

- Fay, J. A. 1994. *Introduction to Fluid Mechanics*. MIT Press, Cambridge, MA.
- Fox, R. W. and McDonald, A. T. 1992. *Introduction to Fluid Mechanics*, 4th ed. John Wiley & Sons, New York.
- Moody, L. F. 1944. Friction factors for pipe flow. *Trans. ASME*. 66(8):671–684.
- Schlichting, H. 1979. *Boundary-Layer Theory*, 7th ed. McGraw-Hill, New York.
- Sherman, F. S. 1990. *Viscous Flow*. McGraw-Hill, New York.
- White, F. M. 1991. *Viscous Fluid Flow*. McGraw-Hill, New York.

## Further Information

A comprehensive source of basic information is the *Handbook of Fluid Dynamics*, edited by Victor L. Streeter (McGraw-Hill, New York, 1960).

Timely reviews of important topics are published in the *Annual Review of Fluid Mechanics* series (Annual Reviews, Palo Alto, CA). Each volume contains a cumulative index.

The *Journal of Fluids Engineering*, published quarterly (American Society of Mechanical Engineers, New York), contains articles with content ranging from fundamentals of fluid mechanics to fluid machinery.

The monthly *AIAA Journal* and semimonthly *Journal of Aircraft* (American Institute for Aeronautics and Astronautics, New York) treat aerospace applications of fluid mechanics.

Transportation aspects of fluid mechanics are covered in SAE publications (Society of Automotive Engineers, Warrendale, PA).



Hoffman, J. D. “Compressible Fluids”  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

- 34.1 General Features
- 34.2 Basic Equations
- 34.3 Steady Quasi–One-Dimensional Flow
- 34.4 Equations of State and Thermodynamics
- 34.5 Stagnation Properties
- 34.6 Isentropic Flow
- 34.7 Nozzles
- 34.8 Shock Waves
- 34.9 Friction and Heat Transfer

**Joe D. Hoffman**

*Purdue University*

Throughout history, up until the early 20th century, engineering applications involving fluids were limited to the flow of liquids or the low-speed flow of gases, where the density of the fluid remained constant. In such flows, the famous Bernoulli equation governed the interchange of mechanical energy, kinetic energy, and potential energy:

$$\frac{P}{\rho} + \frac{1}{2}V^2 + gz = 0 \quad (34.1)$$

where  $P$  is fluid pressure,  $\rho$  is fluid density,  $V$  is velocity,  $g$  is acceleration of gravity, and  $z$  is elevation above some reference location.

Beginning in the early 20th century with the development of steam turbines, high-speed flows in which the density changes appreciably, sometimes by several orders of magnitude, have become the rule rather than the exception. The changes in density must be accounted for in such flows. The science concerned with the phenomena arising from the flow of compressible fluids is generally called *gas dynamics*.

The objectives of this section are to discuss the differences between the behavior of incompressible fluids and compressible fluids, to present the general features of the flow of compressible fluids, to present the equations governing steady quasi–one-dimensional flow, and to discuss isentropic flow, nozzles, shock waves, friction, and heat transfer.

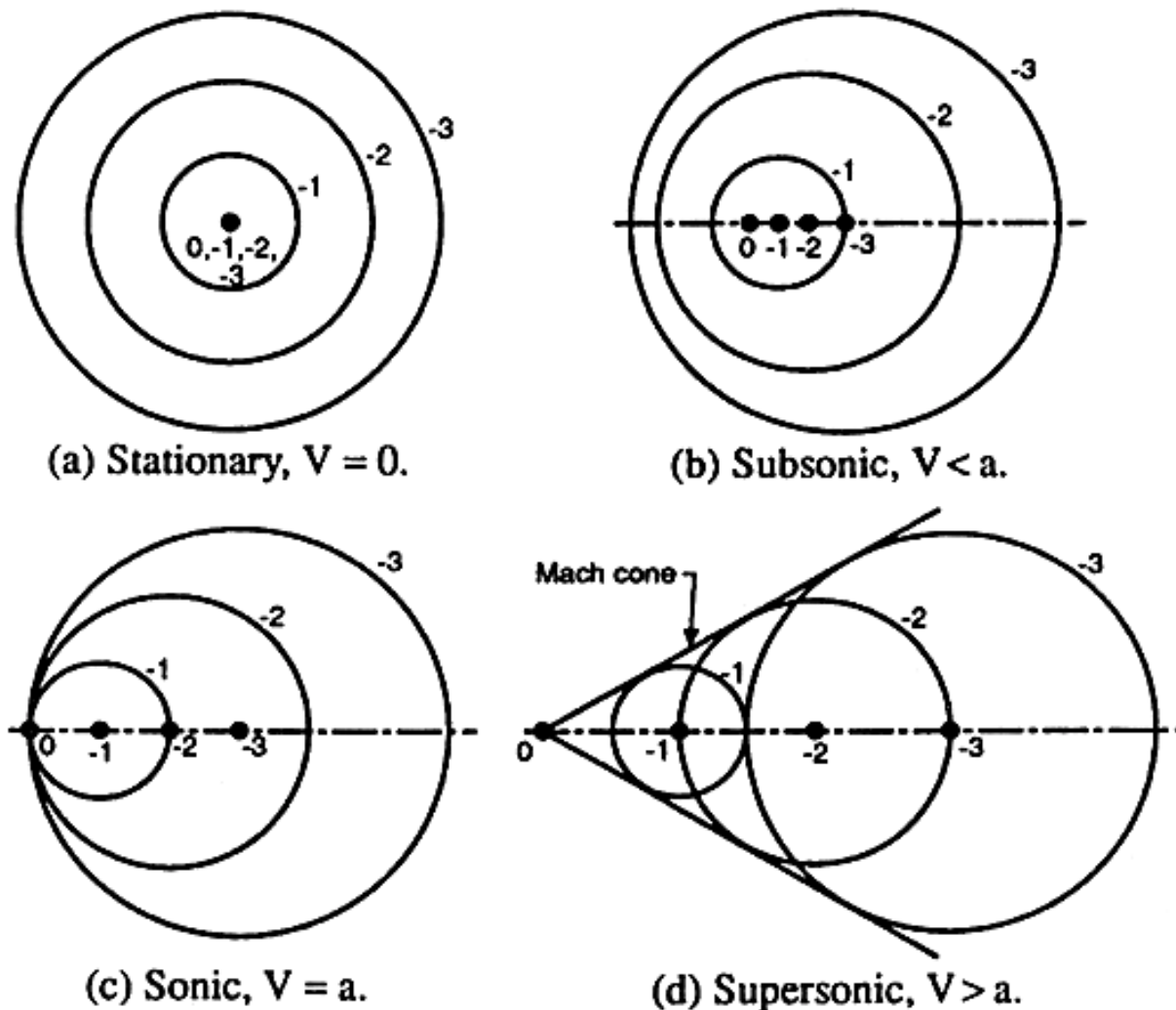
## 34.1 General Features

The distinguishing feature of a compressible fluid in comparison to an incompressible fluid is that the density changes significantly when any of the other fluid properties change. This effect gives rise to several unique features, which are discussed in this section.

The speed of propagation of small disturbances, which is called the **speed of sound** and denoted by the symbol  $a$ , is infinite in an ideal incompressible fluid. Although no real fluid is truly incompressible, liquids are nearly so. At standard temperature and pressure,  $a \cong 5000$  ft/s for many common liquids, whereas  $a$  is in the vicinity of 1000 ft/s for most common gases.

There are major differences in flow patterns for compressible fluids, depending on whether the fluid flow velocity,  $V$ , is less than, equal to, or greater than the speed of sound,  $a$ . These differences can be illustrated by considering the motion of a point disturbance with the velocity  $V$  moving through a compressible fluid with the speed of sound  $a$ . Figure 34.1 illustrates the acoustic wave pattern at time 0 due to small disturbances created by a moving point source at  $-1$ ,  $-2$ , and  $-3$  time increments. When  $V = 0$ , the disturbances clear away from the point source uniformly. When  $0 < V < a$ , the disturbances clear away in all directions, but unsymmetrically. When  $V = a$ , a plane wave front is attached to the point source. When  $V > a$ , all the disturbances are contained within a cone, called the *Mach cone*, which has the point source at its apex. The disturbance is not felt outside of the Mach cone.

**Figure 34.1** Regimes of compressible flow.

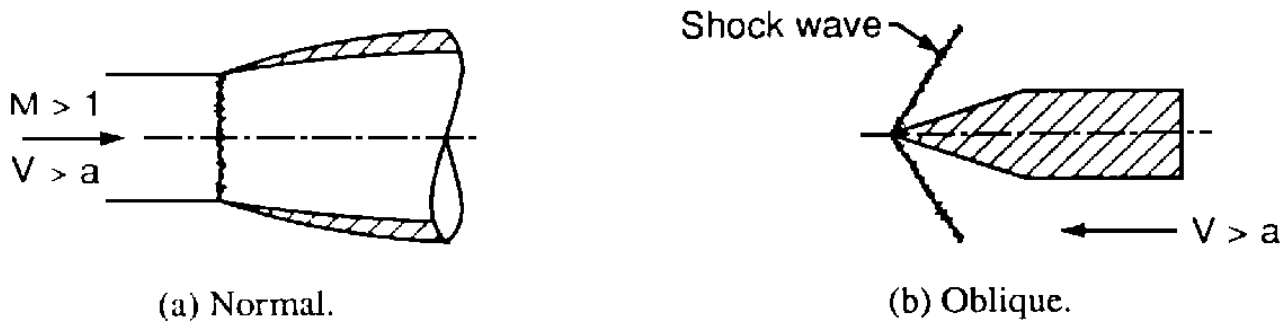


The ratio of the velocity,  $V$ , to the speed of sound,  $a$ , is called the *Mach number* in honor of the Austrian physicist Ernst Mach. Thus,

$$M = \frac{V}{a} \quad (34.2)$$

A major phenomenon peculiar to the flow of compressible fluids is the appearance of shock waves, which are waves of infinitesimal thickness across which finite changes in flow properties occur. Shock waves appear in supersonic flows where  $V > a$ . Figure 34.2(a) illustrates a normal shock wave standing at the inlet of a supersonic diffuser, and Fig. 34.2(b) illustrates an oblique shock wave attached to the nose of a supersonic projectile.

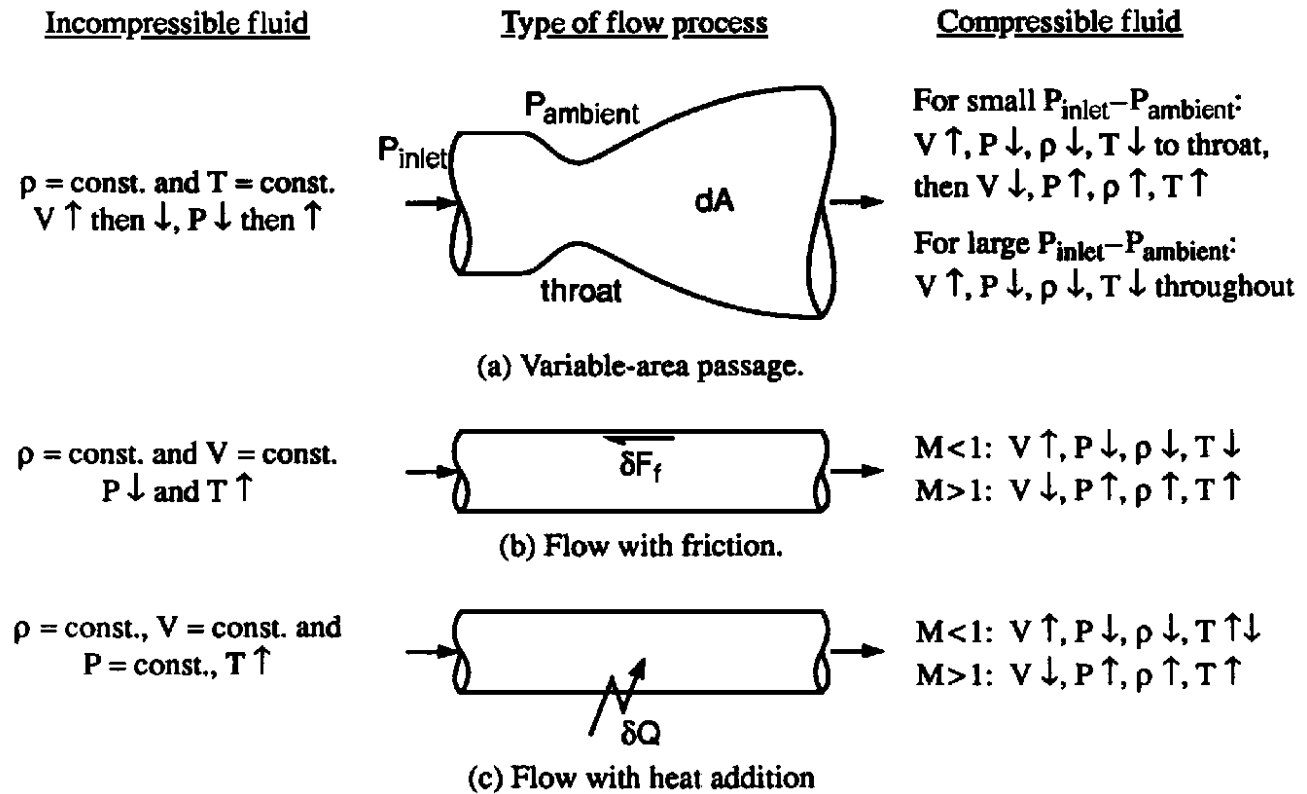
**Figure 34.2** Shock waves.



Another phenomenon that is peculiar to compressible flows is **choking**. Choking occurs when the velocity in a flow passage reaches the speed of sound. At that condition, further changes in the downstream conditions cannot be transmitted upstream since  $V > a$ . Consequently, the upstream conditions, including the mass flow rate, are fixed. Choking cannot occur in incompressible fluids, although the phenomenon of cavitation exerts a similar effect.

As a final illustration of the effects of compressibility, the contrasting behavior of incompressible and compressible fluids in several common processes is illustrated in Fig. 34.3. When an incompressible fluid flows through a variable-area passage where  $P_{\text{inlet}} > P_{\text{ambient}}$ , as illustrated in Fig. 34.3(a),  $V$  increases and  $P$  decreases (i.e., nozzle action) to the minimum area, called the *throat*, after which  $V$  decreases and  $P$  increases (i.e., diffuser action). When a compressible fluid flows through such a passage, the fluid behavior is more complicated. For small  $P_{\text{inlet}} - P_{\text{ambient}}$ , the fluid behavior is similar to an incompressible flow, except that  $\rho$  decreases to the throat then increases. However, for large values of  $P_{\text{inlet}} - P_{\text{ambient}}$ , the flow chokes at the throat where  $M = 1$ , and the fluid continues to accelerate in the diverging section. For intermediate values of  $P_{\text{inlet}} - P_{\text{ambient}}$ , the flow chokes at the throat and continues to accelerate until a shock wave occurs. The flow decelerates after the shock wave.

**Figure 34.3** Flow processes



When an incompressible fluid flows adiabatically (i.e.,  $\delta Q = 0$ ) through a constant-area duct with wall friction, as illustrated in Fig. 34.3(b),  $\rho$  and  $V$  remain constant,  $P$  decreases, and  $T$  increases. The behavior of a compressible fluid in a constant-area duct with friction depends upon whether the incoming flow is subsonic or supersonic. For subsonic inflow,  $V$  increases, and  $P$ ,  $\rho$ , and  $T$  decrease. For supersonic inflow,  $V$  decreases, and  $P$ ,  $\rho$ , and  $T$  increase. If the ambient pressure is low enough, both cases will choke at  $M = 1$ .

When an incompressible fluid flows without friction through a constant-area duct with heat addition, as illustrated in Fig. 34.3(c),  $\rho$ ,  $V$ , and  $P$  remain constant, and  $T$  increases. The heat addition has no effect on the fluid flow. The behavior of a compressible fluid in a constant-area duct with heat addition depends upon whether the incoming flow is subsonic or supersonic. For subsonic inflow,  $V$  increases,  $P$  and  $\rho$  decrease, and  $T$  increases up to Mach numbers in the vicinity of 0.85, then decreases. For supersonic inflow,  $V$  decreases, and  $P$ ,  $\rho$ , and  $T$  increase. Opposite effects occur for heat removal.

## 34.2 Basic Equations

The basic equations of fluid dynamics, for all fluids and all flow processes, are (1) the law of conservation of mass, (2) Newton's second law of motion, and (3) the first law of thermodynamics. When applied to a flowing fluid, these basic laws of physics are called the *continuity equation*, the *momentum equation*, and the *energy equation*, respectively. These basic laws govern the behavior of all fluid flows: unsteady or steady; one-, two-, or three-dimensional; ideal fluid (inviscid) or real

fluid (viscous); laminar or turbulent; inert or chemically reacting; thermal equilibrium or nonequilibrium, and so on.

The integral forms of the basic equations for control volumes are presented in **Chapter 33**, Eqs. (33.3) to (33.5). These forms of the basic equations are valid for both incompressible and compressible fluids. The differential forms of the basic equations for compressible fluids differ somewhat from those for incompressible fluids. Those equations are not presented here due to their length. They are given by Zucrow and Hoffman [1976] and Fox and McDonald [1992]. For a Newtonian fluid (i.e., one for which stress is directly proportional to rate of strain), the resulting equations are the Navier-Stokes equations. For an ideal fluid (i.e., one in which shear stresses are assumed to be zero, an inviscid fluid), the resulting equations are the Euler equations. Both of these sets of equations comprise a system of coupled, highly nonlinear, partial differential equations that are quite difficult to solve. Their solution is generally obtained by numerical methods using large computers. The art and science concerned with solving these equations is called *computational fluid dynamics* (CFD) [Anderson *et al.*, 1984].

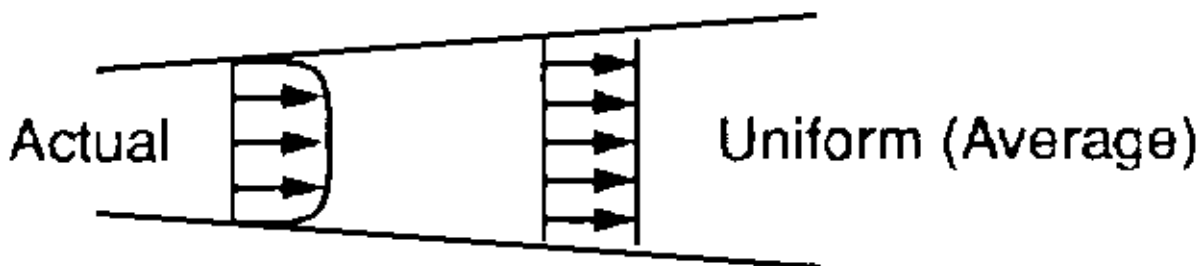
In addition to these basic laws, constitutive relationships are required to relate the response of the fluid to loads. Thermal and caloric equations of state, stress-strain relationships, and the laws of heat conduction and radiation are examples of constitutive relationships.

### 34.3 Steady Quasi–One-Dimensional Flow

---

Many features of the steady flow of a compressible fluid can be illustrated by considering the quasi–one-dimensional flow model. Consider flow through a passage or a streamtube of a larger flowfield. The quasi–one-dimensional flow model for such a flow is illustrated in Fig. 34.4. Actual flow properties vary over the inflow and outflow areas. In the quasi–one-dimensional flow model, all fluid properties are assumed to be uniform at their average values over each cross section. This is an approximate model. The integral equations for a control volume can be applied to the approximate model.

**Figure 34.4** Uniform flow concept.



The basic equations for steady quasi–one-dimensional flow in the absence of friction are:

$$\dot{m} = \rho AV = \text{constant} \quad (34.3)$$

$$dP + \rho V dV = 0 \quad (34.4)$$

$$dh + d(V^2/2) = \delta Q + \delta W \quad (34.5)$$

where  $\dot{m}$  is mass flow rate,  $A$  is cross-sectional flow area,  $h$  is enthalpy,  $Q$  is heat transfer to the fluid, and  $W$  is work done on the fluid. Equations (34.3) to (34.5) are the continuity, momentum, and energy equations, respectively.

## 34.4 Equations of State and Thermodynamics

---

The basic equations must be supplemented by equations of state (i.e., constitutive relationships). As a minimum, thermal and caloric equations of state are required:

$$T = T(P, \rho) \quad \text{and} \quad h = h(P, \rho) \quad (34.6)$$

For real gases and vapors these relationships can be in the form of equations, tables, graphs, or computer programs. For many gases the ideal gas law is a reasonable representation of the thermal equation of state:

$$P = \rho RT \quad (34.7)$$

For an ideal gas,  $h = \int C_p(T) dT$ , where  $C_p(T)$  is the constant-pressure specific heat. For a calorically perfect gas,  $C_p = \text{constant}$  and the caloric equation of state is:

$$h = C_p T \quad (34.8)$$

The second law of thermodynamics states that

$$ds \geq \frac{\delta Q}{T} \quad (34.9)$$

where  $s$  is entropy. For an adiabatic process (i.e.,  $\delta Q = 0$ ),  $ds \geq 0$ . For an adiabatic frictionless process,  $ds = 0$  and  $s = \text{constant}$ . Such a process is called an *isentropic process*. The isentropic process is the ideal reference process for many real processes.

The entropy equation (which is obtained from the second law of thermodynamics) for a thermally and calorically perfect gas is

$$\Delta s = C_p \ln \frac{T_2}{T_1} - R \ln \frac{P_2}{P_1} \quad (34.10)$$

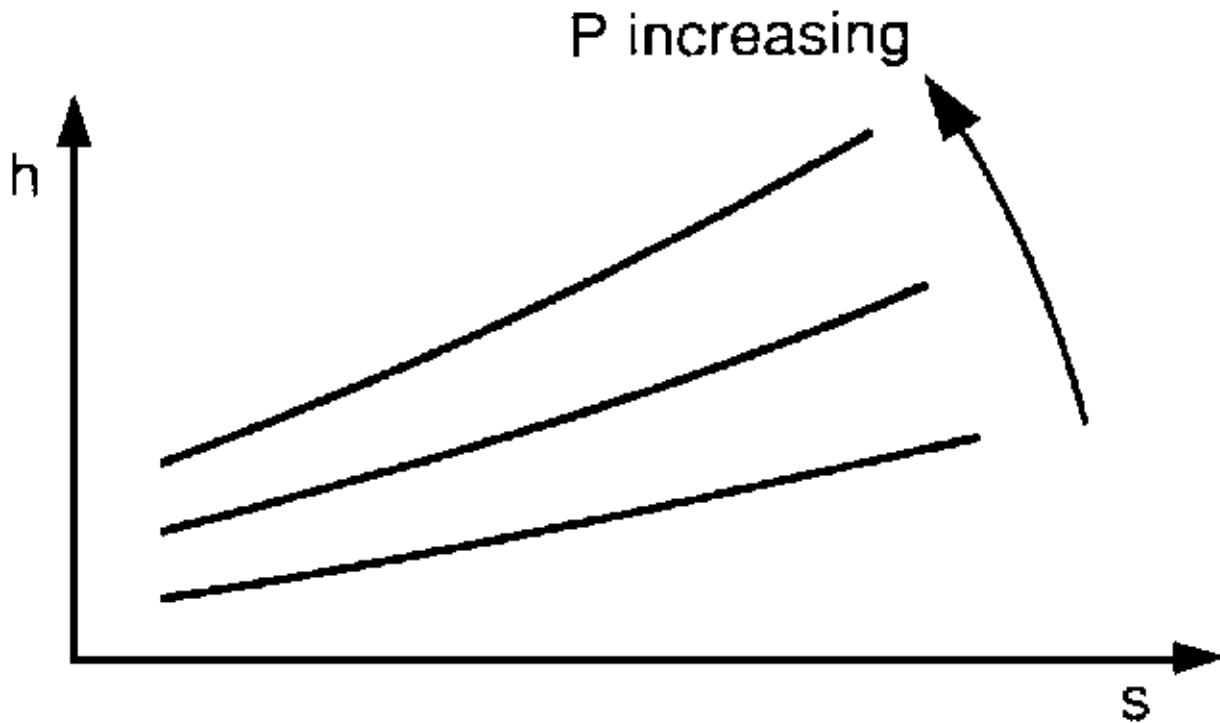
where the subscripts 1 and 2 denote initial and final states in a process. For an isentropic process,  $\Delta s = 0$  and

$$\frac{P_2}{P_1} = \left( \frac{T_2}{T_1} \right)^{C_p/R} = \left( \frac{T_2}{T_1} \right)^{\gamma/(\gamma-1)} \quad (34.11)$$

where  $C_p = \gamma R/(\gamma - 1)$  and  $\gamma$  is the ratio of specific heats.

Many compressible flow processes can be illustrated on the Mollier diagram, which is an  $h$ - $s$  state diagram. Isobars on the Mollier diagram are illustrated in Fig. 34.5. The isentropic process is a vertical line on the Mollier diagram. Many ideal processes are isentropic.

**Figure 34.5** The Mollier diagram.



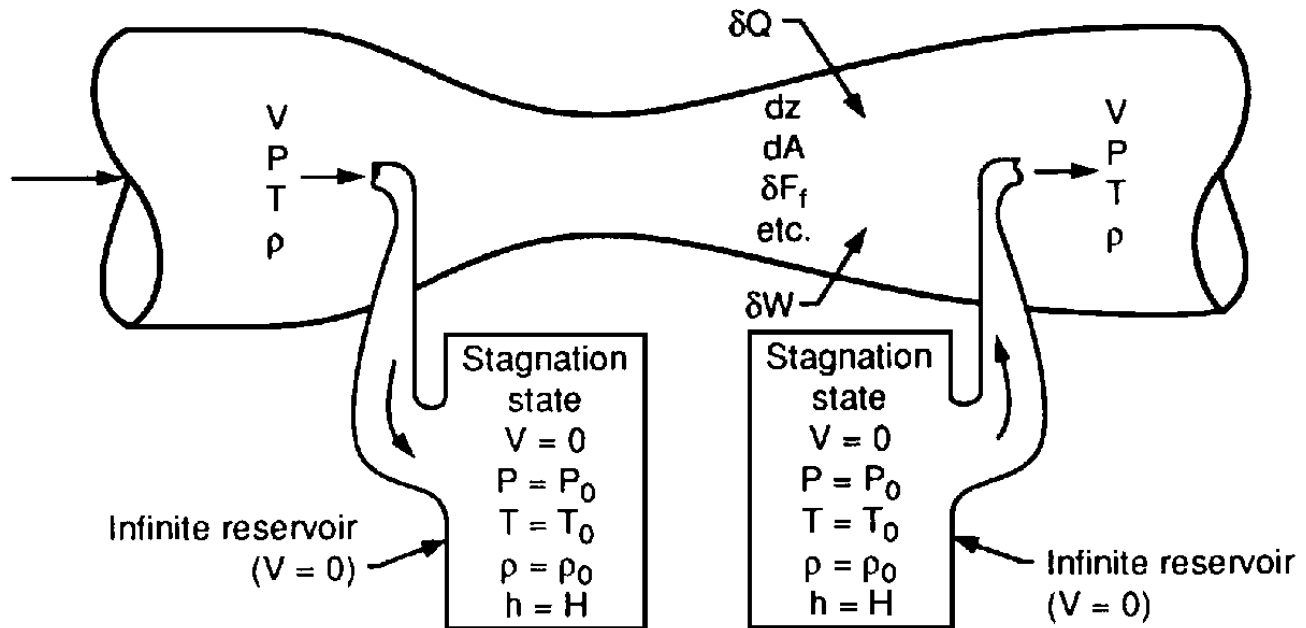
## 34.5 Stagnation Properties

Stagnation properties are the properties of a flow when brought to rest isentropically. They can be defined at every point in any flow, regardless of the type of processes occurring in the flow itself. The stagnation process is illustrated in Fig. 34.6. The properties at the stagnation state are denoted with the subscript 0. The stagnation pressure  $P_0$  is a measure of the mechanical energy in a flow, and the stagnation temperature  $T_0$  is a measure of the thermal energy in a flow. Stagnation properties are illustrated on the Mollier diagram in Fig. 34.7. Changes in stagnation properties between two points in a flow are a direct result of the flow processes occurring between those two

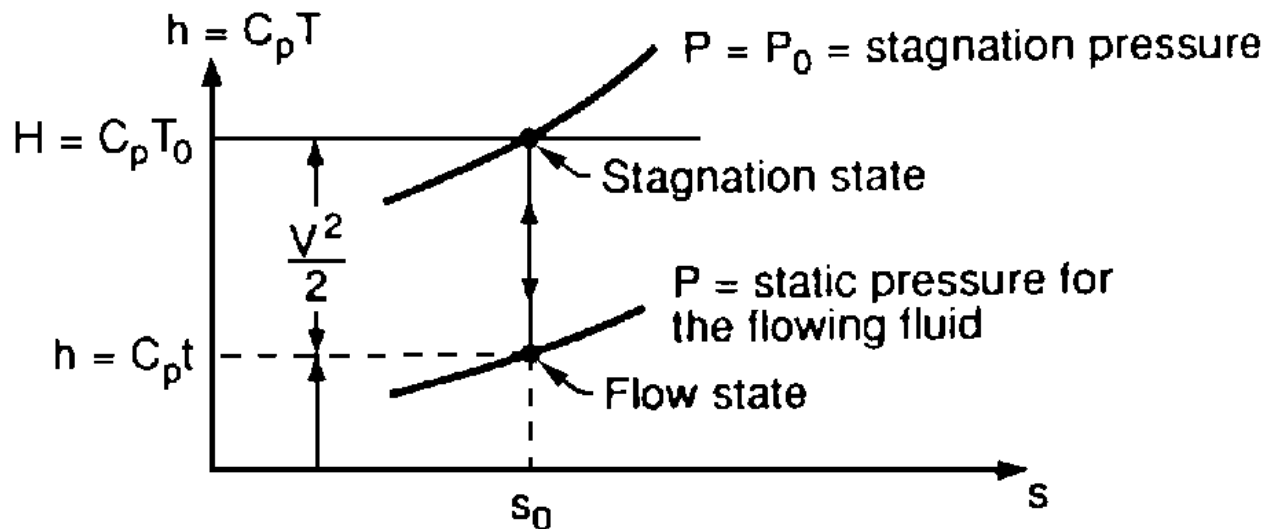


points.

**Figure 34.6** The stagnation process.



**Figure 34.7** Stagnation properties.



## 34.6 Isentropic Flow

An isentropic flow is an ideal flow that is frictionless and adiabatic ( $\delta Q = 0$ ); thus  $ds = 0$  and

$s = \text{constant}$ . On the Mollier diagram the isentropic process is a vertical line connecting the stagnation state with the flow state, as illustrated in Fig. 34.7. The ideal velocity achieved in an isentropic expansion of a perfect gas from a large reservoir where  $V \approx 0$ ,  $P = P_0$ , and  $T = T_0$  can be obtained from Eqs. (34.5) and (34.11) with  $\delta Q = \delta W = 0$ ,  $h = C_p T$ , and  $H = C_p T_0$ . Thus,

$$\begin{aligned} V &= \sqrt{2(H - h)} = \sqrt{2C_p(T_0 - T)} \\ &= \sqrt{2C_p T_0 (1 - T/T_0)} = \sqrt{\frac{2\gamma R T_0}{\gamma - 1} \left(1 - (P/P_0)^{(\gamma-1)/\gamma}\right)} \end{aligned} \quad (34.12)$$

The theoretical maximum speed that can be achieved in an isentropic expansion to  $P = 0$  is

$$V_{\max} = \sqrt{\frac{2\gamma R T_0}{\gamma - 1}} \quad (34.13)$$

From Eq. (34.12), the stagnation temperature  $T_0$  corresponding to a flow state  $(P, T, V)$  is

$$\begin{aligned} T_0 &= T \left(1 + \frac{V^2}{2C_p T}\right) = T \left(1 + \frac{\gamma - 1}{2} \frac{V^2}{\gamma R T}\right) \\ &= T \left(1 + \frac{\gamma - 1}{2} \frac{V^2}{a^2}\right) = T \left(1 + \frac{\gamma - 1}{2} M^2\right) \end{aligned} \quad (34.14)$$

From Eq. (34.11),

$$P_0 = P \left(1 + \frac{\gamma - 1}{2} M^2\right)^{\gamma/(\gamma-1)} \quad (34.15)$$

Equations (34.14) and (34.15) are two of the most useful relationships in gas dynamics. They define the stagnation pressure and temperature at any point in any flow in terms of static pressure, static temperature, and Mach number at the point. Combining Eqs. (34.7), (34.14), and (34.15) yields

$$\rho_0 = \rho \left(1 + \frac{\gamma - 1}{2} M^2\right)^{1/(\gamma-1)} \quad (34.16)$$

If the pressure difference across a passage is large enough to choke a flow, the Mach number will be 1.0 at the choking location, which is the exit if the passage converges and is the minimum area if the passage converges and then diverges. The flow properties at the choking location are the

critical properties, denoted by the superscript \*. From Eqs. (34.14) to (34.16),

$$T^* = \frac{2}{\gamma + 1} T_0, \quad P^* = \left( \frac{2}{\gamma + 1} \right)^{\gamma/(\gamma-1)} P_0, \quad (34.17)$$

and

$$\rho^* = \left( \frac{2}{\gamma + 1} \right)^{1/(\gamma-1)} \rho_0$$

The area at the choking location is the critical area,  $A^*$ . The relationship between  $A$ ,  $A^*$ , and the Mach number  $M$  is:

$$\frac{A}{A^*} = \frac{1}{M} \left[ \frac{2}{\gamma + 1} \left( 1 + \frac{\gamma - 1}{2} M^2 \right) \right]^{(\gamma+1)/2(\gamma-1)} \quad (34.18)$$

The mass flow rate in a passage can be evaluated in several ways.

$$\dot{m} = \rho AV = \frac{PAV}{RT} = PAM \frac{\gamma}{\sqrt{\gamma RT}} \quad (34.19)$$

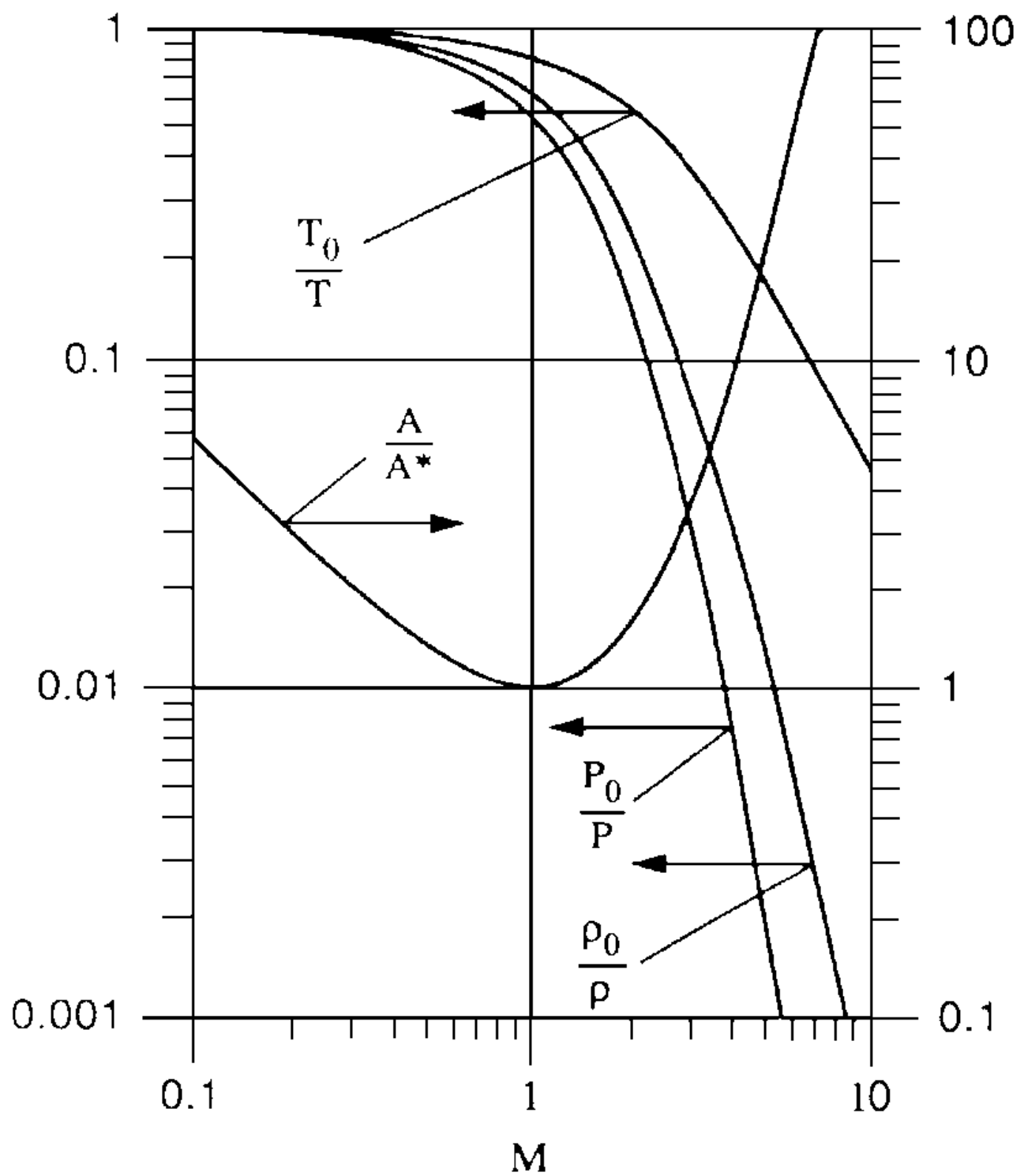
$$\begin{aligned} \dot{m} &= PAM \frac{\gamma}{\sqrt{\gamma RT_0}} \left( 1 + \frac{\gamma - 1}{2} M^2 \right)^{1/2} \\ &= \frac{\gamma P_0 AM}{\sqrt{\gamma RT_0}} \left( 1 + \frac{\gamma - 1}{2} M^2 \right)^{-(\gamma+1)/2(\gamma-1)} \end{aligned} \quad (34.20)$$

Equations (34.19) and (34.20) are valid at every point in any flow. When the flow is choked, the maximum flow rate,  $\dot{m}^*$ , is achieved:

$$\dot{m}^* = \gamma \left( \frac{2}{\gamma + 1} \right)^{(\gamma+1)/2(\gamma-1)} \frac{P_0 A^*}{\sqrt{\gamma RT_0}} \quad (34.21)$$

Isentropic flow in a variable-area passage is the ideal model for many real devices, such as nozzles, diffusers, turbines, and compressors. The variations of velocity  $V$ , pressure  $P$ , density  $\rho$ , temperature  $T$ , and area  $A$  for  $\gamma = 1.4$  are illustrated in [Fig. 34.8](#) as functions of Mach number  $M$ .

**Figure 34.8** Isentropic flow.



## 34.7 Nozzles

A nozzle is a variable-area passage that accelerates fluid from low speed to high speed. It transforms thermal energy into kinetic energy. Nozzles are one of the more common fluid flow devices. They are used in turbines, rocket engines, aircraft gas turbine engines, and wind tunnels. The basic operating principles of nozzles are presented in this section.

The reference ideal nozzle process is an isentropic expansion from the inlet stagnation condition to the desired exit condition, as illustrated on the Mollier diagram in Fig. 34.7. Nozzle flows can be analyzed by the equations presented in the previous section. There are two types of nozzles: converging nozzles, which are used when the nozzle pressure ratio,  $NPR = P_0/P_{\text{ambient}}$ , is modest; and converging-diverging nozzles, which are used when  $NPR$  is large.

Figure 34.9 illustrates a converging nozzle and the corresponding pressure distributions for a range of  $NPR$ s. As  $P_{\text{ambient}}$  is lowered, flow accelerates through the nozzle until  $P_{\text{exit}} = P_{\text{ambient}}$ . When  $P_{\text{ambient}} = P^*$  the flow is choked,  $M_{\text{exit}} = 1.0$ , and further decreases in  $P_{\text{ambient}}$  have no effect on the internal nozzle flow.

**Figure 34.9** Converging nozzle.

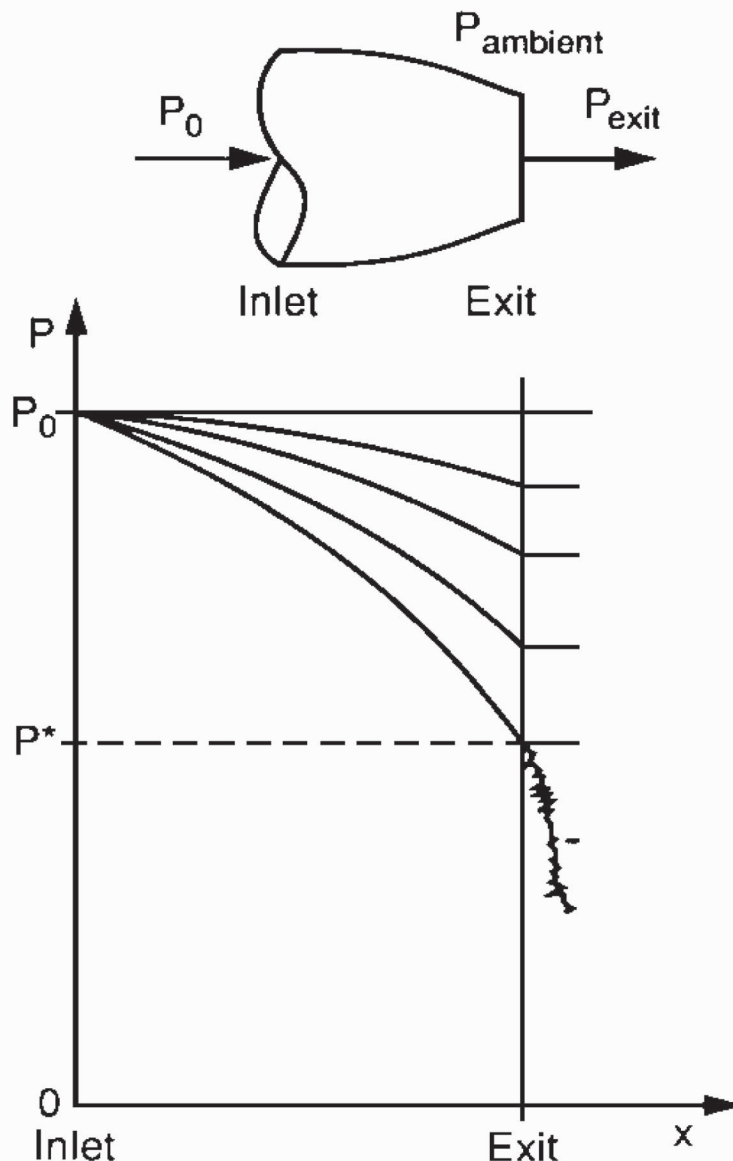
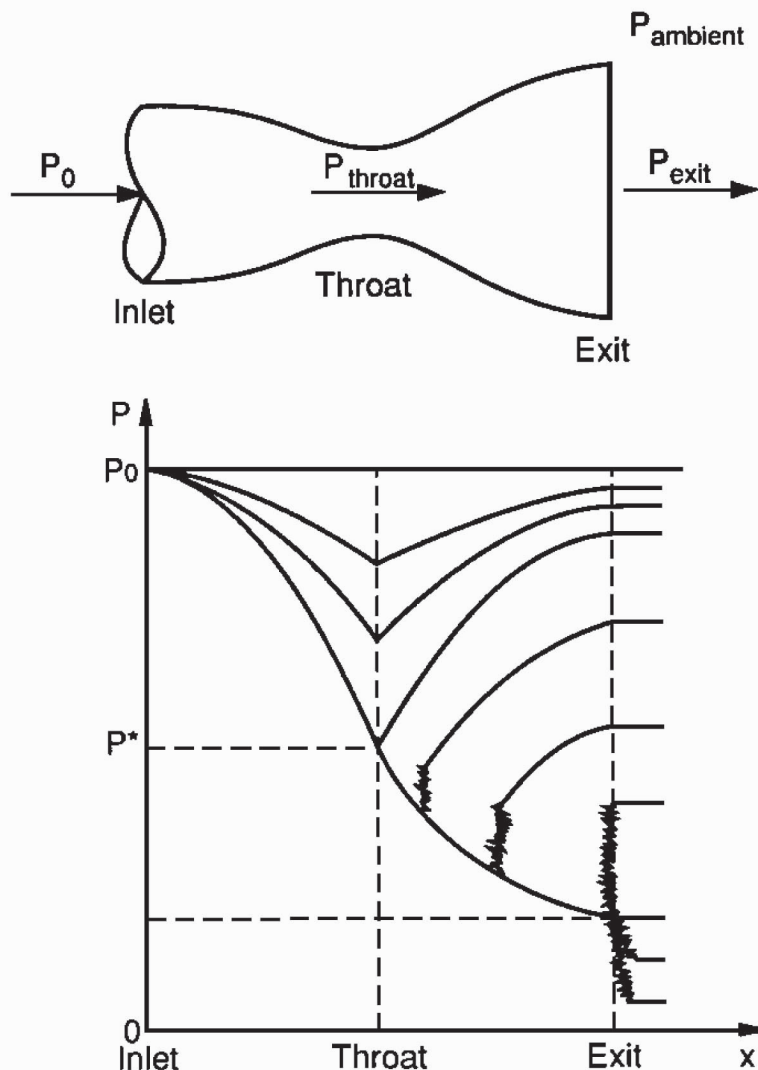


Figure 34.10 illustrates a converging-diverging nozzle and the corresponding pressure distributions for a range of  $NPR$ s. As  $P_{\text{ambient}}$  is lowered, the flow accelerates to the throat (the section of minimum area) then decelerates until  $P_{\text{exit}} = P_{\text{ambient}}$ . As  $P_{\text{ambient}}$  is lowered, the flow eventually chokes,  $M_{\text{throat}} = 1.0$  and  $P_{\text{throat}} = P^*$ . The flow in the diverging portion of the nozzle decelerates until  $P_{\text{exit}} = P_{\text{ambient}}$ . No further changes occur in the converging section of the nozzle as  $P_{\text{ambient}}$  is decreased further. Further decreases in  $P_{\text{ambient}}$  cause the flow to accelerate to supersonic downstream of the throat. Initially the supersonic flow is terminated by a shock wave followed by a subsonic deceleration to  $P_{\text{ambient}}$ . As  $P_{\text{ambient}}$  is decreased further, the shock wave moves downstream and eventually stands at the nozzle exit. The nozzle flow is then said to be "started." The supersonic exit flow adjusts to  $P_{\text{ambient}}$  through the shock wave in the nozzle exit plane. As  $P_{\text{ambient}}$  is reduced further, the shock wave becomes an oblique shock wave attached to the nozzle exit, which becomes weaker and weaker until it disappears and  $P_{\text{exit}} = P_{\text{ambient}}$ . This condition is called *optimum expansion*. Further decreases in  $P_{\text{ambient}}$  cause expansion waves outside of the nozzle.

**Figure 34.10** Converging-diverging nozzle.



## 34.8 Shock Waves

---

It has been observed for many years that a compressible fluid flowing supersonically can, under certain conditions, experience an abrupt change of state called a *shock wave*. As illustrated in Fig. 34.2, shock waves can be normal or oblique to the incoming flow direction. The equations for property ratios across a normal shock wave, from state 1 to state 2, are summarized below [Zucrow and Hoffman, 1976]. These equations are plotted in Fig. 34.11 for  $\gamma = 1.4$ .

$$M_2 = \left( \frac{M_1^2 + \frac{2}{\gamma - 1}}{\frac{2\gamma}{\gamma - 1}M_1^2 - 1} \right)^{1/2} \quad (34.22)$$

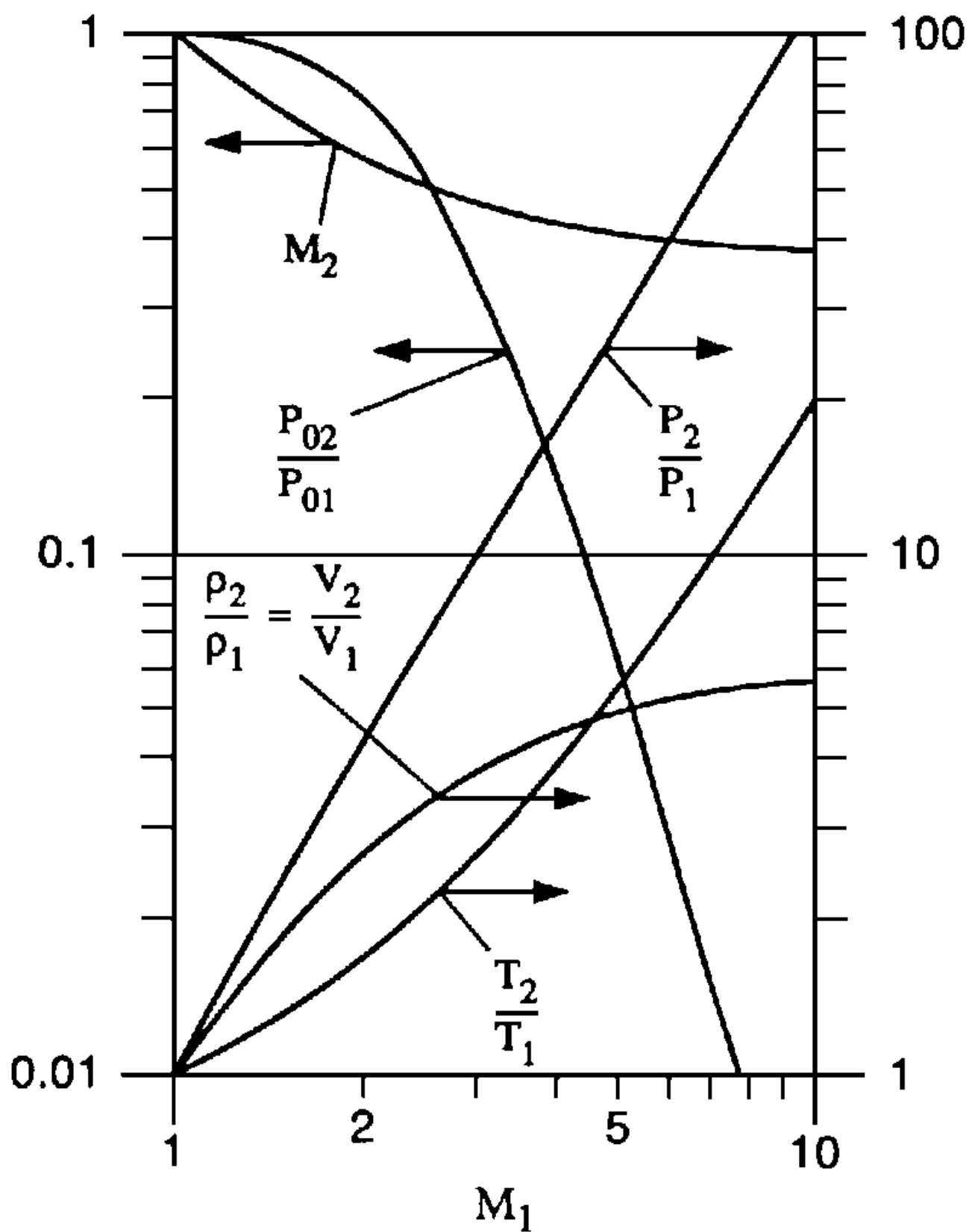
$$\frac{P_2}{P_1} = \frac{2\gamma}{\gamma + 1}M_1^2 - \frac{\gamma - 1}{\gamma + 1} \quad (34.23)$$

$$\frac{T_2}{T_1} = \left( \frac{2\gamma}{\gamma + 1}M_1^2 - \frac{\gamma - 1}{\gamma + 1} \right) \left( \frac{\gamma - 1}{\gamma + 1} + \frac{2}{(\gamma + 1)M_1^2} \right) \quad (34.24)$$

$$\frac{\rho_2}{\rho_1} = \frac{V_1}{V_2} = \frac{(\gamma + 1)M_1^2}{2 + (\gamma - 1)M_1^2} \quad (34.25)$$

$$\frac{P_{02}}{P_{01}} = \left[ \left( \frac{\gamma - 1}{\gamma + 1} + \frac{2}{(\gamma + 1)M_1^2} \right)^\gamma \left( \frac{2\gamma}{\gamma + 1}M_1^2 - \frac{\gamma - 1}{\gamma + 1} \right) \right]^{-1/(\gamma - 1)} \quad (34.26)$$

**Figure 34.11** Normal shock waves.



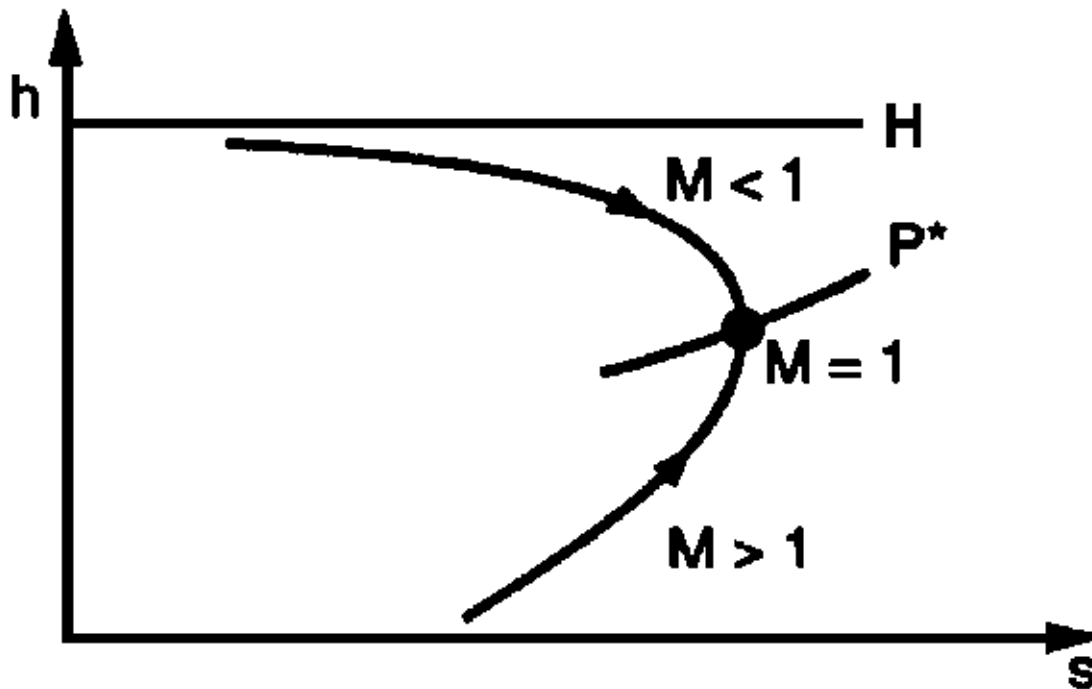


The analysis of oblique shock waves is considerably more complicated, since the angle of the shock wave and the upstream Mach number both influence the property ratios.

## 34.9 Friction and Heat Transfer

In the adiabatic (i.e.,  $\delta Q = 0$ ) flow of an incompressible fluid with friction,  $\rho$  and  $V$  are constant,  $P$  decreases, and  $T$  increases. However, the increase in  $T$  does not influence  $\rho$ ,  $V$ , or  $P$ . The effects of friction on the adiabatic flow of a compressible fluid are illustrated in Fig. 34.12, which presents the flow path on a Mollier diagram. The upper portion of the curve corresponds to subsonic flow and the lower portion corresponds to supersonic flow. The flow chokes at the nose of the curve where  $M = 1$ , which is the maximum entropy point. For subsonic flow,  $P$ ,  $\rho$ ,  $T$ , and  $P_0$  decrease and  $M$  and  $V$  increase along the flow path. For supersonic flow,  $P$ ,  $\rho$ , and  $T$  increase and  $M$ ,  $V$ , and  $P_0$  decrease along the flow path. A complete analysis is given in Zucrow and Hoffman [1976].

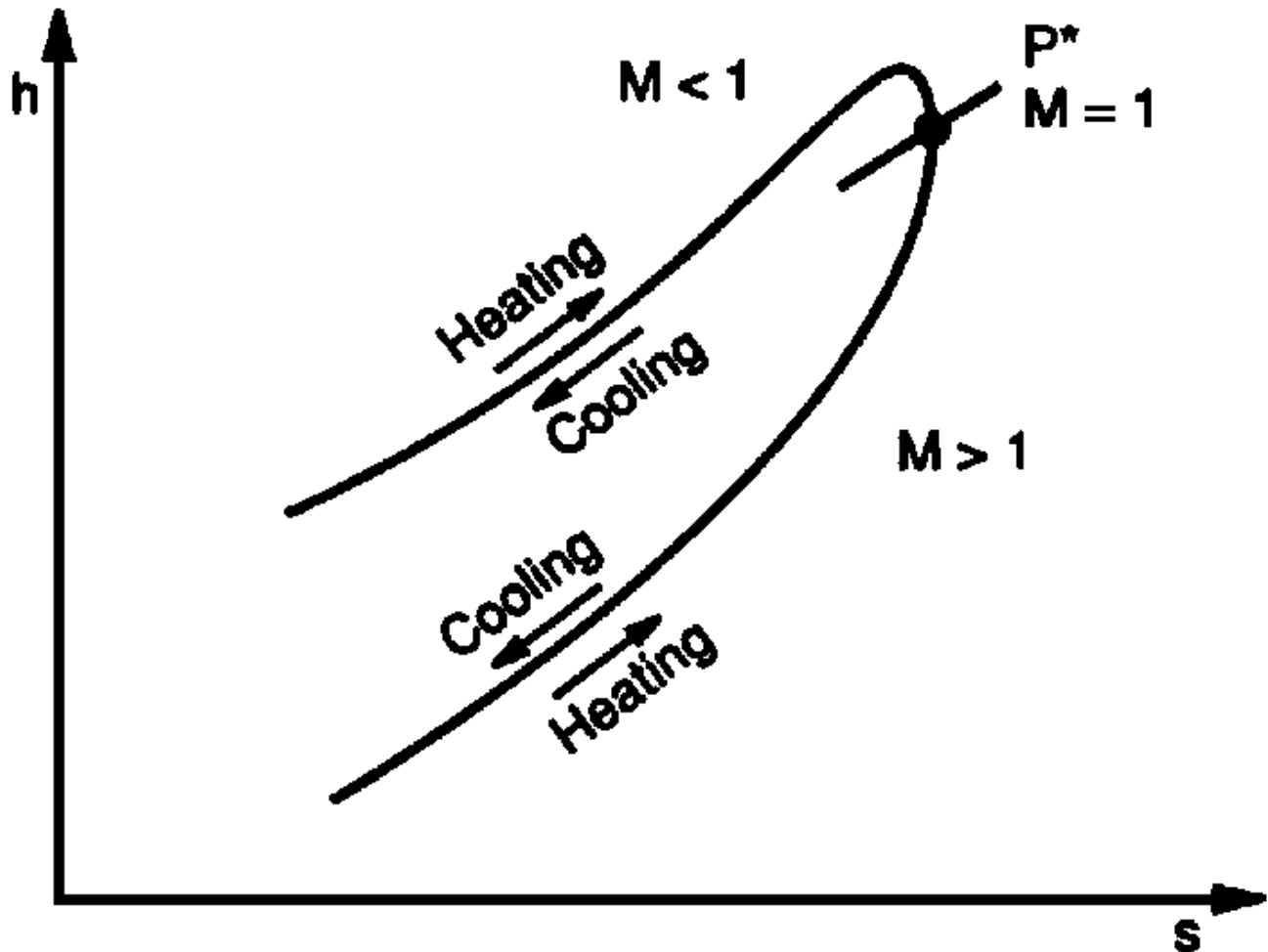
**Figure 34.12** Flow with friction.



In the frictionless flow of an incompressible flow with heat addition,  $T$  increases with heat addition, but  $\rho$ ,  $V$ , and  $P$  remain constant. The effects of heat transfer on the frictionless flow of a compressible fluid are illustrated in Fig. 34.13, which presents the flow path on a Mollier diagram. The upper portion of the curve corresponds to subsonic flow and the lower portion corresponds to supersonic flow. The flow chokes at the nose of the curve where  $M = 1$ , which is the maximum entropy point. For subsonic flow,  $M$ ,  $V$ , and  $T_0$  increase and  $P$ ,  $\rho$ , and  $P_0$  decrease along the flow path. Initially  $T$  increases until  $M \approx 0.85$ , then  $T$  decreases. For supersonic flow,  $P$ ,  $\rho$ ,  $T$ , and  $T_0$

increase and  $M$ ,  $V$ , and  $P_0$  decrease along the flow path. All of these effects are reversed for heat removal. A complete analysis is given in Zucrow and Hoffman [1976].

**Figure 34.13** Flow with heat transfer.



### Defining Terms

**Choking:** The condition in a compressible flow in which the Mach number is 1.0 at the critical point, the passage is passing the maximum flow rate, and the flow field upstream of the critical point is independent of the downstream conditions.

**Speed of sound:** The speed of propagation of small disturbances in a compressible fluid.

### References

Anderson, J. D. 1990. *Modern Compressible Flow*. McGraw-Hill, New York.

- Anderson, D. A., Tannehill, J. C., and Pletcher, R. H. 1984. *Computational Fluid Mechanics and Heat Transfer*. Hemisphere, New York.
- Fox, R. W. and McDonald, A. T. 1992. *Introduction to Fluid Mechanics*, 4th ed. John Wiley & Sons, New York.
- Owczarek, J. A. 1964. *Fundamentals of Gas Dynamics*. International Textbook Company, Scranton, PA.
- Shapiro, A. H. 1953. *The Dynamics and Thermodynamics of Compressible Fluid Flow, Volumes I and II*. Ronald Press, New York.
- Zucrow, M. J. and Hoffman, J. D. 1976. *Gas Dynamics, Volumes I and II*. John Wiley & Sons, New York.

### **Further Information**

For further information, see the sources listed in **Chapter 33, "Incompressible Fluids."**

Doraiswamy, D. "The Rheology of Non-Newtonian Fluids"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# The Rheology of Non-Newtonian Fluids

---

## 35.1 Kinematics, Flow Classification, and Material Functions

### 35.2 Fluids

### 35.3 Constitutive Equations

### 35.4 Some Useful Correlations for Material Functions

Equivalence of Dynamic and Steady Shear Properties • Dependence of Viscosity on Temperature •  
Dependence of Viscosity on Molecular Weight and Concentration

**Deepak Doraiswamy**

*E.I. du Pont de Nemours & Co.*

Rheology may be defined as the study of the deformation and flow of matter under the influence of imposed stresses. In Newtonian fluids the stress varies linearly with the strain (or deformation) rate at constant temperature, with the constant of proportionality being the viscosity. All fluids that do not follow this simple behavior are classified as non-Newtonian. The practicing engineer is often concerned with characterizing the rheological behavior of fluids by means of rheological material functions in well-defined simple flows. This enables development of a constitutive equation that is the relationship between the stress tensor and the rate of deformation tensor. Such an equation should, in principle, enable characterization of the fluid structure and calculation of the fluid stresses for the kinematics of interest, which could involve a complex time-dependent velocity field. A physical restriction on constitutive equations, one which translates into a mathematical requirement, is that they must satisfy the principle of **material objectivity**—that is, they should be independent of the reference frame used to describe the fluid motion. An important consequence of this constraint is that all scalars associated with material functions should depend on **invariants** of the associated tensors. The following section is concerned with the rheological description and characterization of incompressible non-Newtonian liquids.

## 35.1 Kinematics, Flow Classification, and Material Functions

---

It is convenient to use two reference flows—simple shear flow and simple elongational flow—as a practical basis for evaluation of complex situations. The associated material functions in steady flow are the most widely used in rheological characterization of fluids. The kinematics of a flow field may be defined in terms of the fluid velocity field or in terms of particle displacements. It is usually more convenient to work with the deformation rate tensor  $\dot{\gamma}_{ij}$ , which, in Cartesian

coordinates  $x_i$ , is related to the velocity gradient tensor  $v_{ij}$  as follows:

$$\dot{\gamma}_{ij} = v_{ij} + v_{ji} = \frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \quad (35.1)$$

It is often useful to isolate the rotational component of the fluid motion, which is defined by the vorticity tensor:

$$\omega_{ij} = \frac{\partial v_i}{\partial x_j} - \frac{\partial v_j}{\partial x_i} \quad (35.2)$$

The deformation rate tensor  $\dot{\gamma}_{ij}$  is a function only of the present time  $t$ —that is,  $\dot{\gamma}_{ij} = \dot{\gamma}_{ij}(t)$ . It is valid for arbitrarily large deformations since it does not depend on past times  $t'$ . If the kinematics are to be described using deformations instead of deformation rates, it is necessary to consider a spatial reference state at a past time  $t'$ . The infinitesimal strain tensor  $\gamma_{ij}(t, t')$ , which is a measure of the strain at time  $t'$  relative to the reference state at time  $t$ , is related to the deformation rate tensor  $\dot{\gamma}_{ij}(t)$  by a simple integration  $\gamma_{ij}(t, t') = \int_t^{t'} \dot{\gamma}_{ij}(t'') dt''$  only in the limit of vanishingly small deformations. For finite deformations  $\gamma_{ij}(t, t')$  violates the principle of material objectivity, and appropriate frame invariant measures of strain need to be defined; these are described later in the section on integral constitutive equations.

A homogeneous flow field is one in which the velocity gradient is constant at all points. In simple shear flow there is a nonzero component of velocity in only one direction. In a Cartesian ( $x, y, z$ ) coordinate system the velocity field is defined by  $v_x = \dot{\gamma}y$ ,  $v_y = v_z = 0$ . The scalar  $\dot{\gamma}$  is termed the *shear rate* and, for the most general situation, is related to the magnitude of the second invariant of  $\dot{\gamma}_{ij}$  by  $\dot{\gamma} = (II_{\dot{\gamma}})^{1/2}$ . The stress tensor  $\tau_{ij}$  and the deformation rate tensor  $\dot{\gamma}_{ij}$  enable definition of three independent material functions that are sufficient to characterize simple shear flow:

$$\begin{aligned} \eta(\dot{\gamma}) &= \frac{\tau_{xy}}{\dot{\gamma}}, \\ \psi_1(\dot{\gamma}) &= \frac{(\tau_{xx} - \tau_{yy})}{\dot{\gamma}^2}, \quad \text{and} \\ \psi_2(\dot{\gamma}) &= \frac{(\tau_{yy} - \tau_{zz})}{\dot{\gamma}^2} \end{aligned} \quad (35.3)$$

where  $\eta(\dot{\gamma})$  is the shear viscosity and  $\psi_1$  and  $\psi_2$  are the first and second normal stress coefficients, respectively. Unlike most non-Newtonian systems, Newtonian fluids do not display stresses normal to the shear stress, and the coefficients  $\psi_1$  and  $\psi_2$  are therefore zero.

For dilute solutions it is useful to define the intrinsic viscosity  $[\eta]$  (which has dimensions of reciprocal concentration) as follows:

$$[\eta] = \lim_{c \rightarrow 0} \frac{\eta - \eta_s}{c\eta_s} \quad (35.4)$$

where  $\eta_s$  is the solvent viscosity and  $c$  is the mass concentration of the solute. At very low shear rates, the intrinsic viscosity approaches a limiting value  $[\eta]_0$  known as the zero-shear rate intrinsic viscosity.

Extensional flow is defined kinematically by a rate of deformation tensor that has only diagonal components. The most common extensional flow is the simple uniaxial extension, for which the velocity field has the form  $v_x = -1/2\dot{\epsilon}x$ ,  $v_y = -1/2\dot{\epsilon}y$ , and  $v_z = +\dot{\epsilon}z$  in Cartesian coordinates. It is approximated by the stretching of a filament by forces exerted at both ends. The scalar coefficient  $\dot{\epsilon}$  is called the principal extension rate. This flow is completely characterized by a single material function, the elongational (or extensional) viscosity  $\bar{\eta}$ , which, for typical non-Newtonian systems (like polymer melts), can be several orders of magnitude higher than the shear viscosity  $\eta$ :

$$\bar{\eta}(\dot{\epsilon}) = \frac{(\tau_{zz} - \tau_{xx})}{\dot{\epsilon}} = \frac{(\tau_{zz} - \tau_{yy})}{\dot{\epsilon}} \quad (35.5)$$

The extensional viscosity has a value three times that of the shear viscosity for a Newtonian fluid (called the Trouton relation).

**Viscometric** and steady extensional flows are motions of constant strain rate history in which the material properties are independent of time (although they usually have a strong dependence on strain rate). Time-dependent material properties are required for describing non-Newtonian behavior because the effect of past deformations (varying strain history) on the structure of the fluid (and consequently the stress) cannot be ignored.

A class of transient shearing flows that is widely used in rheological measurements is small-amplitude oscillatory shear flow, in which the stresses and material properties depend on time (or frequency). These involve measurement of the stress response of a fluid to an imposed sinusoidal shearing strain or shear rate of the form  $\dot{\gamma}_{xy} = \gamma_o \omega \cos \omega t$ , where  $\gamma_o$ ,  $\omega$ , and  $t$  are the maximum imposed amplitude, the frequency, and the time, respectively. The shear stress response is then given by

$$\tau_{xy} = \gamma_o G'(\omega) \sin \omega t + \gamma_o G''(\omega) \cos \omega t \quad (35.6)$$

which provides a definition of the storage modulus,  $G'(\omega)$ , and the loss modulus,  $G''(\omega)$  (or the analogous dynamic viscosity functions:  $\eta' = G'/\omega$  and  $\eta'' = G''/\omega$ ). For a purely elastic material the loss modulus  $G''$  is zero, and for a purely viscous material the storage modulus  $G'$  is zero.

Another commonly used time-dependent material property is the relaxation modulus  $G(t, \gamma_o)$ , which involves measurement of the shear stress relaxation of a material after the sudden imposition of a step strain,  $\gamma_o$ . It is more suited to describing solid-like materials than liquid-like materials and is defined by  $G(t, \gamma_o) = \tau_{xy}/\gamma_o$ .

For stress growth upon inception of steady shear flow defined by

$$\begin{aligned}\dot{\gamma}_{xy} &= 0, & t < 0 \\ &= \dot{\gamma}_o, & t \geq 0\end{aligned}\quad (35.7)$$

the associated material functions are

$$\begin{aligned}\eta^+(t, \dot{\gamma}_o) &= \frac{\tau_{xy}}{\dot{\gamma}}, \\ \psi_1^+(t, \dot{\gamma}_o) &= \frac{(\tau_{xx} - \tau_{yy})}{\dot{\gamma}_o^2}, \quad \text{and} \\ \psi_2^+(t, \dot{\gamma}_o) &= \frac{(\tau_{yy} - \tau_{zz})}{\dot{\gamma}_o^2}\end{aligned}\quad (35.8)$$

where the plus sign indicates that the shear rate is applied at positive times. The inverse experiment is stress relaxation after cessation of steady shear flow defined by

$$\begin{aligned}\dot{\gamma}_{xy} &= \dot{\gamma}_o, & t < 0 \\ &= 0, & t \geq 0\end{aligned}\quad (35.9)$$

in which the material functions can be defined in a similar manner:

$$\begin{aligned}\eta^-(t, \dot{\gamma}_o) &= \frac{\tau_{xy}}{\dot{\gamma}}, \\ \psi_1^-(t, \dot{\gamma}_o) &= \frac{(\tau_{xx} - \tau_{yy})}{\dot{\gamma}_o^2}, \quad \text{and} \\ \psi_2^-(t, \dot{\gamma}_o) &= \frac{(\tau_{yy} - \tau_{zz})}{\dot{\gamma}_o^2}\end{aligned}\quad (35.10)$$

Analogous transient quantities,  $\bar{\eta}^+(t, \dot{\epsilon}_o)$  and  $\bar{\eta}^-(t, \dot{\epsilon}_o)$ , can be defined for the stress growth or inception of simple extensional flow with an elongation rate  $\dot{\epsilon}_o$

$$\bar{\eta}^+(t, \dot{\epsilon}_o) \quad \text{or} \quad \bar{\eta}^-(t, \dot{\epsilon}_o) = \frac{\tau_{zz} - \tau_{xx}}{\dot{\epsilon}_o} \quad (35.11)$$

All the above properties become independent of the imposed strain in the limit of zero strain (that is, as  $\gamma_o \rightarrow 0$ ,  $\dot{\gamma}_o \rightarrow 0$ , or  $\dot{\epsilon}_o \rightarrow 0$ ), in which case the material functions depend only on the time. The behavior in the limit of small deformations is termed *linear viscoelasticity* and the related material properties are defined appropriately, for example,  $G(t) = G(t, \gamma_o)_{\gamma_o \rightarrow 0}$ . Linear viscoelasticity reveals information about material behavior in the unstrained state in which the molecular conformations and entanglements have their equilibrium values and is used more for material characterization and quality control applications than for process



modeling.

In addition to the kinematics, the material properties depend on the chemical constitution of the fluid (e.g., molecular weight, molecular weight distribution, polymer branching) and the physical state of the fluid (typically measured by the temperature and concentration); consequently, they have great utility in characterization and processing operations. Some of the more commonly used correlations for material functions are discussed later. The experimental science of determining rheological material properties such as those considered in this section is termed **rheometry**; further details are provided in Dealy [1982, 1994].

## 35.2 Fluids

---

Non-Newtonian fluids may be broadly classified by their ability to retain the memory of a past deformation (which is usually reflected in a time dependence of the material properties). Fluids that display memory effects usually exhibit elasticity. A fluid is identified as viscoelastic if the stresses in it persist after the deformation has ceased (typically manifested in the decay of the shear stress and primary normal stress difference after cessation of steady shear flow). The duration of time over which appreciable stresses persist after cessation of deformation gives an estimate of the relaxation time  $\lambda$  of the material. A dimensionless group commonly used for evaluating the role of fluid viscoelasticity is the Deborah number,  $De$ , defined by

$$De = \frac{\lambda}{T} \quad (35.12)$$

where  $\lambda$  is the relaxation time for the fluid (which can be estimated experimentally using an appropriate constitutive equation) and  $T$  is the characteristic time constant for the process of interest. Low values of the Deborah number therefore correspond to fluid-like behavior and high values to solid-like behavior.

The non-Newtonian materials most often encountered by the engineer are polymer melts, polymer solutions, and multiphase systems. Polymer melts and solutions are usually viscoelastic, that is, they are capable of storing elastic energy. Their viscosity–shear rate behavior typically exhibits a constant "zero shear rate viscosity" at low shear rates, followed by a shear thinning region where the viscosity decreases with shear rate. Unlike that of polymer solutions, the viscosity of melts rarely exhibits a second plateau—the infinite-shear viscosity—at the highest shear rates typically used in measurements. Also, melts usually display much higher viscosities than solutions. It is often convenient to classify solutions into dilute and concentrated systems; in the former, unlike the latter, the individual polymer chains rarely overlap. Based on simple scaling arguments it is usually assumed that the dilute solution regime ends when  $[\eta]_0 c \sim 1$ , where  $c$  is the polymer concentration (typically 5% by weight). The source of elasticity differs between melts and solutions. In polymer solutions elasticity is mainly associated with changes in the orientation and configuration of molecules as a result of polymer-solvent interactions. In the case of polymer melts (or concentrated solutions) it is the result of mobility constraints imposed by polymer-polymer interactions. Not all fluids that retain a memory of past deformations exhibit elasticity and, in some cases (typically in concentrated suspensions), time-dependent rheological properties may be related

to changes in fluid structure that are purely dissipative, as in thixotropic and rheopectic fluids. Under steady shear conditions, the viscosity decreases with time for thixotropic fluids, whereas it increases with time for rheopectic fluids. Suspensions also frequently display "yield behavior," that is, they do not flow unless a critical (or yield) stress is applied. A number of equations describing the yield behavior of suspensions are included in [Table 35.1](#). At low volume fractions ( $< 5\%$ ), the viscosity  $\eta_{sp}$  of a rigid suspension of spheres is related to the viscosity of the suspending fluid  $\eta_s$  by the Einstein equation:

$$\eta_{sp} = \eta_s(1 + 2.5\phi) \quad (35.13)$$

where  $\phi$  is the volume fraction. At higher concentrations the empirical Maron-Pierce equation is useful for estimating the viscosity of suspensions of rigid particles of narrow size distribution in viscous fluids at low deformation rates:

$$\eta_{sp} = \frac{\eta_s}{(1 - \phi/A)^2} \quad (35.14)$$

The empirical constant  $A$  may be loosely identified with the maximum packing fraction for the particulate system. It has a value of 0.68 for smooth spheres and may be approximated by the expression  $A = 0.54 - 0.013a$  for aspect ratio  $a$  in the range  $6 < a < 30$ . The distribution of particle sizes has little effect on suspension viscosity when the volumetric loading of solids is below 20%; at higher concentration levels the effects are very pronounced. In general, the addition of suspended solids increases the shear thinning as one moves into the shear-thinning range of deformation rates and the importance of elasticity appears to decrease with increased solids content. The rheology of spherical and anisotropic suspensions (with and without Brownian motion) is discussed by Gupta [1994].

**Table 35.1** Common Generalized Newtonian Fluids Described by the Functions  $\eta(\dot{\gamma})$  or  $\eta(\tau)$ .

Model	Equation	Comments
Power-law model (Ostwald–de Waele)	$\eta = K\dot{\gamma}^{n-1}$ $K$ = consistency index $n$ = power law index	For "pseudoplastic" liquids, $n < 1$ For "dilatant" liquids, $n > 1$
Carreau	$\frac{\eta - \eta_\infty}{\eta_o - \eta_\infty} = [1 + (\lambda\dot{\gamma})^2]^{(n-1)/2}$ $\eta_o$ = zero shear viscosity $\eta_\infty$ = infinite shear viscosity	Describes smooth transition from $\eta_o$ to $\eta_\infty$ typically observed in polymer solutions.
Bingham	$\eta = \infty, \quad \tau \leq \tau_y$ $\eta = \mu_0 + \frac{\tau_y}{\dot{\gamma}}, \quad \tau \geq \tau_y$ $\tau_y$ = yield stress $\mu_0$ = viscosity	Describes yield behavior typically observed in suspensions.

Herschel-Bulkley	$\eta = \infty, \quad \tau \leq \tau_y$ $\eta = \frac{\tau_y}{\dot{\gamma}} + K\dot{\gamma}^{n-1}, \quad \tau \geq \tau_y$	Describes behavior of shear-thinning suspensions.
Ellis	$\frac{\eta_o}{\eta} = 1 + \left( \frac{\tau}{\tau_{1/2}} \right)^{\alpha-1}$ $\tau_{1/2} = \text{value of } \tau \text{ at which } \eta = \eta_o/2$	Predicts zero-shear viscosity but difficult to use since it is not explicit in shear rate.

Note: All scalar stresses refer to the magnitude or second invariant of the stress tensor  $\tau_{ij}$ .

For multiphase systems comprising fluid drops in a suspending liquid at low deformation rates and volume fractions, the Taylor analogue to the Einstein equation may be used:

$$\eta_{sp} = \eta_s \left( 1 + \frac{1 + 2.5r}{1 + r} \phi \right) \quad (35.15)$$

where  $r$  is the ratio of the viscosity of the disperse phase to that of the continuous phase. Few generalizations can be made about blends of immiscible deformable systems in view of complications arising due to droplet breakup, coalescence, and morphology effects; further details are provided in Han [1981].

## 35.3 Constitutive Equations

No single constitutive equation is suitable for all purposes and the selection of one depends on the particular situation concerned. The form of the equation will reflect the type of flow (e.g., shear or elongation), the type of material (e.g., melt, concentrated solution or dilute solution), the solution scheme (differential or integral model) best suited to the problem of interest, and the particular situation to be portrayed (stress overshoot, flow instability, die-swell, etc.). The basic assumption of all constitutive equations is that the stress at any location and at any time in the flowing fluid depends on the entire flow history of the fluid element occupying that material point only (and not of adjacent elements). Constitutive equations may be broadly classified into rate (or differential) equations and integral equations (although many of the nonlinear differential forms may also have an equivalent integral expression). The development and use of a number of commonly used differential and integral models have been described by Astarita and Marrucci [1974], Bird *et al.* [1987a], and Larson [1988].

For a Newtonian fluid the scalar fluid viscosity  $\eta$  is defined by:

$$\tau_{ij} = \eta \dot{\gamma}_{ij} \quad (35.16)$$

A generalized Newtonian fluid is described by a constitutive equation in which the viscosity is only a function of the magnitude of the second invariant of the stress or the shear rate. Common empiricisms for the function  $\eta(\dot{\gamma})$  or  $\eta(\tau)$  (where  $\tau$  is the magnitude or second invariant of  $\tau_{ij}$ ) are summarized in Table 35.1. The generalized Newtonian fluid is best suited to describing steady

state shear flows or small deviations from such flows as long as the Deborah number is sufficiently low. Like the Newtonian fluid, it cannot describe normal stress effects or time-dependent elastic effects. In elongational flows and in rapidly changing flows, the generalized Newtonian models should not be used.

Linear viscoelasticity is primarily concerned with the description of fluid deformations that are very small or very slow. The theory of linear viscoelasticity does not satisfy the principle of frame invariance except in the zero deformation limit. Differential formulations of linear viscoelastic models combine classical ideas of Hookean solids and Newtonian fluids and are represented by the Maxwell model,

$$\tau_{ij} + \lambda_1 \frac{\partial \tau_{ij}}{\partial t} = \eta_0 \dot{\gamma}_{ij} \quad (35.17)$$

and the Jeffreys model,

$$\tau_{ij} + \lambda_1 \frac{\partial \tau_{ij}}{\partial t} = \eta_0 \left( \dot{\gamma}_{ij} + \lambda_2 \frac{\partial \dot{\gamma}_{ij}}{\partial t} \right) \quad (35.18)$$

in which  $\lambda_1$  and  $\lambda_2$  are the relaxation time and the retardation time, respectively. The most general linear viscoelastic model, which includes both the above forms, is the generalized Maxwell model. It is conveniently represented in an integral form (which can therefore accommodate an infinite number of time constants) and is given by:

$$\tau_{ij} = \int_{-\infty}^t G(t-t') \dot{\gamma}_{ij}(t') dt' = \int_{-\infty}^t M(t-t') \gamma_{ij}(t, t') dt' \quad (35.19)$$

in which  $G(t-t')$  is the relaxation modulus and  $M(t-t') = dG(t-t')/dt'$  is the memory function. Various relationships between linear viscoelastic properties and material structure are provided by Ferry [1980].

The differences between the various nonlinear viscoelastic equations based on continuum mechanics are due primarily to the types of time derivative that arise as a result of rewriting them in an appropriate reference frame in order to make them objective. One of the most general frame invariant differential formulations for the stress tensor  $\tau_{ij}$  is the 8-constant Oldroyd model, which may be expressed by the following equation:

$$\begin{aligned} \tau_{ij} + \lambda_1 \frac{\mathcal{D}}{\mathcal{D}t} \tau_{ij} + \frac{\lambda_3}{2} (\dot{\gamma}_{ik} \tau_{kj} + \tau_{ik} \dot{\gamma}_{kj}) + \frac{\lambda_5}{2} \tau_{kk} \dot{\gamma}_{ij} + \frac{\lambda_6}{2} \tau_{ij} \dot{\gamma}_{ji} \delta_{ij} \\ = \eta_0 \left[ \dot{\gamma}_{ij} + \lambda_2 \frac{\mathcal{D}}{\mathcal{D}t} \tau_{ij} + \lambda_4 \dot{\gamma}_{ik} \dot{\gamma}_{kj} + \frac{\lambda_7}{2} \dot{\gamma}_{ij} \dot{\gamma}_{ji} \delta_{ij} \right] \end{aligned} \quad (35.20)$$

where  $\mathcal{D}/\mathcal{D}t$  is the corotational or Jaumann time derivative defined by:

$$\frac{\mathcal{D}}{\mathcal{D}t}\tau_{ij} = \frac{\partial\tau_{ij}}{\partial t} + v_i \frac{d\tau_{ij}}{dx_i} + \frac{1}{2}(\omega_{ik}\tau_{kj} - \tau_{ik}\omega_{kj}) \quad (35.21)$$

The 8-constant Oldroyd model includes a number of common differential models, some of which are summarized in Table 35.2 [Bird *et al.*, 1987a]. The steady state material functions for the various models can be obtained by assigning relevant values of the constants from Table 35.2 into the following expressions [Bird *et al.*, 1987a]:

$$\frac{\eta}{\eta_o} = \frac{1 + [\lambda_2(\lambda_3 + \lambda_5) + \lambda_4(\lambda_1 - \lambda_3 - \lambda_5) + \lambda_7(\lambda_1 - \lambda_3 - \frac{3}{2}\lambda_5)]\dot{\gamma}^2}{1 + [\lambda_1(\lambda_3 + \lambda_5) + \lambda_3(\lambda_1 - \lambda_3 - \lambda_5) + \lambda_6(\lambda_1 - \lambda_3 - \frac{3}{2}\lambda_5)]\dot{\gamma}^2} \quad (35.22)$$

$$\frac{\psi_1}{2\eta_o\lambda_1} = \frac{\eta(\dot{\gamma})}{\eta_o} - \frac{\lambda_2}{\lambda_1} \quad (35.23)$$

$$\frac{\psi_2}{2\eta_o\lambda_1} = -\frac{\psi_1}{2\eta_o\lambda_1} + \frac{(\lambda_1 - \lambda_3)\eta(\dot{\gamma})}{\lambda_1\eta_o} - \frac{(\lambda_2 - \lambda_4)}{\lambda_1} \quad (35.24)$$

$$\frac{\bar{\eta}}{3\eta_o} = \frac{1 - (\lambda_2 - \lambda_4)\dot{\epsilon} + (\frac{3}{2}\lambda_5 - \lambda_1 + \lambda_3)(2\lambda_2 - 2\lambda_4 - 3\lambda_7)\dot{\epsilon}^2}{1 - (\lambda_1 - \lambda_3)\dot{\epsilon} + (\frac{3}{2}\lambda_5 - \lambda_1 + \lambda_3)(2\lambda_1 - 2\lambda_3 - 3\lambda_6)\dot{\epsilon}^2} \quad (35.25)$$

The Oldroyd 8-parameter model illustrates how the number of constants becomes prohibitively large if only frame invariance considerations are used to formulate constitutive equations based purely on continuum mechanics considerations. It provides useful qualitative descriptions but is not quantitatively accurate. Care must be exercised in using the Oldroyd models in elongational flows since the elongational viscosity can become infinite if the parameters are not chosen carefully. Various empirical differential models have also been suggested. The White-Metzner model, for example, is obtained by making the relaxation time and the viscosity in the upper convected Maxwell model a function of the shear rate. It is useful in describing flows where the coupling of shear thinning and elasticity is significant. The Giesekus model can be obtained from molecular arguments and, unlike the Oldroyd model, is quadratic in stress. Although the model is considerably more difficult to use, it predicts decreasing viscosity and normal stress coefficients with increasing shear rate and a finite second normal stress coefficient. Differential constitutive equations are often the most convenient and practical way for determining the role of viscoelasticity in real-world systems. They are best suited to describing flows in which both shearing and extension are involved or shearing flows in which normal stresses are important.

**Table 35.2** Common Constitutive Equations Included in the 8-Constant Oldroyd Model

Model Name	Time Constants						
	$\lambda_1$	$\lambda_2$	$\lambda_3$	$\lambda_4$	$\lambda_5$	$\lambda_6$	$\lambda_7$
Oldroyd 6-constant						0	0
Oldroyd 4-constant			0	0		0	0
Oldroyd fluid A			$2\lambda_1$	$2\lambda_2$	0	0	0
Oldroyd fluid B			0	0	0	0	0
Corotational Jeffreys			$\lambda_1$	$\lambda_2$	0	0	0
Second-order fluid	0		0		0	0	0
Upper convected Maxwell		0	0	0	0	0	0

A major advantage of integral formulations is that they are usually explicit in stress. The integral constitutive equations are conveniently represented in terms of a finite strain tensor and not the velocity field that has been considered so far. The infinitesimal strain tensor  $\gamma_{ij}$  described earlier is valid only for vanishingly small displacements. In order to describe the strain or displacement for finite deformations, it is necessary to consider a reference configuration. If a particle occupies position  $x_i$  at time  $t$  and position  $x'_i$  at past time  $t'$ , specification of the displacement functions  $x_i(x'_i, t', t)$  or  $x'_i(x_i, t, t')$  is equivalent to specifying the velocity field. This enables definition of the displacement gradient tensor  $\Delta_{ij}$  and the inverse displacement gradient tensor  $\Delta_{ij}^{-1}$  given by:

$$\Delta_{ij}(x_i, t, t') = \frac{\partial x'_i}{\partial x_j} \quad (35.26)$$

and

$$\Delta_{ij}^{-1}(x_i, t, t') = \frac{\partial x_i}{\partial x'_j} \quad (35.27)$$

The principle of material objectivity then permits definition of finite strain tensors called the *Cauchy strain tensor*,  $c_{ij}$ , and the *Finger strain tensor*,  $f_{ij}$ , in terms of displacement gradients rather than the velocity field:

$$c_{ij} = \Delta_{ki} \Delta_{kj} \quad (35.28)$$

$$f_{ij} = \Delta_{ik}^{-1} \Delta_{jk}^{-1} \quad (35.29)$$

Both  $c_{ij}$  and  $f_{ij}$  describe material deformation independent of superimposed rigid rotations and reduce to the unit tensor in the undeformed state. Since rheology is concerned with deviations from the undeformed state, it is sometimes more convenient to work with the relative Cauchy strain tensor  $C_{ij}$  and the relative Finger strain tensor  $F_{ij}$ , defined as follows:

$$C_{ij} = c_{ij} - \delta_{ij} = \Delta_{ki} \Delta_{kj} - \delta_{ij} \quad (35.30)$$

$$F_{ij} = \delta_{ij} - f_{ij} = \delta_{ij} - \Delta_{ik}^{-1} \Delta_{jk}^{-1} \quad (35.31)$$

Any of the finite strain tensors ( $f_{ij}$ ,  $c_{ij}$ ,  $F_{ij}$ ,  $C_{ij}$ ) or an appropriate combination of these can be used in integral formulations since they are all frame invariant measures of the fluid strain at time  $t'$  relative to that at the time  $t$ .

The most general formulation for stress, which contains a number of constitutive equations, is the memory integral expansion in which the stress is expressed as a functional of the strain history [Bird *et al.*, 1987a]. It is obtained by integration over all past times  $t'$  following the particle that ends up at position  $x$  at present time  $t$ :

$$\begin{aligned} \tau_{ij}(x_i, t) = & \int_{-\infty}^t M_I(t-t') F'_{ij} dt' \\ & + \int_{-\infty}^t \int_{-\infty}^t M_{II}(t-t', t-t'') (F'_{ik} F''_{kj} + F''_{ik} F'_{kj}) dt' dt'' + \dots \end{aligned} \quad (35.32)$$

where  $F'_{ij} = F_{ij}(x_i, t, t')$ ,  $F''_{ij} = F_{ij}(x_i, t, t'')$ , and  $M_I, M_{II}, \dots$  are kernel functions that account for memory effects. The single integral forms are the most practical and are obtained by setting all higher-dimensional integrals equal to zero. The most popular empirical integral model is the factorized K-BKZ (Kaye-Bernstein, Kearsley, Zapas) model given by:

$$\tau_{ij}(x_i, t) = \int_{-\infty}^t M(t-t') \left[ \frac{\partial W}{\partial I_f} F'_{ij}(x_i, t, t') + \frac{\partial W}{\partial I_c} C'_{ij}(x_i, t, t') \right] dt' \quad (35.33)$$

in which  $M(t-t') (\equiv M_I)$  is the linear viscoelastic memory function defined in Eq. (35.19),  $C'_{ij} = C_{ij}(x_i, t, t')$ , and  $W(I_f, I_c)$  is a potential function that needs to be experimentally determined; it is a function of the two strain invariants  $I_f = f_{ii}$  and  $I_c = c_{ii}$ . The K-BKZ equation, which is based on rubber elasticity theory and makes no molecular assumptions, quantitatively describes important rheological phenomena such as stress growth and relaxation. The Lodge constitutive equation (which is the integral equivalent of the upper convected Maxwell model) is a special case of the K-BKZ equation and is obtained by setting  $W = I_c$ :

$$\tau_{ij}(x_i, t) = \int_{-\infty}^t M(t-t') C'_{ij}(x_i, t, t') dt' \quad (35.34)$$

In the limit of vanishingly small displacements the general linear viscoelastic model [Eq. (35.19)] is obtained from Eq. (35.34) by setting  $C_{ij}$  equal to the infinitesimal strain tensor,  $\gamma_{ij}$ . Integral models provide a framework for including a wide class of nonlinear viscoelastic behavior. By selecting suitable empirical forms for the kernel functions they enable description of fluid rheology using a small finite number of constants; these have the advantage of having physical meaning and can be determined from rheometric experiments. Quantitative determination of the kernel functions is, however, a nontrivial problem.

In addition to the continuum models described above, constitutive equations based on molecular theories (which result in integral forms similar to those discussed) have also been proposed. Most of these models can be classified in one of the following categories: (a) bead-spring or random coil theories for dilute polymer solutions, (b) hydrodynamic interaction theories for moderately concentrated solutions that account for the indirect drag that one part of a polymer chain exerts on another through the solvent, and (c) molecular entanglement, network, or reptation theories for concentrated solutions and melts. Although molecular models have considerable potential for material characterization and rheological flow modeling, their applicability is restricted at this time. Bird *et al.* [1987b] have covered such constitutive equations in great detail.

## 35.4 Some Useful Correlations for Material Functions

Correlations for and between the material functions are very useful in order to enable quick estimates of material properties; they are often used for material characterization and process control applications. The more commonly used correlations are described below [further details on correlations and parameter values are summarized in van Krevelen and Hoftyzer (1976), Graessley (1974) and Bird *et al.*, (1987a)].

### Equivalence of Dynamic and Steady Shear Properties

The Cox-Merz empiricism predicts that the magnitude of the complex viscosity ( $|\eta^*|$ ) is equal to that of the viscosity at equal values of the frequency and shear rate. That is,

$$\eta(\dot{\gamma}) = |\eta^*(\omega)|_{\omega=\dot{\gamma}} = \left( \sqrt{\eta'(\omega)^2 + \eta''(\omega)^2} \right)_{\omega=\dot{\gamma}} \quad (35.35)$$

Laun's rule is another useful correlation for predicting the first normal stress difference from dynamic data:

$$\psi_1 = \frac{2\eta''(\omega)}{\omega} \left[ 1 + \left( \frac{\eta''}{\eta'} \right)^{0.7} \right] \bigg|_{\omega=\dot{\gamma}} \quad (35.36)$$

### Dependence of Viscosity on Temperature

If  $\eta_o(T)$  is the zero-shear viscosity at the desired temperature  $T$ , and  $\eta_o(T_o)$  the zero shear viscosity at an arbitrary reference temperature  $T_o$ , a shift factor  $a_T$  may be approximated as follows:

$$a_T = \frac{\eta_o(T)T_o\rho_o}{\eta_o(T_o)T\rho} \cong \frac{\eta_o(T)}{\eta_o(T_o)} \quad (35.37)$$

where  $\rho$  and  $\rho_o$  are the densities at the temperatures  $T$  and  $T_o$ , respectively. An Arrhenius-type



equation is often used to determine the temperature dependence of  $a_T$  as long as the temperature is at least 100 K above the glass transition temperature,  $T_g$ :

$$a_T = \exp \left[ \frac{\Delta E}{R} \left( \frac{1}{T} - \frac{1}{T_o} \right) \right] \quad (35.38)$$

where  $\Delta E$  is the activation energy and  $R$  is the Boltzmann constant. For typical polymer melts  $\Delta E/R$  is  $\sim 5000$  K. For temperatures between  $T_g$  and  $T_g + 100$ , the Williams-Landel-Ferry expression can be used to estimate  $a_T$ :

$$\log a_T = \frac{-c_1^\circ(T - T_o)}{c_2^\circ + (T - T_o)} \quad (35.39)$$

If  $T_o$  is taken to be the glass-transition temperature, the following values can be used for quick estimates:  $c_1^\circ = 17.44$  and  $c_2^\circ = 51.6$  K.

The time-temperature superposition principle states that varying the temperature at fixed shear rate (or time) is equivalent to varying the shear rate at fixed temperature. By defining a reduced viscosity  $\eta_r$  and a reduced shear rate  $\dot{\gamma}_r$  as follows,

$$\eta_r = \eta(\dot{\gamma}, T) \frac{\eta_o(T_o)}{\eta_o(T)} \cong \frac{\eta(\dot{\gamma}, T)}{a_T}; \quad \dot{\gamma}_r = a_T \dot{\gamma} \quad (35.40)$$

it is possible to reduce viscosity data at different temperatures and shear rates to a single master curve, provided the different isotherms are of similar shape [Bird *et al.*, 1987a]. This approach enables extending the shear rate range of an available experimental configuration for obtaining viscometric data. Analogous procedures can be used for other material properties.

## Dependence of Viscosity on Molecular Weight and Concentration

Molecular weights of polymers can be determined using solution viscosity measurements. At very low concentration or in the zero concentration limit, the intrinsic viscosity of linear, monodisperse solutions of polymers is related to the molecular weight,  $M$ , by the Mark-Houwink equation:

$$[\eta]_o = K M^a \quad (35.41)$$

where  $K$  and  $a$  are functions of the polymer, solvent, and temperature; their values are available in the literature [e.g., van Krevelen and Hoftyzer (1976)]. A typical value for  $a$  is 0.7. At higher concentrations ( $c > 1/[\eta]_o$ ), for which overlap between neighboring molecules becomes significant, the viscosity scales as the product  $cM$ . Finally, for pure melts and highly concentrated solutions beyond a critical molecular weight (typically in the range 2000–50000), entanglements

between molecules become significant and the zero shear melt viscosity is given by:

$$\eta_o = KM^{3.4} \quad (35.42)$$

For narrow distributions  $M$  can be approximated by  $M_w$ . For broad distributions, the viscosity depends on a molecular weight average between  $M_w$  and the next higher average ( $z$  average). If branching is present,  $M_w$  is replaced by  $gM_w$ , where  $g$  is a branching index.

## Defining Terms

**Invariants of tensors:** A second-order tensor  $T_{ij}$  having three scalar invariants that are independent of the coordinate system to which the components of  $T_{ij}$  are referred. They are  $I_T = T_{ii}$ ,  $II_T = T_{ij}T_{ji}$  and  $III_T = T_{ij}T_{jk}T_{ki}$ .

**Material objectivity:** A principle that requires that the rheological description of a material should be independent of the reference frame used to describe the fluid motion.

**Relaxation time:** The duration of time over which appreciable stresses persist after cessation of deformation in a fluid (e.g., the characteristic time constant for exponential decay of stress).

**Rheometry:** The experimental science of determining rheological material properties.

**Viscometric flow:** A flow field in which the deformation as seen by a fluid element is indistinguishable from simple shear flow.

## References

- Astarita, G. and Marrucci, G. 1974. *Principles of Non-Newtonian Fluid Mechanics*. McGraw-Hill, London.
- Bird, R. B., Armstrong, R. C., and Hassager, O. 1987a. *Dynamics of Polymeric Liquids, Volume 1, Fluid Mechanics*, 2nd ed. John Wiley & Sons, New York.
- Bird, R. B., Armstrong, R. C., and Hassager, O. 1987b. *Dynamics of Polymeric Liquids, Volume 2, Kinetic Theory*, 2nd ed. John Wiley & Sons, New York.
- Dealy, J. M. 1982. *Rheometers for Molten Plastics*. Van Nostrand Reinhold, New York.
- Dealy, J. M. 1994. Official nomenclature for material functions describing the response of a viscoelastic fluid to various shearing and extensional deformations. *J. Rheol.* 38(1):179–191.
- Ferry, J. D. 1980. *Viscoelastic Properties of Polymers*, 3rd ed. John Wiley & Sons, New York.
- Graessley, W. W. 1974. The entanglement concept in polymer rheology. *Adv. Polym. Sci.* 16:1–179.
- Gupta, R. K. 1994. Particulate suspensions. In *Flow and Rheology in Polymer Composites Manufacturing*, ed. S. G. Advani, pp. 9–51. Elsevier, New York.
- Han, C. D. 1981. *Multiphase Flow in Polymer Processing*. Academic, New York.
- Larson, R. G. 1988. *Constitutive Equations for Polymer Melts and Solutions*. Butterworth, Boston.
- van Krevelen, D. W. and Hoftyzer, P. J. 1976. *Properties of Polymers*. Elsevier, Amsterdam.

## Further Information

The *Journal of Non-Newtonian Fluid Mechanics* (issued monthly), the *Journal of Rheology* (issued bimonthly), and *Rheologica Acta* (issued bimonthly) report advances in non-Newtonian fluids.

Dealy, J. M. and Wissbrun, K. F. 1990. *Melt Rheology and its Role in Plastics Processing*. Van Nostrand Reinhold, New York. Industrially relevant non-Newtonian systems (including liquid crystalline polymers) are discussed.

Crochet, M. J., Davies, A. R., and Walters, K. 1984. *Numerical Simulation of Non-Newtonian Flow*. Elsevier, Amsterdam.

Tanner, R. I., 1978. *Engineering Rheology*. Clarendon, Oxford.

Munson, B. R., Cronin, D. J. “Airfoils/Wings”  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

- 36.1 Nomenclature
- 36.2 Airfoil Shapes
- 36.3 Lift and Drag Characteristics for Airfoils
- 36.4 Lift and Drag of Wings

**Bruce R. Munson**

*Iowa State University*

**Dennis J. Cronin**

*Iowa State University*

A simplified sketch of a wing is shown in [Fig. 36.1](#). An **airfoil** is any cross section of the wing made by a plane parallel to the  $xz$  plane. The airfoil size and shape usually vary along the span.

Airfoils and wings are designed to generate a lift force,  $L$ , normal to the free stream flow that is considerably larger than the drag force,  $D$ , parallel to the free stream flow. The lift and drag are strongly dependent on the geometry (shape, size, orientation to the flow) of the wing and the speed at which it flies,  $V_o$ , as well as other parameters, including the density,  $\rho$ ; viscosity,  $\mu$ ; and speed of sound,  $a$ , of the air. The following sections discuss some properties of airfoils and wings.

## 36.1 Nomenclature

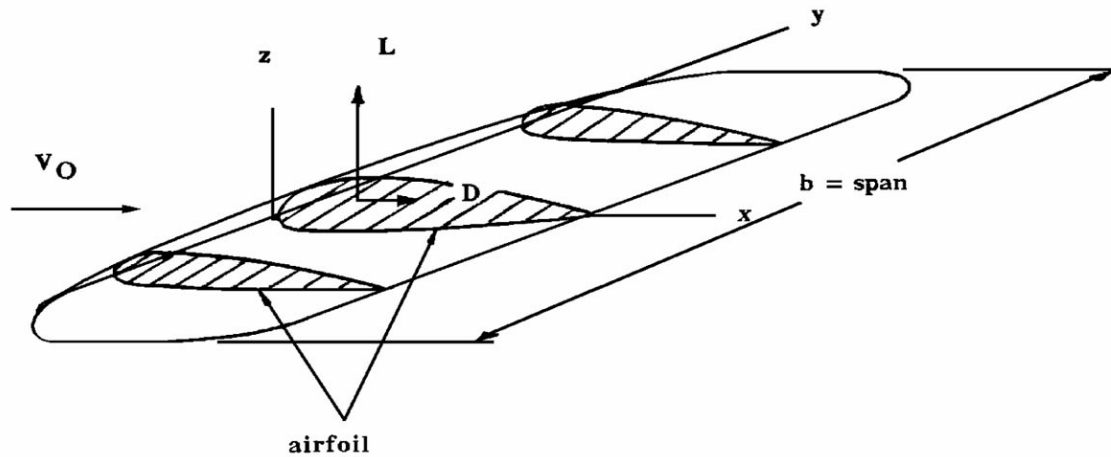
---

The shape, size, and orientation of an airfoil can be given in terms of the following parameters ([Fig. 36.2](#)): the **chord** length,  $c$ , the chord line that connects the leading and trailing edges; the **angle of attack**,  $\alpha$ , relative to the free stream velocity,  $V_o$ ; the mean **camber** line that is halfway between the upper and lower surfaces; and the thickness distribution,  $t$ , which is the distance between the upper and lower surfaces perpendicular to the camber line.

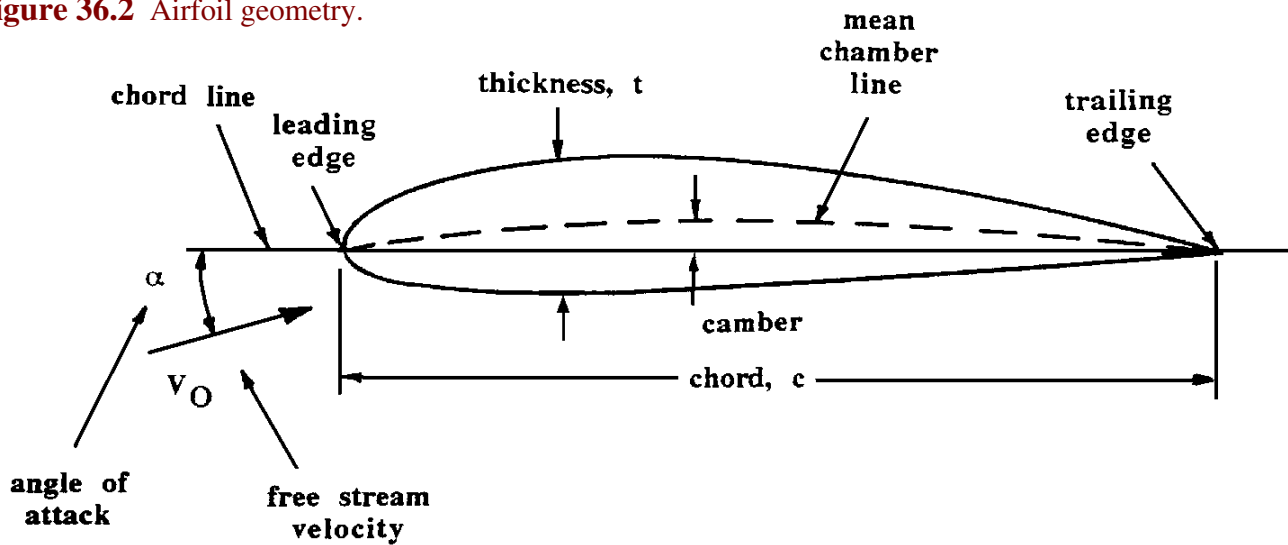
Various classes of airfoils have been developed over the years. These include the classic National Advisory Committee for Aeronautics four-, five-, and six-digit series airfoils (for example, the NACA 2412 airfoil used on the Cessna 150 or the NACA 64A109 used on the Gates Learjet [[Anderson, 1991](#)]) as well as numerous other modern airfoils [[Hubin, 1992](#)].

Performance characteristics of wings are normally given in terms of the dimensionless **lift coefficient** and **drag coefficient**,

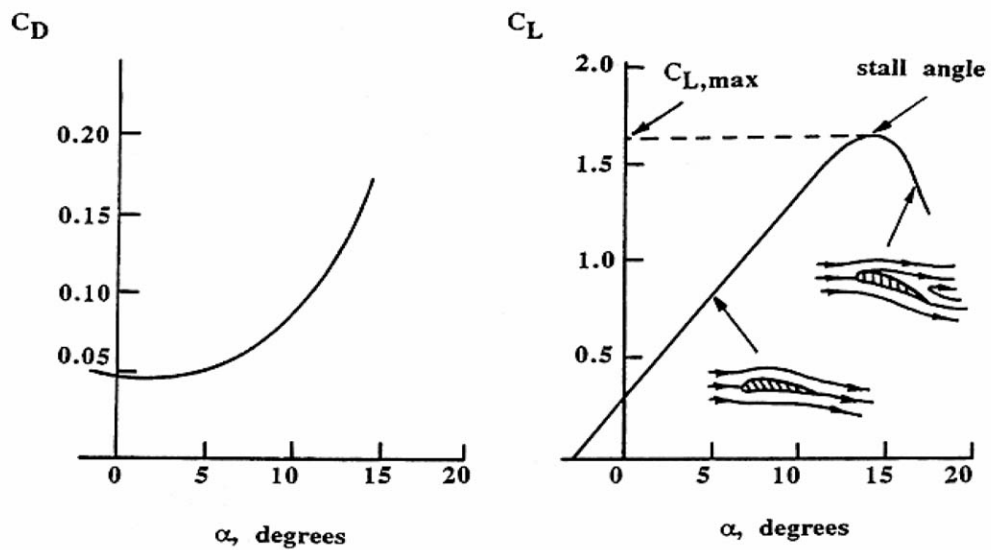
**Figure 36.1** Wing geometry.



**Figure 36.2** Airfoil geometry.



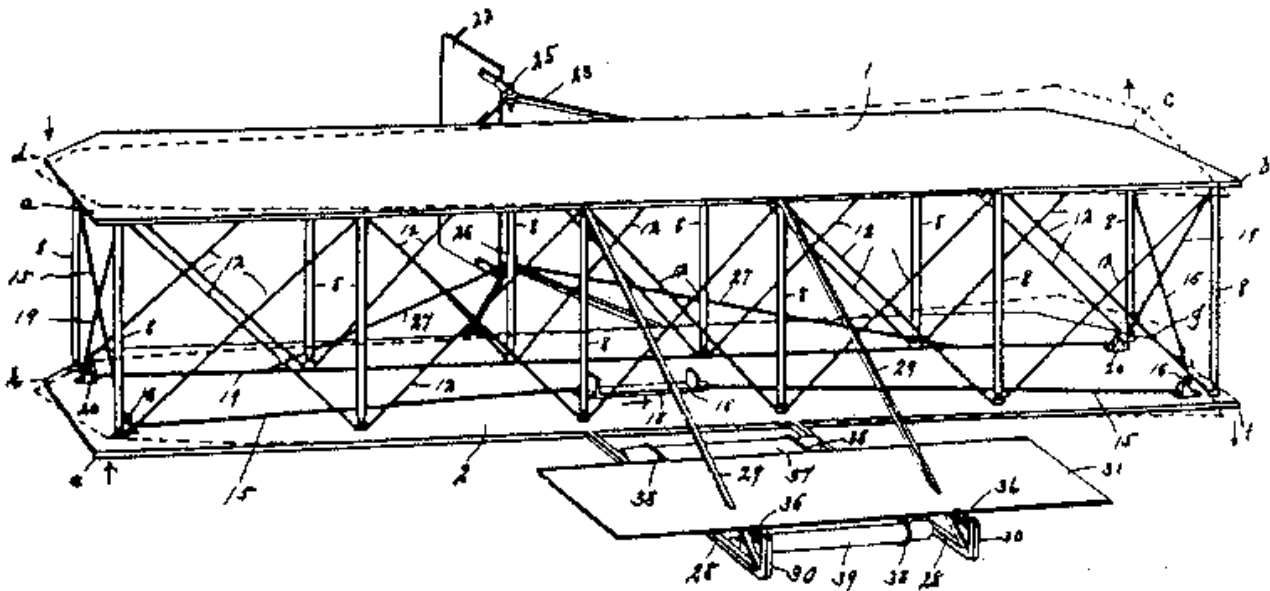
**Figure 36.3** Typical lift and drag coefficients as a function of angle of attack.



$$C_L = \frac{L}{q_o S} \quad (36.1)$$

$$C_D = \frac{D}{q_o S} \quad (36.2)$$

where  $q_o = \frac{1}{2} \rho V_o^2$  is the **dynamic pressure** and  $S$  is the planform area of the wing. The planform area is the area seen by looking onto the wing from above: the span times the chord for a rectangular wing. Typical characteristics for lift and drag coefficients as a function of the angle of attack are shown in Fig. 36.3. An efficient wing has a large lift-to-drag ratio—that is, a large  $C_L/C_D$ .



#### FLYING MACHINE

Orville & Wilbur Wright

Patented May 22, 1906

#821,393

An excerpt:

Our invention relates to that class of flying-machines in which the weight is sustained by the reactions resulting when one or more aeroplanes are moved through the air edgewise at a small angle of incidence, either by the application of mechanical power or by the utilization of the force of gravity.

The objects of our invention are to provide means for maintaining or restoring the equilibrium or lateral balance of the apparatus, to provide means for guiding the machine both vertically and horizontally, and to provide a structure combining lightness, strength, convenience of construction, and certain other advantages which hereinafter appear.

The Wrights became interested in flight after reading of successful glider flights in Germany. They built and flew three glider bi-planes before attempting their unassisted powered flight at Kitty Hawk, N. Carolina in 1903. The U.S. Army was interested in powered flight and eventually awarded the Wrights a contract for the first military aircraft in 1909. (©1992, DewRay Products, Inc. Used with permission.)

As the angle of attack is increased from small values, the lift coefficient increases nearly linearly with  $\alpha$ . At larger angles there is a sudden decrease in lift and a large increase in drag. This condition indicates that the wing has stalled. The airflow has separated from the upper surface, and an area of reverse flow exists (Fig. 36.3). **Stall** is a manifestation of boundary layer separation. This complex phenomenon is a result of viscous effects within a thin air layer (the boundary layer) near the upper surface of the wing in which viscous effects are important [Schlichting, 1979].

In addition to knowing the lift and drag for an airfoil, it is often necessary to know the location where these forces act. This location, the **center of pressure**, is important in determining the moments that tend to pitch the nose of the airplane up or down. Such information is often given in terms of a **moment coefficient**,

$$C_M = \frac{M}{q_o S c} \quad (36.3)$$

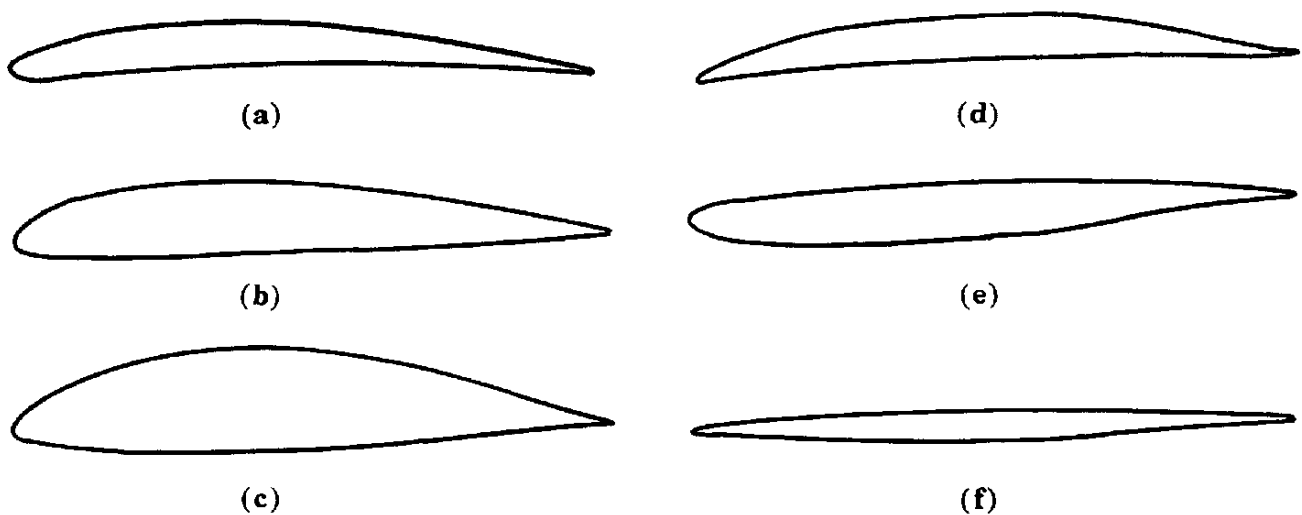
where  $M$  is the moment of the lift and drag forces about some specified point, often the leading edge.

For a given geometry, the lift and drag coefficients and the center of pressure (or moment coefficient) may depend on the flight speed and properties of the air. This dependence can be characterized in terms of the Reynolds number based on chord length,  $Re = \rho V_o c / \mu$ , and the Mach number,  $Ma = V_o / a$ . For modern commercial aircraft the Reynolds number is typically on the order of millions ( $10^6$ ). Mach numbers range from less than 1 (subsonic flight) to greater than 1 (supersonic flight).

## 36.2 Airfoil Shapes

As shown in Fig. 36.4, typical airfoil shapes have changed over the years in response to changes in flight requirements and because of increased knowledge of flow properties. Early airplanes used thin airfoils (maximum thickness 6 to 8% of the chord length), with only slight camber [Fig. 36.4(a)]. Subsequently, thicker (12 to 18% maximum thickness) airfoils were developed and used successfully on a variety of low-speed aircraft [Fig. 36.4(b) and (c)].

**Figure 36.4** Various airfoil shapes.





A relatively recent development (c. 1970s) has been design and construction of laminar flow airfoils that have smaller drag than previously obtainable [Fig. 36.4(d)]. This has resulted from new airfoil shapes and smooth surface construction that allow the flow over most of the airfoil to remain laminar rather than become turbulent. The performance of such airfoils can be quite sensitive to surface roughness (e.g., insects, ice, and rain) and Reynolds number effects.

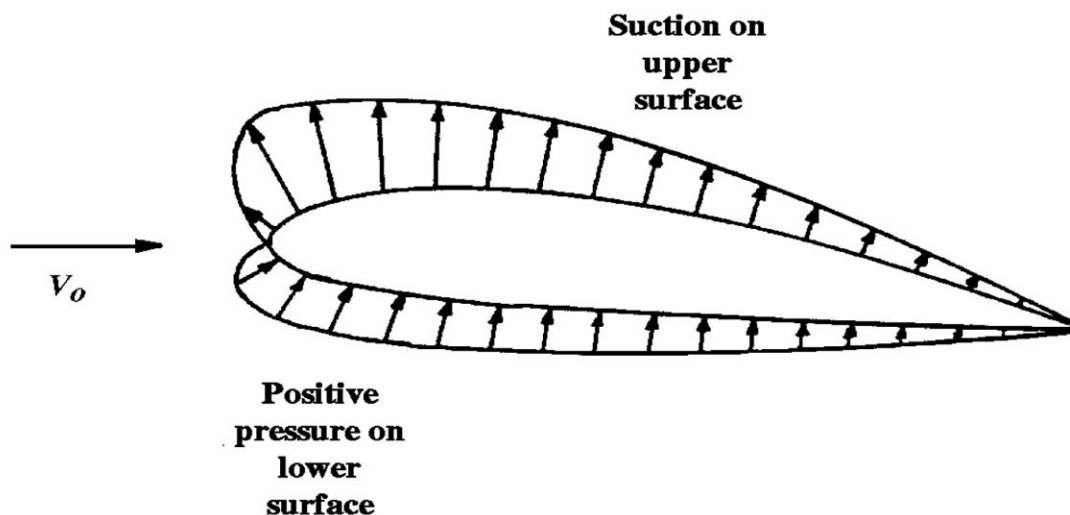
With the advent of commercial and business jet aircraft, it became necessary to develop airfoils that operate properly at Mach numbers close to unity. Since air, in general, accelerates as it passes around an airfoil, the flow may be locally supersonic near portions of the upper surface, even though the flight speed is subsonic. Such supersonic flow can cause shock waves (discontinuities in the flow) that degrade airfoil performance. Airfoils have been developed (denoted *supercritical airfoils*) to minimize the effect of the shock wave in this locally supersonic region [Fig. 36.4(e)].

Certain flow phenomena, such as shock waves, occur in supersonic flight that do not occur for subsonic flight [Anderson, 1991]. The result is that supersonic airfoils tend to be thinner and sharper than those for subsonic flight [see Fig. 36.4(f)].

### 36.3 Lift and Drag Characteristics for Airfoils

Lift and drag forces on airfoils are the result of pressure and viscous forces that the air imposes on the airfoil surfaces. Pressure is the dominant factor that produces lift. A typical pressure distribution is shown in Fig. 36.5. In simplistic terms, the air travels faster over the upper surface than it does over the lower surface. Hence, from **Bernoulli's principle** for steady flow, the pressure on the upper surface is lower than that on the bottom surface.

**Figure 36.5** Typical pressure distribution on an airfoil.



For an unstalled airfoil, most of the drag is due to viscous forces. This skin friction drag is a result of the shear stress distribution on the airfoil. For a stalled airfoil, pressure forces contribute significantly to the drag.

For a two-dimensional body such as an airfoil (a wing of infinite span), the section lift, drag, and moment coefficients (about a defined point) are based on the lift, drag, and moment per unit span,  $L'$  [force/length],  $D'$  [force/length], and  $M'$  [force · length/length], respectively. That is,

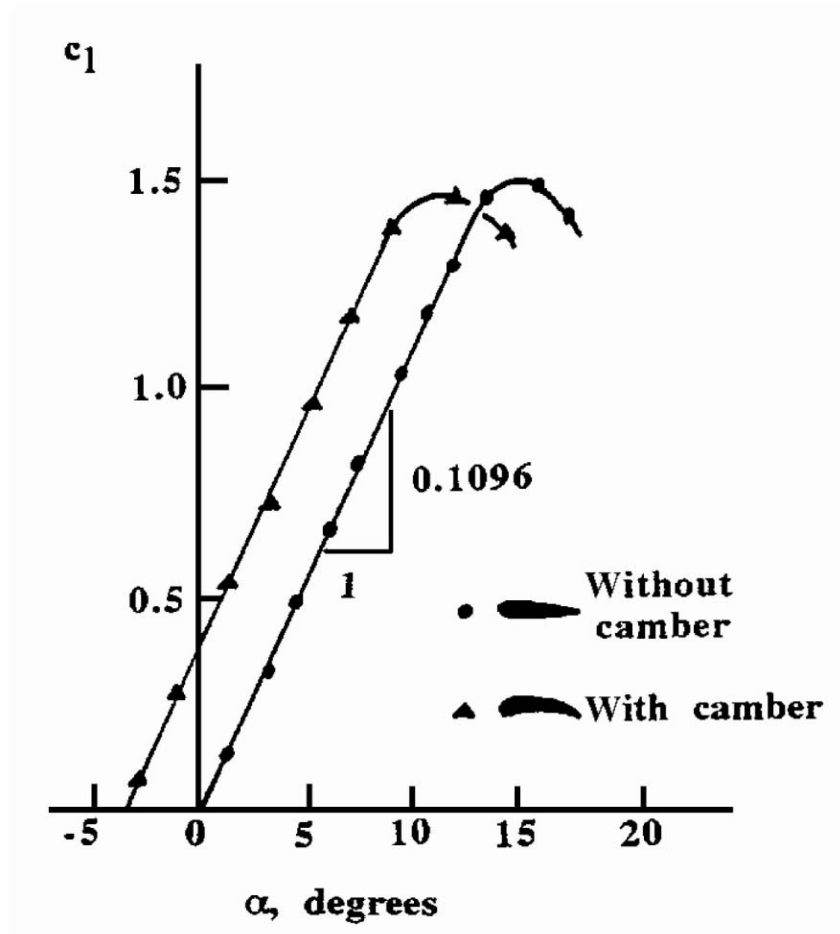
$$c_l = \frac{L'}{q_o c} \quad (36.4)$$

$$c_d = \frac{D'}{q_o c} \quad (36.5)$$

$$c_m = \frac{M'}{q_o c^2} \quad (36.6)$$

For most airfoils, the lift coefficient is nearly linear with angle of attack up to the stall angle. According to simple airfoil theory [and verified by experiment (Fig. 36.6)], the lift-curve slope,  $dc_l/d\alpha$ , is approximately equal to  $2\pi$  with  $\alpha$  in radians (or  $dc_l/d\alpha = 0.1096 \text{ deg}^{-1}$  when  $\alpha$  is in degrees) [Anderson, 1991].

**Figure 36.6** Effect of camber on airfoil lift coefficient.



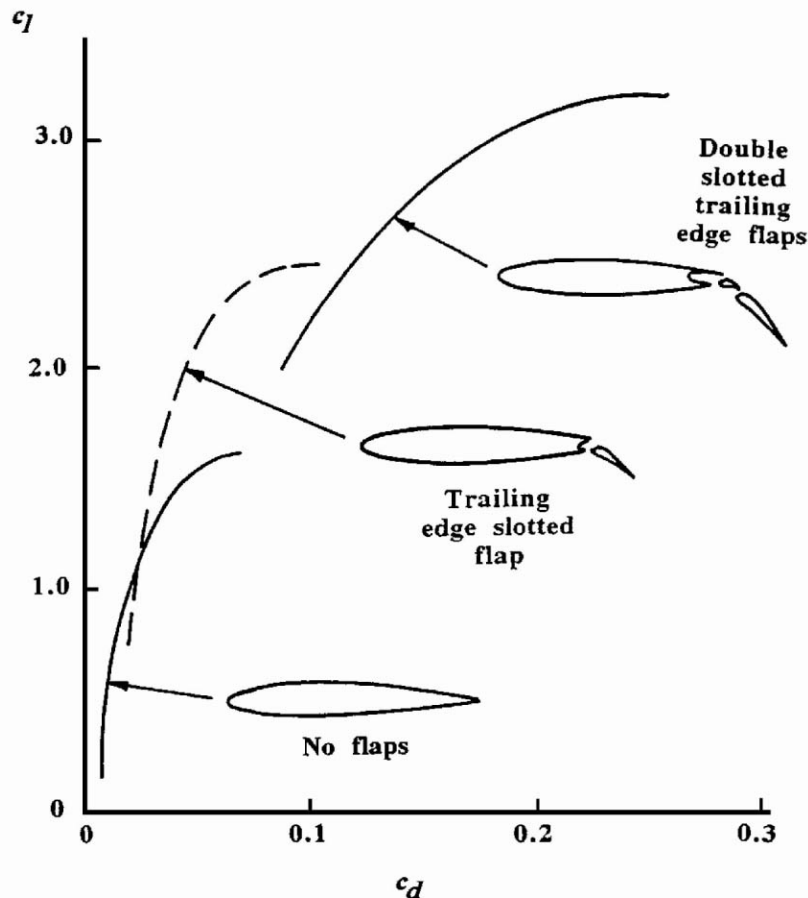
For symmetrical (no camber) airfoils, zero angle of attack produces zero lift; for airfoils with camber, the zero-lift condition ( $\alpha = \alpha_{0L}$ ) occurs at nonzero angle of attack (Fig. 36.6).

The maximum lift coefficient for an airfoil ( $c_l = c_{l,\max}$ ) is typically on the order of unity and occurs at the critical angle of attack ( $\alpha = \alpha_{CR}$ ) (Fig. 36.3). That is, the lift generated per unit span is on the order of the dynamic pressure times the planform area:  $L' = c_l q_o c \approx q_o c$ . The drag

coefficient, on the other hand, is on the order of 0.01. Hence, the maximum lift-to-drag ratio is on the order of 100.

As illustrated in Fig. 36.7, the airfoil geometry may be altered by using movable trailing or leading edge **flaps**. Such devices can significantly improve low-speed (i.e., landing or takeoff) performance by increasing the maximum lift coefficient, thereby reducing the required landing or takeoff speed.

**Figure 36.7** Effect of flaps on airfoil lift coefficient.

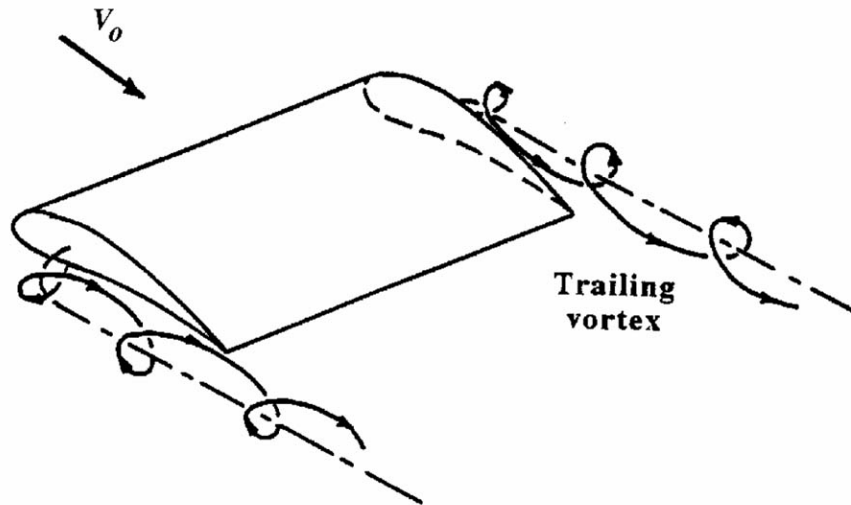


## 36.4 Lift and Drag of Wings

All wings have a finite span,  $b$ , with two wing tips. The flow near the tips can greatly influence the flow characteristics over the entire wing. Hence, a wing has different lift and drag coefficients than those for the corresponding airfoil. That is, the lift and drag coefficients for a wing are a function of the aspect ratio,  $AR = b^2/S$ . For a wing of rectangular planform (i.e., constant chord), the aspect ratio is simply  $b/c$ .

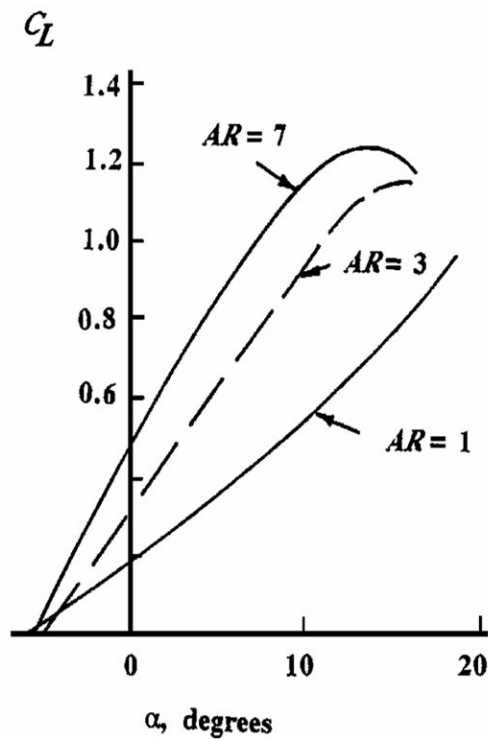
Because of the pressure difference between the lower and upper surfaces of a wing, the air tends to "leak" around the wing tips (bottom to top) and produce a swirling flow—the trailing or wing tip vortices shown in Fig. 36.8. This swirl interacts with the flow over the entire length of the wing, thereby affecting its lift and drag. The trailing vortices create a flow that makes it appear as though the wing were flying at an angle of attack different from the actual angle. This effect produces additional drag termed the *induced drag*.

**Figure 36.8** Trailing vortex

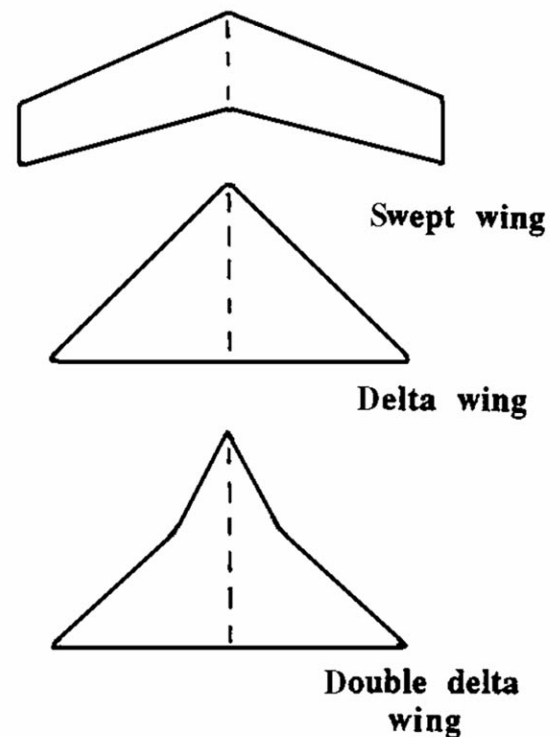


As shown by theory and experiment [Anderson, 1991], the larger the aspect ratio is, the larger the lift coefficient is and the smaller the drag coefficient is (Fig. 36.9). A wing with an infinite aspect ratio would have the properties of its airfoil section. Very efficient flyers (i.e., soaring birds and sailplanes) have long, slender (large  $AR$ ) wings.

**Figure 36.9** Lift coefficient as a function of aspect ratio.



**Figure 36.10** Various wing planforms.



For many wings the chord length decreases from the wing root (next to the aircraft body) to the wing tip. In addition, the shape of the airfoil section may change from root to tip, as may the local angle of attack (i.e., the wing may have some "twist" to it).

Many modern high-speed wings are swept back with a V-shaped planform; others are delta wings with a triangular planform (Fig. 36.10). Such designs take advantage of characteristics associated with high-speed compressible flow [Anderson, 1990].

Although wind tunnel tests of model airfoils and wings still provide valuable (and sometimes unexpected) information, modern computational fluid dynamic (CFD) techniques are widely used. Techniques involving paneling methods, finite elements, boundary elements, finite differences, and viscous-inviscid interaction are among the powerful tools currently available to the aerodynamicist [Moran, 1984].

## Defining Terms

**Airfoil:** The cross section of a wing, front to back.

**Angle of attack:** The angle between a line connecting the leading and trailing edges of an airfoil and the free stream velocity.

**Bernoulli's principle:** Conservation of energy principle that states that an increase in flow speed is accompanied by a decrease in pressure and vice-versa.

**Camber:** Maximum distance between the chord line and the camber line.

**Center of pressure:** Point of application of the lift and drag forces.

**Chord:** Distance between the leading and trailing edges of an airfoil.

**Dynamic pressure:** Pressure increase resulting from the conversion of kinetic energy into pressure.

**Flaps:** Leading and trailing edge devices used to modify the geometry of an airfoil.

**Lift and drag coefficients:** Lift and drag made dimensionless by dynamic pressure and wing area.

**Moment coefficient:** Pitching moment made dimensionless by dynamic pressure, wing area, and chord length.

**Planform:** Shape of a wing as viewed from directly above it.

**Span:** Distance between the tips of a wing.

**Stall:** Sudden decrease in lift as angle of attack is increased to the point where flow separation occurs.

## References

- Anderson, J. D. 1990. *Modern Compressible Flow with Historical Perspective*, 2nd ed. McGraw-Hill, New York.
- Anderson, J. D. 1991. *Fundamentals of Aerodynamics*, 2nd ed. McGraw-Hill, New York.
- Hubin, W. N. 1992. *The Science of Flight: Pilot Oriented Aerodynamics*. Iowa State University Press, Ames, IA.
- Moran, J. 1984. *An Introduction to Theoretical and Computational Aerodynamics*. John Wiley & Sons, New York.
- Schlichting, H. 1979. *Boundary Layer Theory*, 7th ed. McGraw-Hill, New York.

## Further Information

- Abbott, I. H. and van Doenhoff, A. E. 1949. *Theory of Wing Sections*. McGraw-Hill, New York.
- Anderson, D. A., Tannehill, J. C., and Pletcher, R. H. 1984. *Computational Fluid Mechanics and Heat Transfer*. Hemisphere, New York.
- Anderson, J. D. 1985. *Introduction to Flight*, 2nd ed. McGraw-Hill, New York.

Braun, E. R., Wang, P. L. "Boundary Layers"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

- 37.1 Theoretical Boundary Layers
- 37.2 Reynolds Similarity in Test Data
- 37.3 Friction in Pipes
- 37.4 Noncircular Channel
- 37.5 Example Solutions

**Edwin R. Braun**

*University of North Carolina, Charlotte*

**Pao-lien Wang**

*University of North Carolina, Charlotte*

## 37.1 Theoretical Boundary Layers

---

In a simple model of a solid, material deformation is proportional to the strain. In a simple model of a fluid, the deformation is proportional to the rate of strain or the change in velocity over a small distance. The mathematical term describing this phenomenon is the last term in the following boundary layer equation:

$$\rho \left[ \frac{\delta u}{\delta t} + u \frac{\delta u}{\delta x} + \nu \frac{\delta u}{\delta y} \right] = -\frac{\delta p}{\delta x} + \frac{\delta}{\delta y} \left( \mu \frac{\delta u}{\delta y} \right) \quad (37.1)$$

where  $u$  is the velocity in the  $x$  direction as a function of  $x$ ,  $y$ , and  $t$ . The values  $\rho$  and  $\mu$  are the density and dynamic viscosity for the fluid, respectively. This equation is good for all situations with no pressure ( $P$ ) change present in the direction normal to the wall.

The left-hand side of Eq. (37.1) represents time kinetic energy in flow. The pressure term is a potential energy term. The rate of strain term that represents this energy dissipates through viscous losses. When the dissipation term is significant compared to the others a boundary layer must be considered as part of the flow analysis.

For a straight-channel, steady (not time-dependent) flow, Eq. (37.1) becomes

$$\mu \frac{d^2 u}{dy^2} = \frac{dp}{dx} \quad (37.2)$$

which has the solution

$$u = -\frac{1}{2\mu} \frac{dp}{dx} (D^2 - y^2) \quad (37.3)$$

where  $d$  is the distance toward the wall measured from the centerline and the velocity  $u$  is zero at the walls. This velocity equation is a parabola. Note that when  $y = D$  Eq. (37.3) becomes

$$u_{cL} = -\frac{D^2}{2\mu} \frac{dp}{dx} \quad (37.4)$$

Thus, if the pressure loss over a distance  $X$  is measured along with the centerline velocity ( $u_{cL}$ ), the viscosity can be determined. Similarly, if the velocity is known at the centerline, the pressure loss per unit length can be calculated.

## 37.2 Reynolds Similarity in Test Data

---

As a boundary layer develops, it starts in a smooth, or laminar, state. Downstream, it transforms into a turbulent state, where the flow is irregular and contains eddies. Various physical conditions will affect the speed of this transition, like wall or surface roughness or upstream turbulence. In smooth walled pipes, laminar flow occurs for Reynolds numbers (Re) of less than 2000, with fully developed turbulence for Re greater than 4000. The Reynolds number is a dimensionless number developed from dynamic similarity principles that represents the ratio of the magnitudes of the inertia forces to the friction forces in the fluid.

$$\text{Re} = \frac{\text{inertia force}}{\text{friction force}}$$

where inertia force  $= \rho V_c^2 L_c^2$  and friction force  $= \mu V_c L_c^2$ . Then,

$$\text{Re} = \frac{\rho V_c L_c}{\mu} = \frac{V_c L_c}{\nu} \quad (37.5)$$

where  $V_c$  and  $L_c$  are characteristic or representative velocities and lengths, respectively. For a pipe or similar narrow channel,  $L_c$  is the internal diameter (ID) of the pipe, and  $V_c$  is the average or bulk velocity obtained by dividing the mass flow rate ( $M$ ) by the cross-sectional area and density of the fluid:

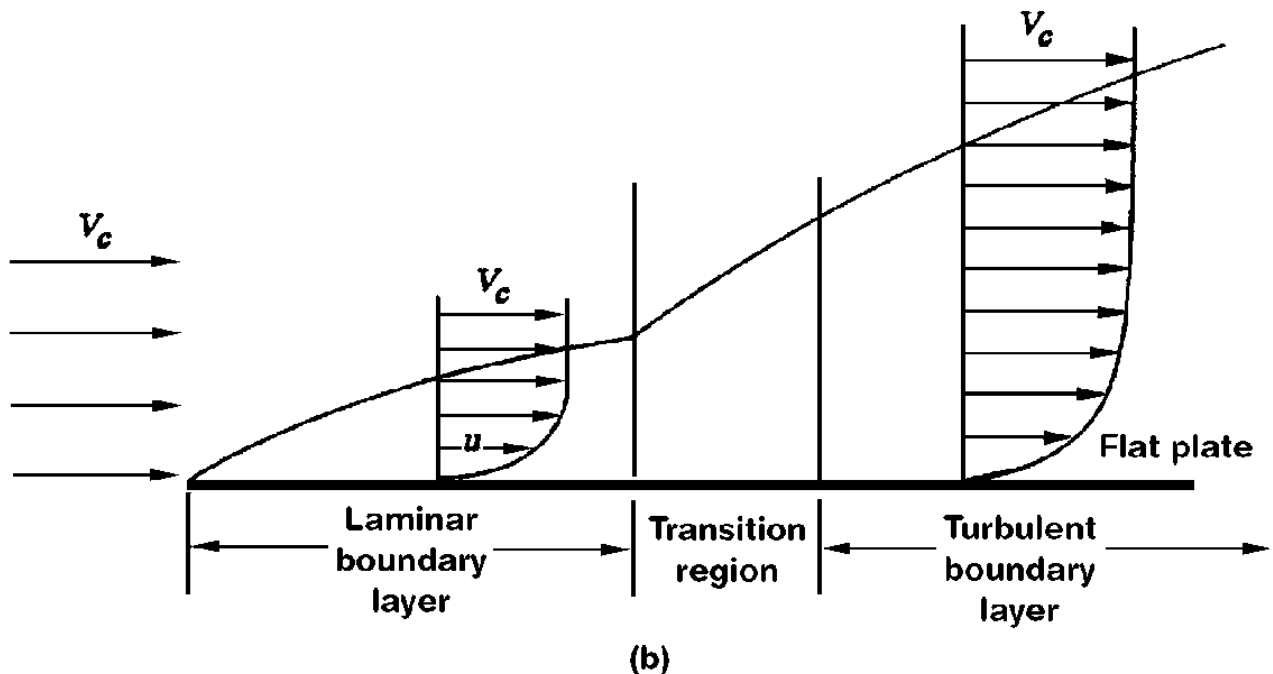
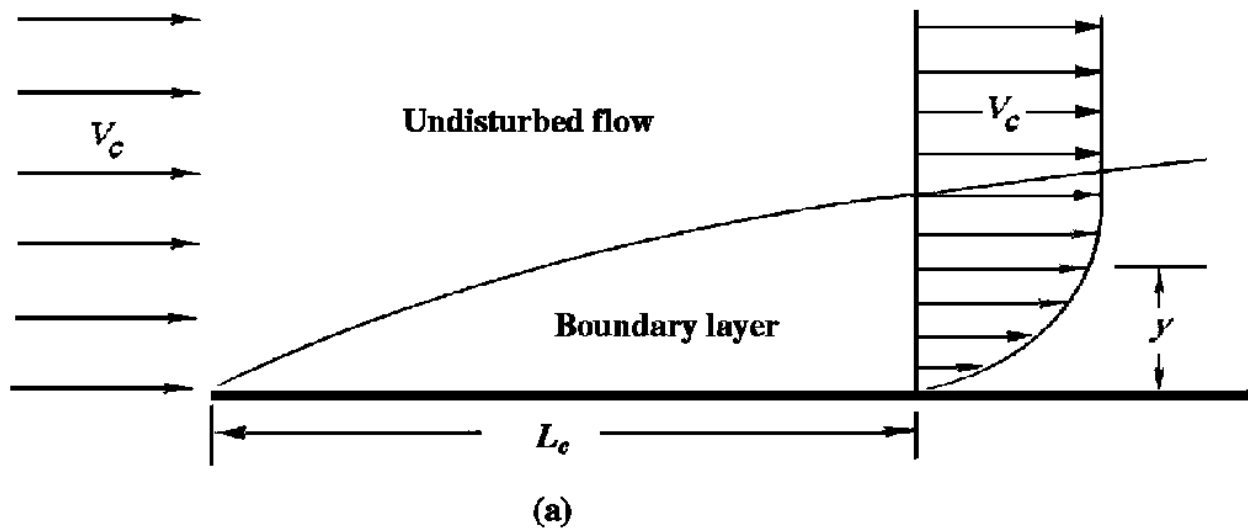
$$V_c = \frac{M}{\rho A} \quad (37.6)$$



Using the Reynolds number as a similarity parameter, test data can be correlated into generalized charts for frictional losses.

For the flat plate (Fig. 37.1) case,  $V_c$  is taken as the free stream velocity outside the boundary layer, and  $L_c$  is the length measured along the wall standing from the leading edge.

**Figure 37.1** (a) Boundary layer along a smooth plane. (b) Laminar and turbulent boundary layers along a smooth, flat plate. (Vertical scales greatly enlarged.)



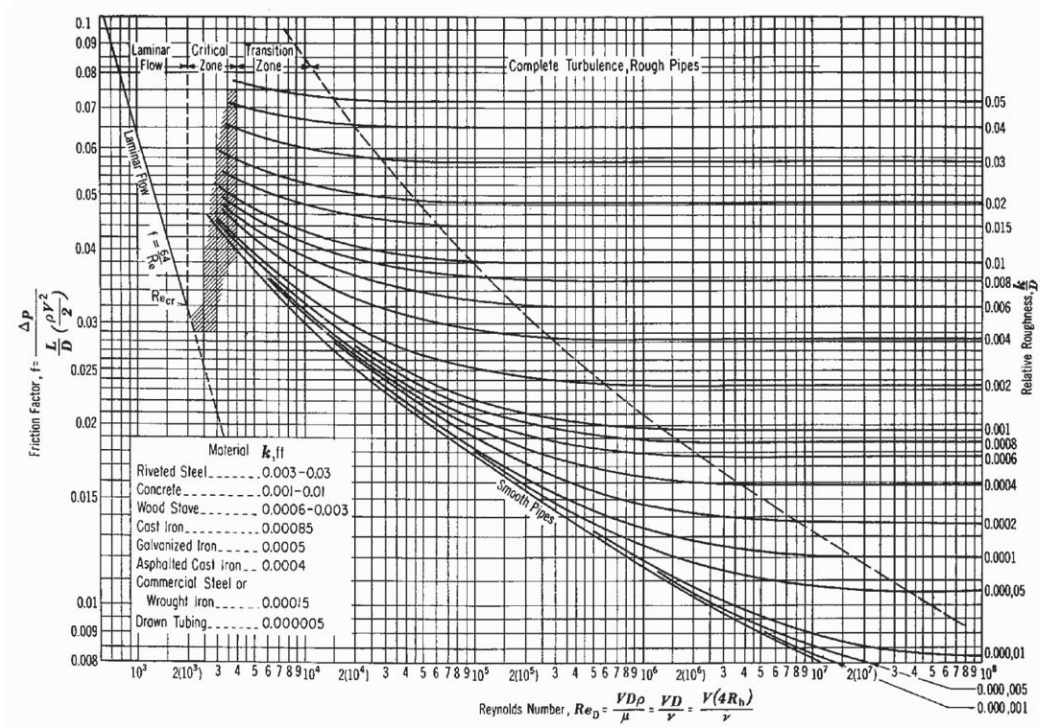
# 37.3 Friction in Pipes

The energy equation for steady flow between any two points in a pipe can be written as

$$\frac{V_2^2 - V_1^2}{2g} + \frac{P_2 - P_1}{\rho g} + Z_2 - Z_1 - h_f = 0 \tag{37.7}$$

where  $h_f$  is a head loss due to friction. This equation neglects other minor losses (such as elbows, valves, exit and entrance losses, and bends). It is useful to define the head loss in terms of a friction factor ( $f$ ) such that this nondimensional friction factor ( $f$ ), known as the Darcy friction factor, can be determined experimentally as a function of the dimensionless Reynolds numbers and a relative roughness parameter  $\varepsilon/D$ , as shown in Fig. 37.2. Rough factors,  $\varepsilon$ , are given in Table 37.1.

**Figure 37.2** Friction factors for commercial pipe. (Source: Moody, L. F. 1944. Friction factors for pipe flow. *Trans. ASME*. 66:672. With permission.)



**Table 37.1** Roughness

Surface	$\varepsilon$ , ft	$\varepsilon$ , m
Glass, plastic	smooth	smooth
Drawn tubing	$5 \cdot 10^{-6}$	$1.5 \cdot 10^{-6}$
Commercial steel, wrought iron or aluminum sheet	$1.5 \cdot 10^{-4}$	$4.6 \cdot 10^{-5}$
Galvanized iron	$5 \cdot 10^{-4}$	$1.2 \cdot 10^{-4}$
Cast iron	$8.5 \cdot 10^{-4}$	$2.4 \cdot 10^{-4}$
Concrete pipe	$4 \cdot 10^{-3}$	$1.2 \cdot 10^{-3}$
Riveted steel pipe	.01	.003
Wood	.001	.0003

There are two equations that describe the data shown in Fig. 37.2. The first is the laminar line. For laminar flow in pipes with  $Re$  less than 2000, it can be shown through analysis that

$$f = \frac{64}{Re} \quad (37.8)$$

The second is the Colebrook equation [Colebrook, 1938]:

$$\frac{1}{\sqrt{f}} = -2 \log_{10} \left[ \frac{\varepsilon/D}{3.7} + \frac{2.51}{Re \sqrt{f}} \right] \quad (37.9)$$

which describes the turbulent region. Note that, as the roughness  $\varepsilon$  approaches zero, we obtain the smooth pipeline and the equation becomes

$$\frac{1}{\sqrt{f}} = 2 \log_{10} \left[ \frac{Re \sqrt{f}}{2.51} \right] \quad (37.10)$$

For fully developed turbulence the  $Re$  approaches zero and the Colebrook equation simplifies to

$$\frac{1}{\sqrt{f}} = 2 \log_{10} \left[ \frac{3.7}{\varepsilon/D} \right] \quad (37.11)$$

For turbulent flows in closed conduits with noncircular cross sections, a modified form of Darcy's equation may be used to evaluate the friction loss:

$$h_f = f \left( \frac{L}{D} \right) \left( \frac{v^2}{2g} \right)$$

where  $D$  is the diameter of the circular conduit.

## 37.4 Noncircular Channel

In the case of noncircular cross sections, a new term,  $R$ , is introduced to replace diameter  $D$ .  $R$  is defined as hydraulic radius, which is the ratio of the cross-sectional area to the wetted perimeter (WP) of the noncircular flow section.

$$R = \frac{A}{WP}$$

For a circular pipe of diameter  $D$  the hydraulic radius  $R$  is

$$R = \frac{A}{WP} = \frac{\pi D^2/4}{\pi D} = \frac{D}{4}$$

or  $D = 4R$ . Substitution of  $4R$  for  $D$  in Darcy's equation yields

$$h_f = f \left( \frac{L}{4R} \right) \left( \frac{v^2}{2g} \right)$$

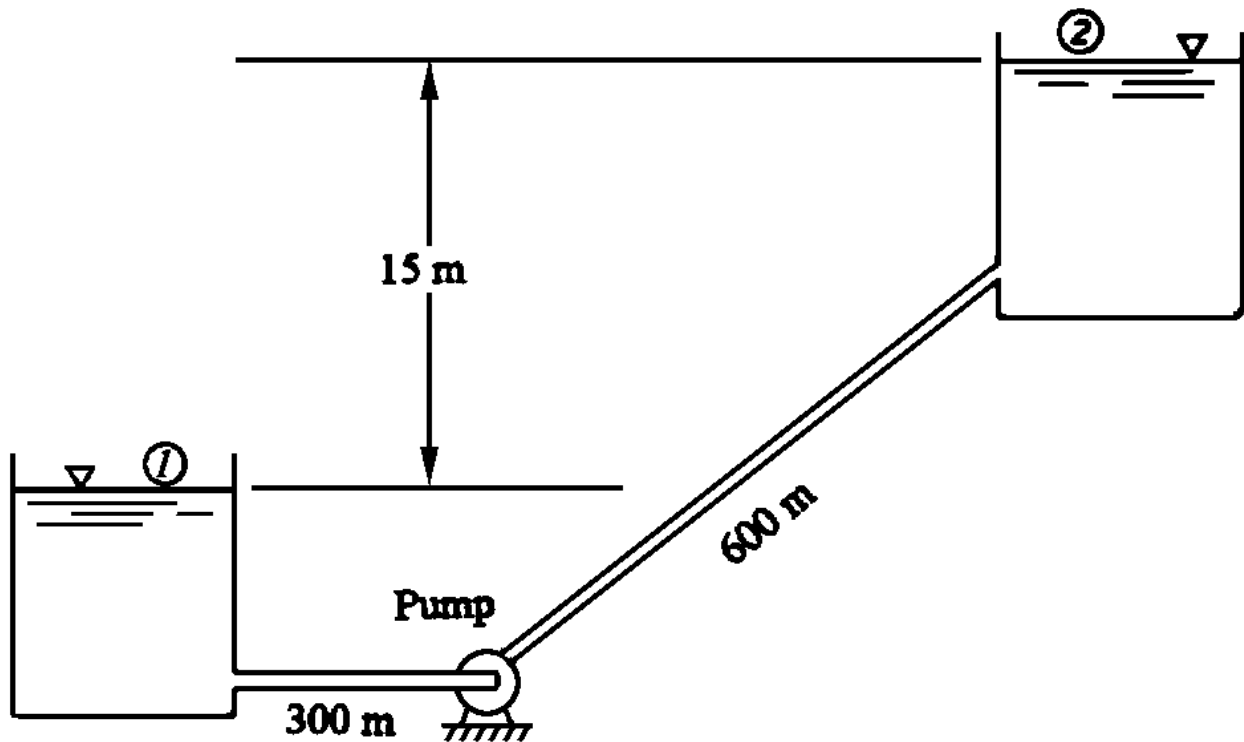
The Reynolds number can be modified as

$$\text{Re} = \frac{v(4R)\rho}{\mu} \quad \text{or} \quad \text{Re} = \frac{v(4R)}{\nu}$$

## 37.5 Example Solutions

**Example 37.1.** Refer to Fig. 37.3. Water at 50° C is flowing at a rate of 0.07 m<sup>3</sup>/s . The pipeline is steel and has an inside diameter of 0.19 m. The length of the pipeline is 900 m. Assume the kinematic viscosity ( $\nu$ ) is  $5.48 \cdot 10^{-7}$  m<sup>2</sup>/s . Find the power input to the pump if its efficiency is 82%; neglect minor losses.

**Figure 37.3** Pipeline in Example 37.1.



Given information is as follows:

$$Q = 0.07 \text{ m}^3; \quad L = 900 \text{ m}; \quad T = 50^\circ\text{C}$$

$$\Delta Z = 15 \text{ m}; \quad \nu = 5.48 \cdot 10^{-7} \text{ m}^2/\text{s}; \quad D = 0.2 \text{ m}$$

$$\gamma \text{ at } 50^\circ\text{C} = 9.69 \text{ kN/m}^3$$

Find power input to pump.

**Solution.** First, determine the Reynolds number:

$$\text{Re} = \frac{vD}{\nu}; \quad v = \frac{Q}{A}$$

$$\begin{aligned} \text{Re} &= \frac{4Q}{\pi D \nu} = \frac{4(0.07)}{\pi(0.2)(5.48 \cdot 10^{-7})} \\ &= 8.13 \cdot 10^5 \end{aligned}$$

Second, determine  $\varepsilon/D$  ratio and friction factor  $f$ . Roughness ( $\varepsilon$ ) for steel pipe =  $4.6 \cdot 10^{-5} \text{ m}$ .

$$\frac{\varepsilon}{D} = \frac{4.6 \cdot 10^{-5} \text{ m}}{0.19 \text{ m}} = 0.000242$$

From Moody's diagram with values of  $N_R$  and  $\varepsilon/D$ ,  $f = 0.0151$ .

Next, determine head loss due to friction:

$$h_f = 0.0151 \left( \frac{L}{D} \right) \left( \frac{v^2}{2g} \right), \quad v = \frac{Q}{A}$$

$$\begin{aligned} h_f &= 0.0151 \frac{8LQ^2}{\pi^2 g D^5} \\ &= 0.0151 \frac{8(900)(0.07)^2}{\pi^2 (9.81)(0.2)^5} \\ &= 0.0151 \frac{35.28}{0.031} = 17.2 \text{ m} \end{aligned}$$

Finally, determine power input into pump:

$$P_A = h_A \gamma Q = 17.2 \text{ m} \left( 9.69 \frac{\text{kN}}{\text{m}^3} \right) \left( 0.07 \frac{\text{m}^3}{\text{s}} \right)$$

$$= 11.67 \frac{\text{kN} \cdot \text{m}}{\text{s}} = 11.67 \text{ kW}$$

$$e_p = \frac{P_A}{P_I}$$

$e_p$  = Pump efficiency

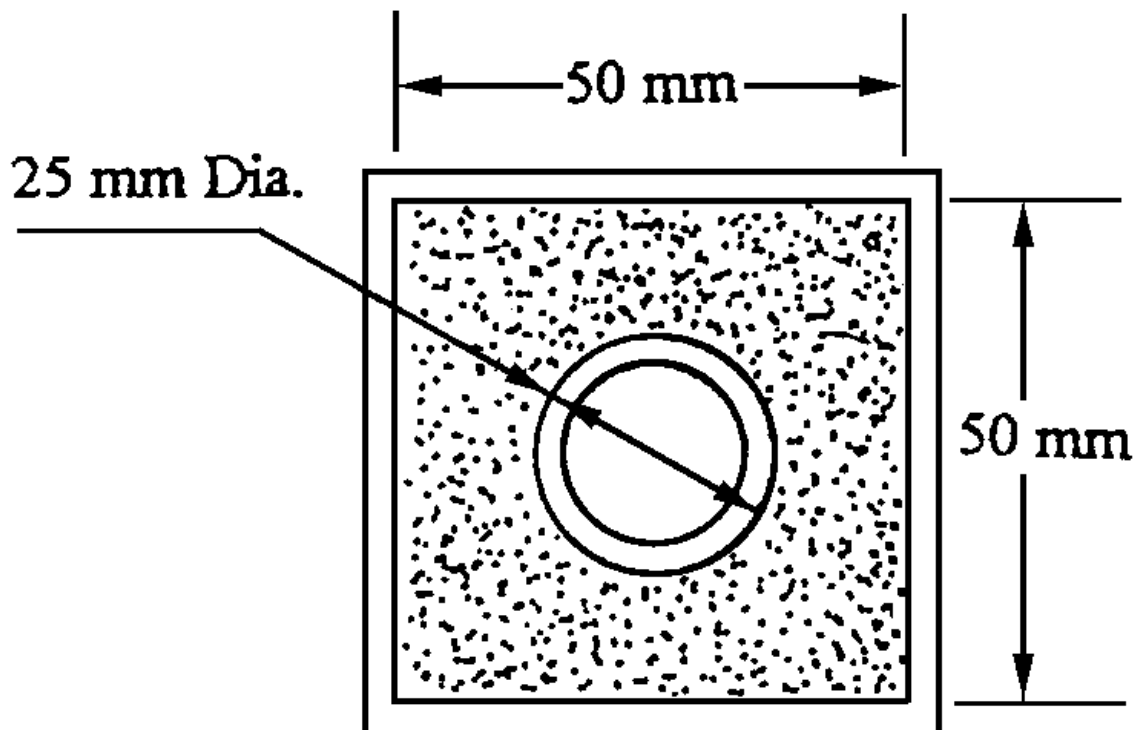
$P_A$  = Power delivered to fluid

$P_I$  = Power input into pump

$$P_I = \frac{P_A}{e_p} = \frac{11.67 \text{ kW}}{0.82} = 14.23 \text{ kW}$$

**Example 37.2.** Air with a specific weight of  $12.5 \text{ N/m}^3$  and dynamic viscosity of  $2.0 \cdot 10^{-5} \text{ N} \cdot \text{s/m}^2$  flows through the shaded portion of the duct shown in Fig. 37.4 at the rate of  $0.04 \text{ m}^3/\text{s}$ . (See Table 37.2 or Fig. 37.5 for dynamic viscosities of some common liquids.) Calculate the Reynolds number of the flow, given that  $\gamma = 12.5 \text{ N/m}^3$ ,  $\mu = 2.0 \cdot 10^{-5} \text{ N} \cdot \text{s/m}^2$ ,  $Q = 0.04 \text{ m}^3/\text{s}$ , and  $L = 30 \text{ m}$ .

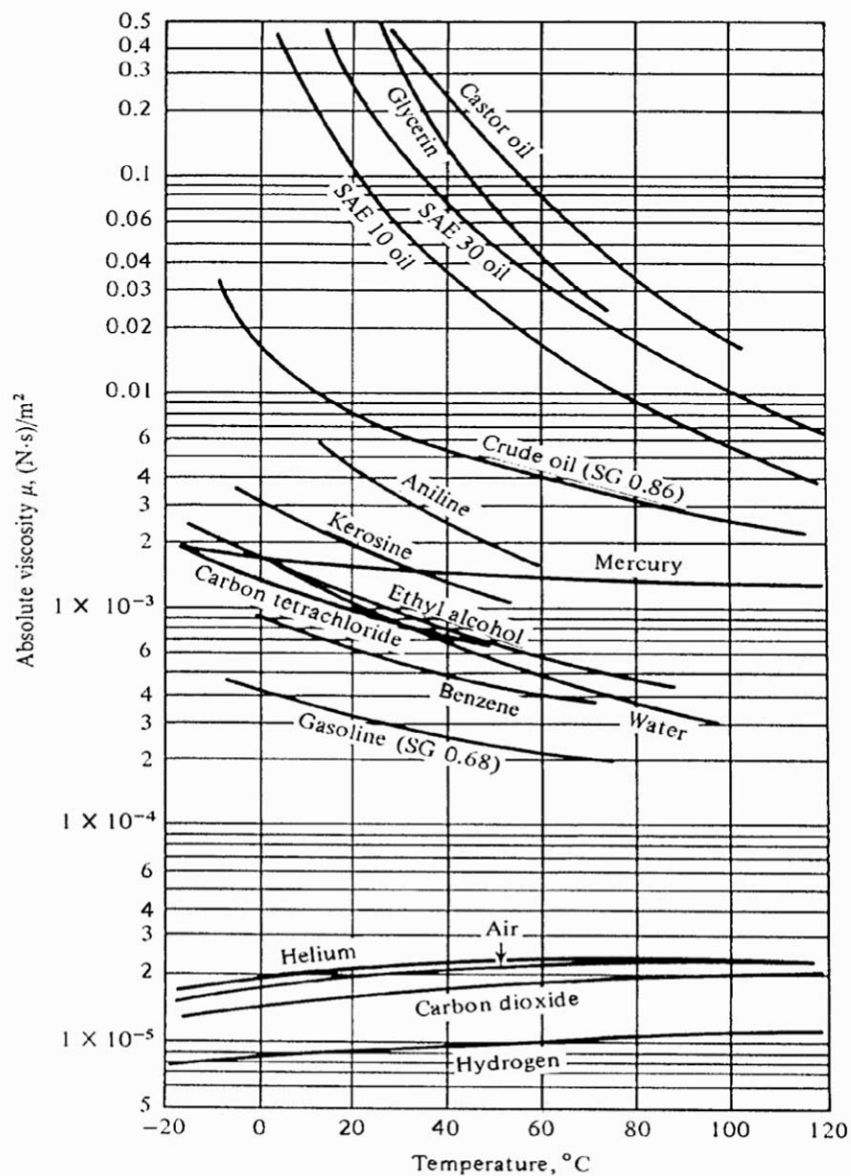
**Figure 37.4** Duct in Example 37.2.



**Table 37.2** Dynamic Viscosity of Liquids ( $\mu$ ) (mPa · s)

Liquid	-25°C	0°C	25°C	50°C	75°C	100°C
Water		1.793	0.890	0.547	0.378	
Mercury			1.526	1.402	1.312	
Methanol	1.258	0.793	0.544			
Isobutyl acetate			0.676	0.493	0.370	0.286
Toluene	1.165	0.778	0.560	0.424	0.333	0.270
Styrene		1.050	0.695	0.507	0.390	0.310
Acetic acid			1.056	0.786	0.599	0.464
Ethanol	3.262	1.786	1.074	0.694	0.476	
Ethylene glycol			16.1	6.554	3.340	1.975

**Figure 37.5** Absolute viscosity of common fluids at 1 atm. (Source: White, F. 1986. *Fluid Mechanics*, 2nd ed. McGraw-Hill, New York. With permission.)



***Solution***

$$\rho = \gamma/g = \frac{12.5 \text{ N/m}^3}{9.81 \text{ m/s}^2} = 1.27 \text{ N} \cdot \text{s}^2/\text{m}^4 \text{ or } \text{kg/m}^3$$

$$\begin{aligned} A \text{ (shaded)} &= (0.05 \text{ m})^2 - \frac{\pi}{4}(0.025 \text{ m})^2 \\ &= 0.0025 \text{ m}^2 - 0.00049 \text{ m}^2 \\ &= 0.002 \text{ m}^2 \end{aligned}$$

$$\begin{aligned} \text{Wet parameter (WP)} &= 4(0.05 \text{ m}) + \pi(0.025 \text{ m}) \\ &= 0.2 \text{ m} + 0.0785 \text{ m} \\ &= 0.279 \text{ m} \end{aligned}$$

$$\begin{aligned} \text{Hydraulic radius (} R \text{)} &= \frac{A}{\text{WP}} \\ &= \frac{0.002 \text{ m}^2}{0.279 \text{ m}} \\ &= 0.00717 \text{ m} \end{aligned}$$

$$v = \frac{Q}{A} = \frac{0.04 \text{ m}^3/\text{s}}{0.002 \text{ m}^2} = 20 \text{ m/s}$$

$$\text{Reynolds number (Re)} = \frac{4Rv\rho}{\mu}$$

$$\begin{aligned} N_R &= \frac{4(0.00717 \text{ m})(20 \text{ m/s})(1.27 \text{ N} \cdot \text{s}/\text{m}^4)}{2.0 \cdot 10^{-5} \text{ N} \cdot \text{s}/\text{m}^2} \\ &= 3.64 \cdot 10^4 \end{aligned}$$



## References

- Colebrook, C. F. 1938. Turbulent flow in pipes with particular reference to the transition points between smooth and rough laws. *ICF Journal*. 2:133–156.
- Moody, L. F. 1944. Friction factors for pipe flow. *Trans. ASME*. 66:672.
- White, F. 1986. *Fluid Mechanics*, 2nd ed. McGraw-Hill, New York.

Tullis, J. P. "Valves"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 38

## Valves

---

### 38.1 Control Valves

Valve Types • Valve Selection

### 38.2 Air Valves

### 38.3 Check Valves

#### **J. Paul Tullis**

*Utah State University, Logan*

Valves are mechanical devices that are installed in pipelines to control flow or pressure. The valves provide control using a variable restriction created by a rotating plug or disk, a sliding sleeve or gate, or the pinching of a flexible membrane. They are an important part of piping systems and care should be taken to ensure that the correct valve is selected. If not properly selected and operated, valves can cause operational problems, including poor control, cavitation, and hydraulic transients that result in problems such as poor performance, accelerated wear, repair, and replacement of the valve.

The primary valve types, classified by their function, are control valves, isolation (block) valves, air release/vacuum breaking valves, and check valves. Control valves can be used to control flow, pressure, liquid level, cavitation, and pressure transients. Isolation valves are frequently placed on each side of control valves and pumps, allowing them to be removed for repair or replacement. Air valves are designed to expel large amounts of air at low pressure during filling and release small amounts of pressurized air during operation. Vacuum relief valves admit air to the pipe while the pipe is being drained to prevent excessive vacuum pressures and reduce the possibility of collapsing thin-walled pipes. Check valves are used to prevent reverse flow.

## **38.1 Control Valves**

---

Selection criteria for control valves should include flow and pressure drop characteristics, torque or thrust requirements, cavitation performance, and a consideration of transients generated by valve operation. Other factors that influence valve performance—such as valve installation (direction and orientation) and the effect of upstream disturbances located close to the valve—should also be considered.

### **Valve Types**

*Butterfly valves* are made with varying disk and seating designs. However, the flow characteristics

of the various valve designs are fairly similar. Regardless of the disk design, all butterfly valves have two crescent-shaped openings that discharge parallel to the wall of the downstream pipe. These valves are a popular style and are relatively inexpensive.

*Cone valves* have a rotating conical plug that presents no obstruction to the flow when the plug is fully open. In a partially open position, these valves have two crescent-shaped throttling ports in series. One style of cone valve uses a solid conical plug for small valves and a fabricated plug with a conical skirt wrapped around the plug for large valves. Another style has a fabricated plug design that allows some of the water to flow around the plug. This slightly increases capacity, but adversely affects its cavitation performance.

*Ball and plug valves* have numerous designs. One style of ball valves has a spherical plug with a cylindrical hole drilled through to form the flow passage. For full-ported designs, the flow passage is the same diameter as the inside pipe diameter. For others the ball is smaller than the pipe diameter and the valve body includes reducing and expanding sections so the inlet and outlet are the same size as the pipe. Both of these designs have two throttling ports in series and have flow characteristics similar to the cone valve. Close machining tolerances are required for the seating surfaces because the spherically shaped plug rotates into the seat. Another design uses a fabricated plug essentially made of two intersecting pipes, one closed and one open. The eccentric plug valve uses a segment of the plug that looks much like the visor on a helmet. This type of valve has only one throttling port.

*Gate valves* have a variety of minor differences that influence their seating performance, but their flow characteristics are similar. All have one crescent-shaped flow opening.

*Sleeve valves* are a relatively new valve style and are used primarily to control cavitation. One form of an in-line sleeve valve consists of a stationary sleeve with numerous holes and a traveling sleeve. The size and spacing of the holes can be varied to provide a variety of flow capabilities. Normally, the jets discharge inward, forcing the cavitation to be concentrated at the center of the discharge pipe, away from the boundaries. The jets can also discharge outward into a pipe or tank.

*Free-discharge sleeve valves* (Howell-Bunger) have a stationary cone, a traveling sleeve, and usually a hood to direct the flow. These valves are used for free-discharge releases at high pressure. Since the high-velocity jet is fully aerated, cavitation is avoided. A dissipation pool is needed to prevent erosion by the jet.

*Globe valves* vary in design more than the other types. There are far too many variations to discuss here in any detail. The conventional type has a single port with a plug that moves linearly. The flow is controlled by the size of the annular opening between the plug and seat. Body styles of globe valves vary significantly. Recent developments include a variety of cavitation trims. One type uses a perforated cylinder, similar to that of the sleeve valve. Cavitation is suppressed but flow capacity is reduced. Other styles of cavitation control trim use both parallel and series flow passages. These are frequently called *stack valves*. These flow passages can be formed by concentric perforated cylinders or with stacks of flat disks with numerous flow passages in parallel containing numerous turns (tortuous paths) or sudden expansions in series. These designs allow very high pressure drops and still control the cavitation to an acceptable level.

## Valve Selection

Selecting the proper flow control valve requires information on maximum and minimum flows and the pressure drop requirements for both the present and projected demands. The following criteria have been suggested for selecting flow control valves based on hydraulic performance [Tullis, 1989]:

1. The valve should not produce excessive pressure drop when full open.
2. The valve should control over at least 50% of its movement; that is, when the valve is closed 50%, the flow should be reduced at least 10%.
3. The maximum flow must be limited so the operating torque does not exceed the capacity of the operator or valve shaft and connections.
4. The valve should not be subjected to excessive cavitation.
5. Pressure transients should not exceed the safe limits of the system. This requires that the valve is sized so it controls the flow over most of its stroke and that the closure speed is controlled to limit the transients.
6. Some valves should not be operated at small openings due to potential seat damage caused by high jet velocity and cavitation.
7. Some valves should not be operated near full open, where they may have poor flow control and/or experience torque reversals leading to fatigue failures.

## Flow Characteristics

The relationship between flow and pressure drop at any valve opening can be expressed by any number of coefficients. The coefficients most commonly used are as follows:

*Discharge coefficient*

$$c_d = \frac{v}{\sqrt{2\Delta P/\rho + v^2}} \quad (38.1)$$

*Loss coefficient*

$$K = \frac{2\Delta P}{\rho v^2} \quad (38.2)$$

*Flow coefficient*

$$c_v = \frac{Q}{\sqrt{\Delta P/SG}} \quad (38.3)$$

*Free-discharge coefficient*

$$c_{df} = \frac{v}{\sqrt{2P_u/\rho + v^2}} \quad (38.4)$$

in which  $v$  is the average velocity at the inlet to the valve,  $\rho$  is the fluid density (specific weight/gravity),  $\Delta P$  is the net pressure drop across the valve,  $SG$  is the specific gravity of the liquid, and  $P_u$  is the gage pressure at the inlet to the valve. In Eq. (38.3)  $\Delta P$  is in psi and  $Q$  is in U.S. gallons per minute (gpm). The other equations are dimensionless. Equation (38.4) can also be applied to predict flow capacity for a valve that is operating in choking cavitation

$$(P_u = P_{\text{absolute}} - P_{\text{vapor}}.)$$

When selecting a control valve, it is necessary to analyze its performance as a part of the piping system and not simply consider it an isolated device. The same valve installed in different systems will have different control characteristics. For example, most valves used to control flow in a short pipe—where friction and minor losses are small—will almost linearly reduce the flow as the valve closes. Installed in a high-loss system, the same valve will not reduce the flow until it has closed a significant percentage of its full stroke. This dependence also influences cavitation and transient problems. [For details, see [Tullis \(1989, ch. 4\)](#).]

## Operating Forces

Proper valve selection includes selecting the operator so the valve can be opened and closed under the most severe flow conditions expected. The forces are caused by friction and hydrodynamic forces. These forces must be known so that the operator is properly sized. The four primary forces that typically affect rotary-actuated valves are packing friction torque, seating torque, bearing friction torque, and hydrodynamic torque. Similar forces affect linear-actuated valves.

For small valves with resilient seats, the seating torque can be larger than all the other torques combined. Some seats and packings are adjustable and allow the friction torque to be changed.

Bearing friction torque is caused by the load placed on the bearing surfaces by pressure differentials across the valve. Since the pressure drop significantly increases as a valve closes, the bearing torque is greatest at small openings.

Hydrodynamic torque is caused by forces induced by the flowing water. The valve opening where maximum torque occurs is dependent on the valve and system. When sizing an operator, one must consider the full range of valve operation to determine the maximum torque (including friction and seating torque).

Some valves experience a torque reversal at large valve openings. Operating a valve near an opening where the torque reverses can result in fatigue failure of the shaft and loosening of connections. Another factor influencing the magnitude of the hydrodynamic torque is the presence of a disturbance located just upstream from the valve. Turbulence and nonuniform flow generated by the disturbance can increase both the average torque value and the magnitude of the torque fluctuation. This can cause fatigue failure of the shaft and connections.

## Cavitation

A valve exposed to excessive cavitation can experience accelerated wear, generate excessive noise

and vibrations, and even lose capacity. The acceptable cavitation level for a valve in a given system varies with valve type, valve function, details of the piping layout, and duration of operation. For example, a control valve required to operate continuously should operate at a light level of cavitation. In contrast, a valve intended for only intermittent use, such as a pressure relief valve, could be designed to operate at choking cavitation (flashing).

A thorough cavitation analysis is often an important part of valve selection. Detailed procedures for analyzing a valve for cavitation and making size and pressure scale effects adjustments are available [Tullis, 1989, 1993]. One of the most unfortunate misunderstandings about valve cavitation arises from the false teaching that cavitation begins when the mean downstream pressure drops to vapor pressure. This is in fact the final stage of cavitation—well beyond the point where damage can occur. Using choking as a design condition is appropriate for some applications, but not all. Numerous options are available for suppressing cavitation by selecting the correct type and size of valve, locating valves or orifices in series, using cavitation-resistant materials, and injecting air.

## 38.2 Air Valves

---

There are three types of automatic air valves: air-vacuum valves, air release valves, and combination valves. The air-vacuum valve generally has a large orifice to expel and admit large quantities of air at low pressure when filling or draining a line. These valves contain a float, which rises and seals as the valve body fills with water. Once the line is pressurized, the valve cannot reopen to remove air that may subsequently accumulate. If the pressure becomes negative during a transient or by draining, the float drops and admits air into the line.

Air release valves contain a small orifice and are intended to release small quantities of pressurized air that accumulate after initial filling. The small orifice is controlled by a plunger activated by a float at the end of a lever arm. As air accumulates in the valve body, the float drops and opens the orifice. As the air is expelled, the float rises and closes off the orifice. The combination valve has two orifices: a large one that functions as an air-vacuum valve and a small one that functions as an air release valve.

Air valves should be sized so that the air is expelled without pressurizing the pipe. Sizing charts are provided by manufacturers. Locating air valves depends primarily on the pipe profile. Preferably, the pipe should be laid to grade with valves placed at all high points or at regular intervals if there are no high points.

Velocity of the flow during filling is important. It should not be so high that pressure surges are generated but it should be high enough to move air to the air valves. It is best to flush trapped air without pressurizing the pipe. Allowing large quantities of air under high pressure to accumulate and move through the pipe can generate severe transients, especially if the compressed air is improperly released or allowed to pass through a control valve.

## 38.3 Check Valves

---

The basic function of a check valve is to allow forward flow under normal conditions and avoid flow reversal to prevent draining of the pipe and reverse rotation of the pump. The characteristics

of check valves that should be considered during valve selection include:

1. The flow required to fully open and firmly backseat the disk
2. The pressure drop at maximum flow
3. The stability of the disk at partial openings and sensitivity of the disk to upstream disturbances
4. Closure speed of check valves compared with the rate of flow reversal of the system
5. Sealing effectiveness and ease of maintenance

The most common type of check valve is the swing check valve. It has a simple design, low pressure drop, and reliable sealing. The swing check valve is relatively easy to repair, available in a wide range of sizes, and economical. It has a heavy disk suspended from a hinge pin located above the flow stream and relies on gravity for closure. The disk and hinge arm can be one or two pieces.

Tilt disk check valves have the hinge pin located in the flow stream, just above the centerline of the disk. They close significantly faster than swing check valves because of the shorter travel distance of the disk. Closer machining tolerances of the sealing surfaces are required than for a swing check because the disk rotates into the seat. Wear of the hinge pins or bushings can cause sealing problems.

The body and closing elements of lift check valves come in a variety of configurations. The disk (closing element) is frequently piston shaped and moves in guide surfaces. Some styles have a spherically shaped disk. The guide surface can be either vertical or inclined at an angle from vertical. They generally create a large pressure drop due to the configuration of the body. Closure speed can be controlled by the strength of the spring.

The double-door check valves have two disks that rely on a spring force for closure. Failure of the spring creates potentially serious operational problems.

Silent check valves have a spring-loaded, circular, flat disk oriented perpendicular to the flow, with the sealing surface around the outer diameter. The silent check valve closes rapidly because the disk is relatively light and the stroke is short.

The nozzle check valve is a streamlined, low-pressure-drop, rapidly closing check valve commonly used in European power plants and water supply systems. The seating surface is an annular ring machined into the valve body. The closing element is a lightweight, spring-loaded annular ring that moves parallel to the valve centerline. Because the seat diameter is large and the flow passage streamlined, only a short stroke is required to allow full flow with limited pressure drop. The short stroke combined with the light weight of the disk, plus the restoring force of the spring, results in very fast disk closure.

Power-assisted check valves can be divided into three categories: (1) valves with dashpots, (2) valves that can function either as a normal uncontrolled check valve or as a control valve in which a mechanical override can be used to limit the maximum disk opening and force the disk closed, and (3) valves that have operators to control both the opening and closing speeds. Power-assisted check valves are frequently activated by a signal that anticipates a transient. For example, a pump discharge valve can be programmed to automatically close at a predetermined rate when power to the pump is interrupted.



One of the most important considerations in check valve selection is the transient pressure rise generated at valve closure. The pressure rise is a function of how fast the valve disk closes and how fast the flow in the system reverses. Systems in which rapid flow reversals occur include parallel pumps and systems that have air chambers or surge tanks close to the check valve [Thorley, 1989].

Another important criterion is disk stability. It is preferable to size check valves so that the disk is fully open and firmly backseated at normal flow. This is especially important when the valve is located immediately downstream from a disturbance.

## References

- Thorley, A. R. D. 1989. Check valve behavior under transient flow conditions: a state-of-the-art review. Vol. III. *ASME*, June.
- Tullis, J. P. 1989. *Hydraulics of Pipelines—Pumps, Valves, Cavitation, Transients*. John Wiley & Sons, New York.
- Tullis, J. P. 1993. *Cavitation Guide for Control Valves*. NUREG/CR-6031, U.S. Nuclear Regulatory Commission, Washington, DC.

## Further Information

- BHRA. 1976. *Proceedings of the 2nd International Conference on Pressure Surges*. BHRA Fluid Engineering, London.
- Crane Co. 1979. *Flow of Fluids Through Valves, Fittings, and Pipe*. Technical Paper No. 410. Crane, New York.
- Driskell, L. 1983. *Control-Valve Selection and Sizing*. Instrument Society of America, Pittsburgh, PA.
- Fisher Controls International. 1977. *Control Valve Handbook*, 2nd ed. Fisher Controls, Marshalltown, IA.
- IAHR. 1992. *Hydraulic Transients with Water Column Separation*. 9th Round Table of the IAHR Group, Valencia, Spain. [There have been eight previous round tables, with the same publishers: 1971 in Milan, Italy; 1974 in Vallombrosa, Italy; 1976 in Royaumont, France; 1979 in Cagliari, Italy; 1981 in Obernach, former West Germany; 1983 in Gloucester, Great Britain; 1985 in Alvarleby, Sweden; and 1987 in Madeira.]
- ISA. 1976. *ISA Handbook of Control Valves*, 2nd ed. Instrument Society of America, Pittsburgh, PA.
- Kalsi Engineering and Tullis Engineering Consultants. 1993. *Application Guide for Check Valves in Nuclear Power Plants*. NP-5479, Rev. 1. Nuclear Maintenance Applications Center, Palo Alto, CA.
- Stone & Webster Engineering and Kalsi Engineering. 1990. *Guide for the Application and Use of Valves in Power Plant Systems*. NP-6516, Research Project 2233-5. Electric Power Research Institute, Palo Alto, CA.
- Wylie, E. B. and Streeter, V. L. 1993. *Fluid Transients in Systems*. Prentice Hall, Englewood Cliffs, NJ.

Boehm, R. F. "Pumps and Fans"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## 39.1 Pumps

Centrifugal and Other Velocity-Head Pumps • Positive-Displacement Pumps • Pump/Flow Considerations

## 39.2 Fans

### **Robert F. Boehm**

*University of Nevada, Las Vegas*

Pumps are devices that impart a pressure increase to a liquid. Fans are used to increase the velocity of a gas, but this is also accomplished through an increase in pressure. The pressure rise found in pumps can vary tremendously, which is a very important design parameter along with the liquid flow rate. This pressure rise can range from simply increasing the elevation of the liquid to increasing the pressure hundreds of atmospheres. Fan applications, on the other hand, deal with generally small pressure increases. In spite of this seemingly significant distinction between pumps and fans, there are many similarities in the fundamentals of certain types of these machines, as well as in their application and theory of operation.

The appropriate use of pumps and fans depends on the proper choice of device and the proper design and installation for the application. A check of sources of commercial equipment shows that many varieties of pumps and fans exist. Each of these has special characteristics that must be appreciated for achieving proper function. Preliminary design criteria for choosing between different types is given by Boehm [1987].

As one might expect, the wise application of pumps and fans requires knowledge of fluid flow fundamentals. Unless the fluid mechanics of a particular application is understood, the design could be less than desirable.

In this section, pump and fan types are briefly defined. In addition, typical application information is given. Also, some ideas from fluid mechanics that are especially relevant to pump and fan operation are reviewed. For more details on this latter topic, see the section of this book that discusses fluid mechanics fundamentals.

## 39.1 Pumps

---

The raising of water from wells and cisterns was the earliest form of pumping [a very detailed history of early applications is given by Ewbank (1842)]. Modern applications are much broader, and these find a wide variety of machines in use.

Modern pumps function on one of two principles. By far the majority of pump installations are of the *velocity-head* type. In these devices the pressure rise is achieved by giving the fluid a

movement. At the exit of the machine, this movement is translated into a pressure increase. The other major type of pump is called a *positive-displacement pump*. These devices are designed to increase the pressure on the liquid while essentially trying to compress the volume. A categorization of pump types has been given by Krutzsch [1986]; an adaptation of this categorization is shown below.

- I. Velocity head
  - A. Centrifugal
    - 1. Axial flow (single or multistage)
    - 2. Radial flow (single or double suction)
    - 3. Mixed flow (single or double suction)
    - 4. Peripheral (single or multistage)
  - B. Special effect
    - 1. Gas lift
    - 2. Jet
    - 3. Hydraulic ram
    - 4. Electromagnetic
- II. Positive displacement
  - A. Reciprocating
    - 1. Piston, plunger
      - a. Direct acting (simplex or duplex)
      - b. Power (single or double acting, simplex, duplex, triplex, multiplex)
    - 2. Diaphragm (mechanically or fluid driven, simplex or multiplex)
  - B. Rotary
    - 1. Single rotor (vane, piston, screw, flexible member, peristaltic)
    - 2. Multiple rotor (gear, lobe, screw, circumferential piston)

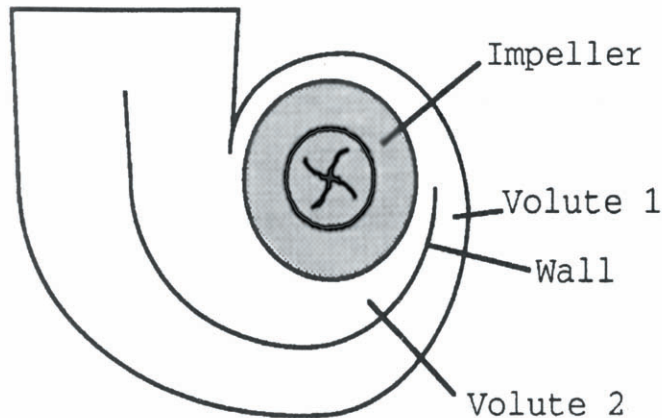
In the next section, some of the more common pumps are described.

## Centrifugal and Other Velocity-Head Pumps

Centrifugal pumps are used in more industrial applications than any other kind of pump. This is primarily because these pumps offer low initial and upkeep costs. Traditionally these pumps have been limited to low-pressure-head applications, but modern pump designs have overcome this problem unless very high pressures are required. Some of the other good characteristics of these types of devices include smooth (nonpulsating) flow and the ability to tolerate nonflow conditions.

The most important parts of the centrifugal pump are the impeller and volute. An *impeller* can take on many forms, ranging from essentially a spinning disk to designs with elaborate vanes. The latter is usual. Impeller design tends to be somewhat unique to each manufacturer, and a variety of designs are available for a variety of applications. An example of an impeller is shown in [Fig. 39.1](#). This device imparts a radial velocity to the fluid that has entered the pump perpendicularly to the impeller. The *volute* (there may be one or more) performs the function of slowing the fluid and increasing the pressure. A good discussion of centrifugal pumps is given by Lobanoff and Ross [1992].

**Figure 39.1** Schematic of a centrifugal pump. The liquid enters perpendicular to the figure, and a radial velocity is imparted by clockwise spin of the impeller.

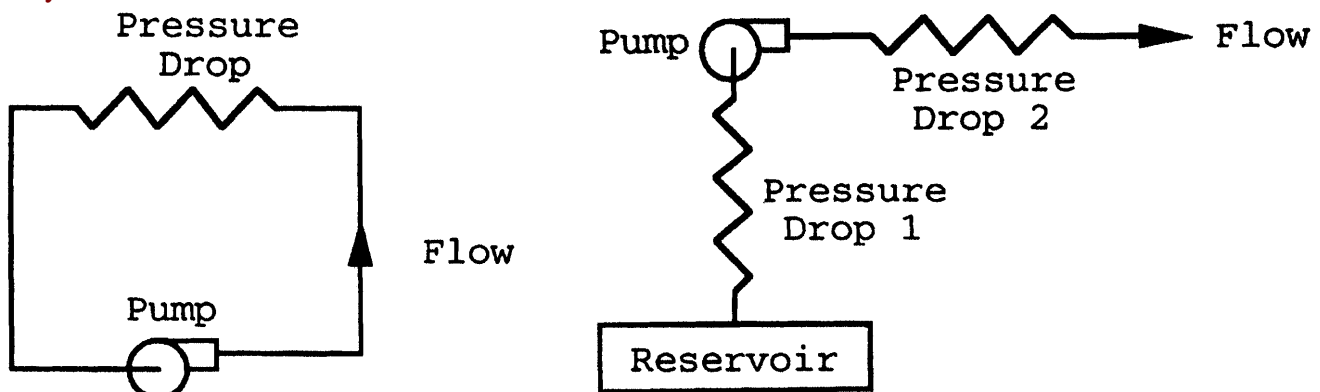


There are other types of velocity-head pumps. *Jet pumps* increase pressure by imparting momentum from a high-velocity liquid stream to a low-velocity or stagnant body of liquid. The resulting flow then goes through a diffuser to achieve an overall pressure increase. *Gas lifts* accomplish a pumping action by a drag on gas bubbles that rise through a liquid.

## Positive-Displacement Pumps

Positive-displacement pumps demonstrate high discharge pressures and low flow rates. Usually this is accomplished by some type of pulsating device. A piston pump is a classical example of a positive-displacement machine. The rotary pump is one type of positive displacement device that does not impart pulsations to the exiting flow [a full description of this type of pumps is given by Turton (1994)]. Several techniques are available for dealing with pulsating flows, including use of double-acting pumps (usually of the reciprocating type) and installation of pulsation dampeners. Positive-displacement pumps usually require special seals to contain the fluid. Costs are higher both initially and for maintenance, compared to most pumps that operate on the velocity-head basis. Positive-displacement pumps demonstrate an efficiency that is nearly independent of flow

**Figure 39.2** Typical pump applications, either in circuits or once-through arrangements, can be represented as combined fluid resistances, as shown. Resistances are determined from fluid mechanics analyses.

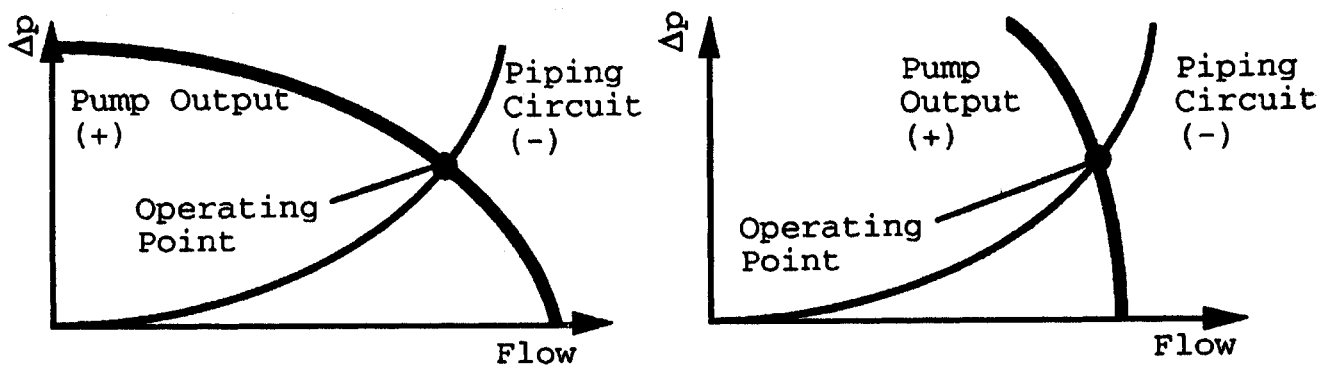


rate, in contrast to the velocity-head type. *Very high head pressures (often damaging to the pump) can be developed if the downstream flow is blocked.* For this reason a pressure-relief-valve bypass must always be used with positive-displacement pumps.

## Pump/Flow Considerations

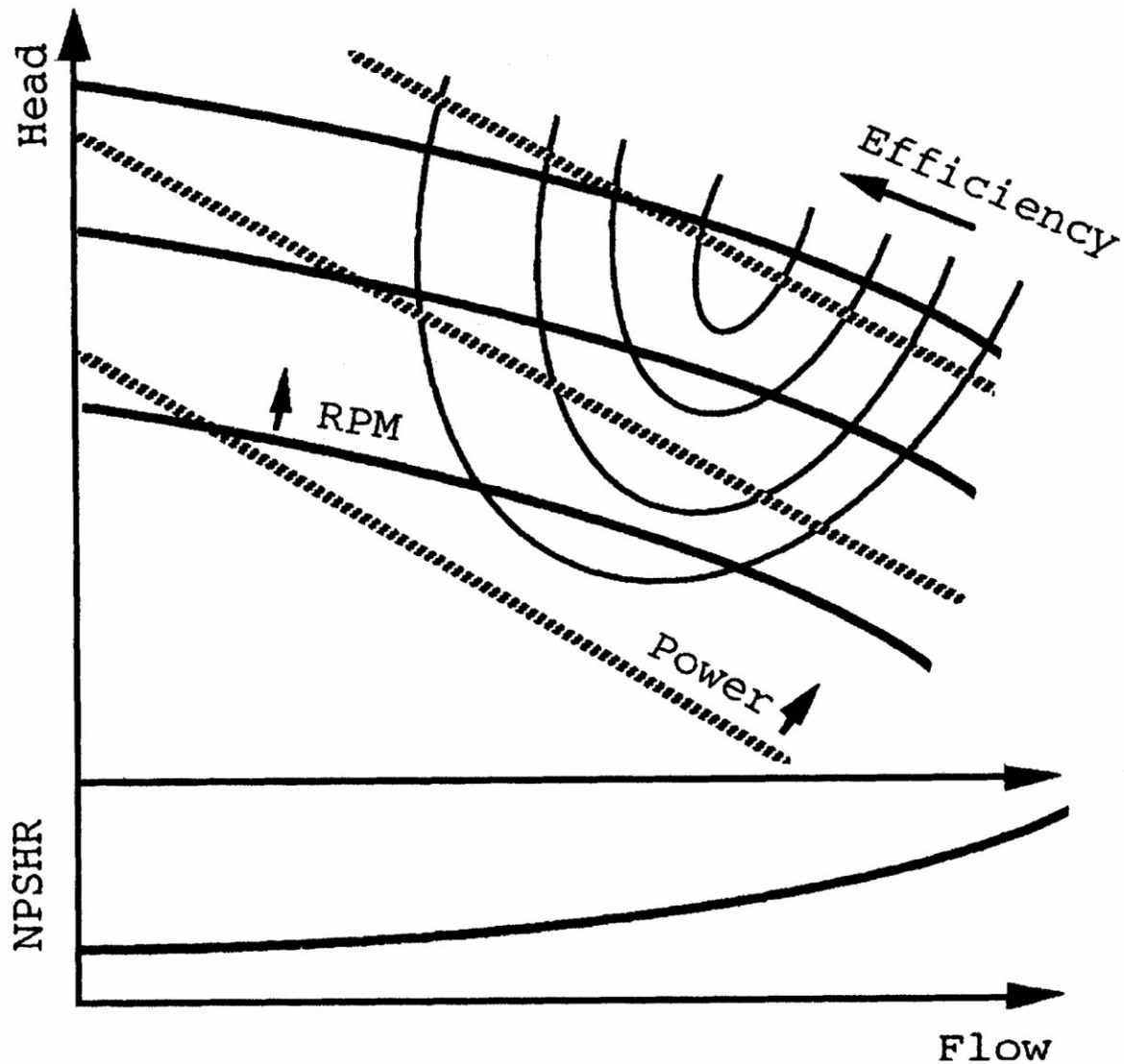
Performance characteristics of the pump must be considered in system design. Simple diagrams of pump applications are shown in Fig. 39.2. First consider the left-hand figure. This represents a flow circuit, and the pressure drops related to the piping, fittings, valves, and any other flow devices found in the circuit, estimated using laws of fluid mechanics. Usually these resistances (pressure drops) are found to vary approximately with the square of the liquid flow rate. Typical characteristics are shown in Fig. 39.3. Most pumps demonstrate a flow-versus-pressure rise variation that is a positive value at zero flow and decreases to zero at some larger flow. Positive-displacement pumps, as shown on the right-hand side of Fig. 39.3, are an exception to this rule in that these devices usually cannot tolerate a zero flow. An important aspect to note is that a closed system can presumably be pressurized. A contrasting situation and its implications are discussed as follows.

**Figure 39.3** Overlay of the pump flow versus head curve with the circuit piping characteristics gives the operating state of the circuit. A typical velocity-head pump characteristic is shown on the left, while a positive-displacement pump curve is shown on the right.



The piping diagram shown on the right-hand side of Fig. 39.2 is a once-through system, another frequently encountered installation. However, the leg of piping represented in "pressure drop 1" can have some very important implications related to **net positive suction head (NPSH)**. In simple terms, NPSH indicates the difference between the local pressure and the thermodynamic saturation pressure at the fluid temperature. If  $NPSH = 0$ , the liquid will vaporize, which can result in a variety of outcomes from noisy pump operation to outright failure of components. This condition is also called **cavitation**. If it occurs, cavitation will first take place at the lowest pressure point within the piping arrangement. Often this point is located at, or inside, the inlet to the pump. Most manufacturers specify how much NPSH is required for satisfactory operation of their pumps. Hence, the **actual NPSH (NPSHA)** experienced by the pump must be larger than the manufacturer's **required NPSH (NPSHR)**. If a design indicates insufficient NPSH, changes should be made in the system, possibly including alternative piping layout such as changes in pipe elevation and/or size, or use of a pump with smaller NPSH requirements.

**Figure 39.4** A full range of performance information should be available from the pump manufacturer, and this may include the parameters shown.

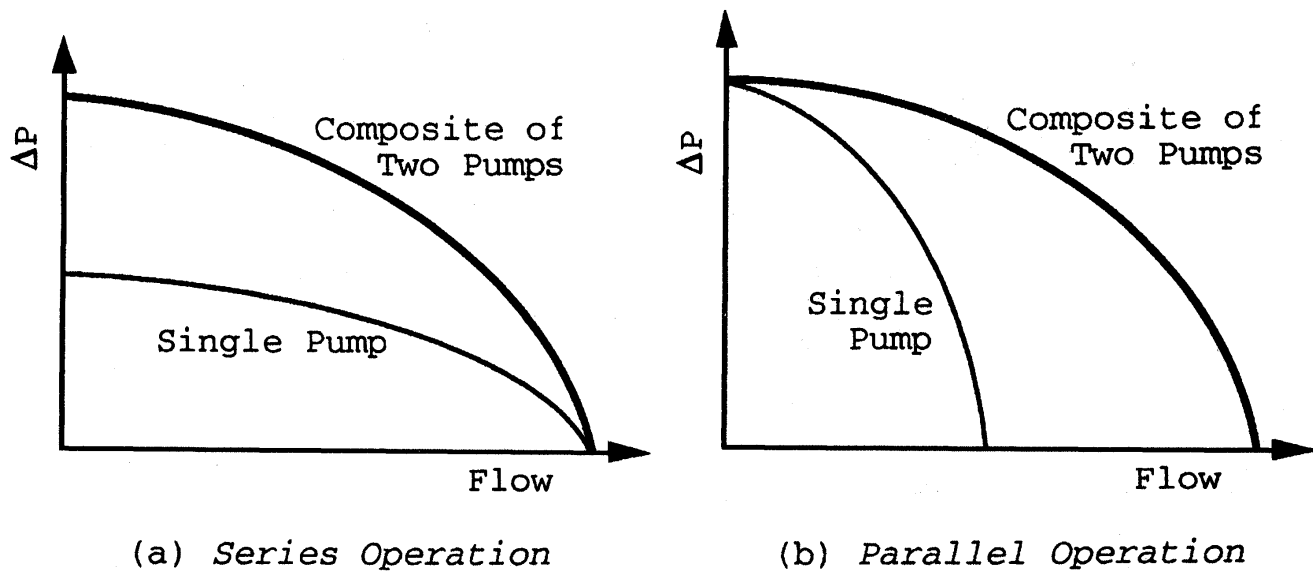


The manufacturer should be consulted for a map of operational information for a given pump. A typical form is shown in [Fig. 39.4](#). This information will allow the designer to select a pump that satisfies the circuit operational requirements while meeting the necessary NPSH and most efficient operation criteria.

Several options are available to the designer for combining pumps in systems. Consider a comparison of the net effect between operating pumps in series or operating the same two pumps in parallel. For pumps with characteristics like centrifugal units, examples of this distinction are shown in [Fig. 39.5](#). It is clear that one way to achieve high pumping pressures with centrifugal pumps is to place a number of units in series. This is an effect related to that found in *multistage* designs.



**Figure 39.5** Series and parallel operation of centrifugal pumps. The resultant characteristics for two identical pumps are shown.



## 39.2 Fans

As noted earlier, fans are devices that cause air to move. This definition is broad and can include, for instance, a flapping palm branch; the discussion here deals only with devices that impart air movement due to *rotation of an impeller inside a fixed casing*. In spite of this limiting definition, a large variety of commercial designs are included.

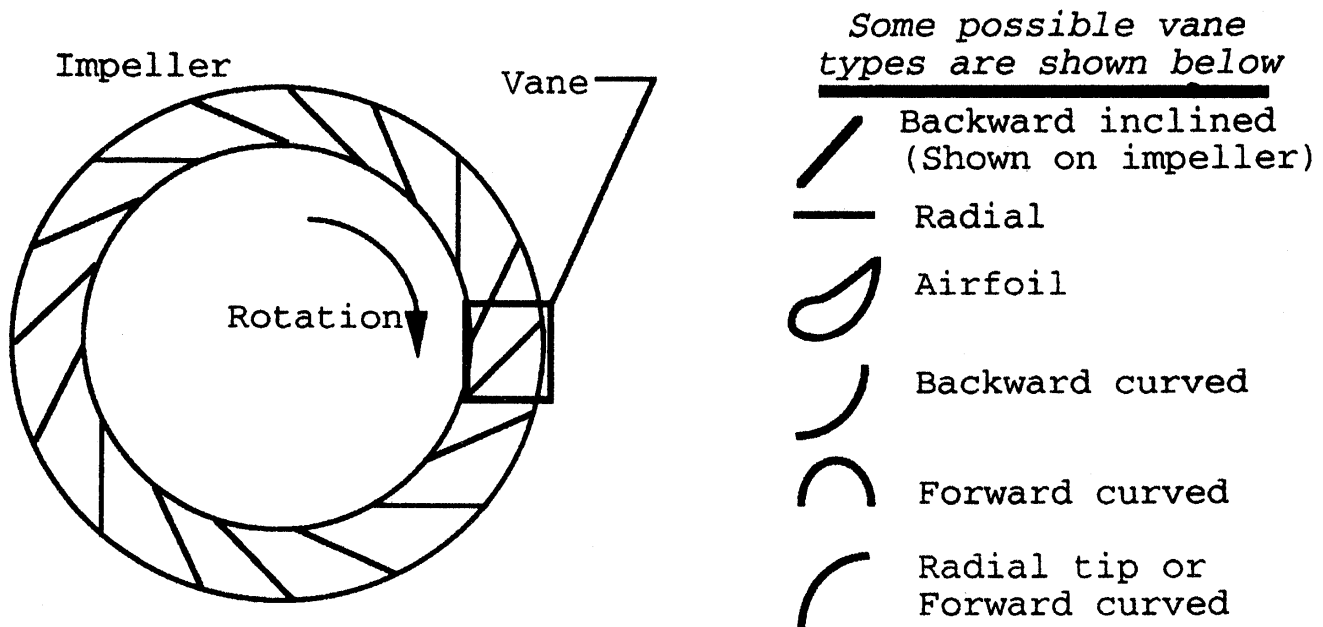
Fans find application in many engineering systems. Along with heating and cooling equipment, they are the heart of *heating, ventilating, and air conditioning* (HVAC) systems. When the physical dimensions of a unit are not a significant limitation (usually the case), centrifugal fans are favored over axial flow units for HVAC applications. Many types of fans are found in *power plants*. Very large fans are used to furnish air to the boiler, as well as to draw or force air through cooling towers and pollution control equipment. *Electronic cooling* finds applications for small units. Because of the great engineering importance of fans, several organizations publish rating and testing criteria [see, for example, ASME (1990)].

Generally fans are classified according to how the air flows through the impeller. These flows may be *axial* (essentially a propeller in a duct), *radial* (conceptually much like the centrifugal pumps discussed earlier), *mixed*, and *cross*. Although there are many other fan designations, all industrial units are one of these classifications. Mixed-flow fans are so named because both axial and radial flow occurs on the vanes. Casings for these devices are essentially like those for axial-flow machines, but the inlet has a radial-flow component. On cross-flow impellers, the gas traverses the blading twice.

A variety of vane types are found on fans, and each particular type is also used for fan classification. Axial fans usually have vanes of airfoil shape or vanes of uniform thickness. Some vane types that might be found on a centrifugal (radial-flow) fan are shown in [Fig. 39.6](#).



**Figure 39.6** Variety of vane types that might be used on a centrifugal fan.



Each type of fan has some specific qualities for certain applications. In general terms, most installations use centrifugal (radial-flow) fans. A primary exception is for very-high-flow, low-pressure-rise situations in which axial (propeller) fans are used.

Similarities exist between fans and pumps because the fluid density essentially does not vary through either type of machine. Of course in pumps this is because a liquid can be assumed to be incompressible. In fans, a gas (typically air) is moved with little pressure change. As a result, the gas density can be taken to be constant. Since most fans operate near atmospheric pressure, the ideal gas equation can be used in determining gas properties.

Flow control in fan applications, where needed, is a very important design concern. Methods for accomplishing this involve use of dampers (either on the inlet or on the outlet of the fan), variable pitch vanes, or variable speed control. Dampers are the least expensive to install but also the most inefficient in terms of energy use. Modern solid state controls for providing a variable frequency power to the drive motor is becoming the preferred control method when a combination of initial and operating costs is considered.

## Defining Terms

**Actual net positive suction head (NPSHA):** The NPSH at the given state of operation of a pump.

**Cavitation:** A state in which local liquid conditions allow vapor voids to form (boiling).

**Net positive suction head (NPSH):** The difference between the local absolute pressure of a liquid and the liquid's thermodynamic saturation pressure based on the liquid's temperature. Applies

to the inlet of a pump.

**Required net positive suction head (NPSHR):** The amount of NPSH required by a specific pump for a given application.

## References

- ASME. 1990. *ASME Performance Test Codes, Code on Fans*. ASME PTC 11-1984 (reaffirmed 1990). American Society of Mechanical Engineers, New York.
- Boehm, R. F. 1987. *Design Analysis of Thermal Systems*. John Wiley & Sons, New York, pp. 17–26.
- Ewbank, T. 1842. *A Description and Historical Account of Hydraulic and Other Machines for Raising Water*, 2nd ed. Greeley and McElrath, New York.
- Krutzsch, W. C. 1986. Introduction: Classification and selection of pumps. In *Pump Handbook*, 2nd ed., ed. I. Karassik *et al.* McGraw-Hill, New York.
- Lobanoff, V. and Ross, R. 1992. *Centrifugal Pumps: Design & Application*, 2nd ed. Gulf, Houston, TX.
- Turton, R. K. 1994. *Rotodynamic Pump Design*. Cambridge University Press, England.

## Further Information

- ASHRAE. 1992. Fans. Chapter 18 of *1992 ASHRAE Handbook, HVAC Systems and Equipment*. American Society of Heating, Refrigerating, and Air Conditioning Engineers, Atlanta, GA.
- Dickson, C. 1988. *Pumping Manual*, 8th ed. Trade & Technical Press, Morden, England.
- Dufour, J. and Nelson, W. 1993. *Centrifugal Pump Sourcebook*, McGraw-Hill, New York.
- Garay, P. N. 1990. *Pump Application Book*. Fairmont Press, Liburn, GA.
- Krivchencko, G. I. 1994. *Hydraulic Machines, Turbines and Pumps*, 2nd ed. Lewis, Boca Raton, FL.
- Stepanoff, A. J. 1993. *Centrifugal and Axial Flow Pumps: Theory, Design, and Application*. Krieger, Malabar, FL.

Lahey, R. T. Jr. "Two-Phase Flow"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

[40.1 Notation](#)[40.2 Conservation Equations](#)

Mass Conservation • Momentum Conservation • Energy Conservation

[40.3 Closure](#)[40.4 Two-Phase Instabilities](#)[40.5 Conclusion](#)**Richard T. Lahey, Jr.***Rensselaer Polytechnic Institute*

Multiphase flows occur in many cases of practical concern. In particular, important vapor/liquid and solid/fluid two-phase flows may occur in thermal energy production and utilization, chemical and food processing, environmental engineering, pharmaceutical manufacturing, petroleum production, and waste incineration technologies.

This chapter summarizes the essential features of two-phase flow and focuses on the engineering analysis of vapor/liquid systems. Readers interested in a more in-depth treatment are referred to the work of Lahey [1992] and Roco [1993].

Two-phase flows are inherently more complicated than single-phase flows. This is due to the fact that the phases may separate and arrange themselves into distinct flow regimes. Moreover, the two phases normally do not travel at the same velocity, nor, in some situations, even in the same direction. As a consequence, there are important phenomena in two-phase flows that do not occur in single-phase flows.

The notation, conservation equations, and their associated closure relations will first be discussed. Next, flooding will be considered and then the conservation equations will be used to analyze some situations of interest in two-phase flows.

---

**40.1 Notation**

The most important parameter that characterizes a two-phase flow is the so-called *local volume fraction* of phase  $k$ ,  $\alpha_k(x, t)$ . This parameter is the time fraction that a probe at location  $x$  and time  $t$  will sense phase  $k$  during a measurement time  $T$ . That is,

$$\alpha_k = \sum_{i=1}^N \Delta t_i / T \quad (40.1)$$

The global volume fraction of phase  $k$  is the integral of the local volume fraction over the cross-sectional area ( $A_{xs}$ ) of the conduit:

$$\langle \alpha_k \rangle = \int \int_{A_{xs}} \alpha_k da / A_{xs} \quad (40.2)$$

By convention, in vapor/liquid two-phase flows, the volume fraction of the vapor phase,  $\alpha_v$ , is called the *void fraction*,  $\alpha$ .

Unlike single-phase flows, there are multiple velocities of interest in two-phase flows—in particular, the superficial velocity,

$$\langle j_k \rangle = Q_k / A_{xs} \quad (40.3)$$

and the phasic velocity,

$$\langle u_k \rangle = Q_k / A_k = \langle j_k \rangle / \langle \alpha_k \rangle \quad (40.4)$$

Also, the two phases normally do not travel with the same velocity. Indeed, in vapor/liquid systems, the vapor often travels faster than the liquid, giving rise to a local relative velocity ( $u_R$ ),

$$u_R = u_v - u_l \quad (40.5a)$$

or slip ratio ( $S$ ),

$$S = u_v / u_l \quad (40.5b)$$

The density of a two-phase mixture is given by

$$\langle \rho \rangle = \left[ \int \int \int_{V_l} \rho_l dv + \int \int \int_{V_v} \rho_v dv \right] / (V_l + V_v) \quad (40.6a)$$

Thus, from Eq. (40.2),

$$\langle \rho \rangle = \rho_l (1 - \langle \alpha \rangle) + \rho_v \langle \alpha \rangle \quad (40.6b)$$

## 40.2 Conservation Equations

---

Equations for the conservation of mass, momentum, and energy are needed to describe a flowing two-phase mixture. The appropriate one-dimensional, two-fluid (i.e., writing the conservation laws for each phase separately), and mixture conservation equations are described in the following

sections. Let us consider the case of a vapor/liquid system, since this is the most complicated case.

## Mass Conservation

For the vapor phase,

$$\frac{\partial}{\partial t} [\rho_v \langle \alpha \rangle A_{xs}] + \frac{\partial}{\partial z} [\rho_v \langle \alpha \rangle \langle u_v \rangle A_{xs}] = \Gamma A_{xs} \quad (40.7a)$$

For the liquid phase,

$$\frac{\partial}{\partial t} [\rho_l (1 - \langle \alpha \rangle) A_{xs}] + \frac{\partial}{\partial z} [\rho_l (1 - \langle \alpha \rangle) \langle u_l \rangle A_{xs}] = -\Gamma A_{xs} \quad (40.7b)$$

where  $\Gamma$  is the amount of liquid evaporated per unit volume per unit time.

Adding Eqs. (40.7a) and (40.7b) together yields the one-dimensional mixture continuity equation,

$$\frac{\partial}{\partial t} [\langle \rho \rangle A_{xs}] + \frac{\partial}{\partial z} [G A_{xs}] = 0 \quad (40.8)$$

where the mass flux is given by

$$G \triangleq w/A_{xs} = \rho_v \langle \alpha \rangle \langle u_v \rangle + \rho_l (1 - \langle \alpha \rangle) \langle u_l \rangle \quad (40.9)$$

## Momentum Conservation

For the vapor phase,

$$\begin{aligned} & \frac{1}{g_c} \left[ \frac{\partial}{\partial t} (\rho_v \langle \alpha \rangle \langle u_v \rangle) + \frac{1}{A_{xs}} \frac{\partial}{\partial z} (\rho_v \langle \alpha \rangle \langle u_v \rangle^2 A_{xs}) \right] \\ &= -\langle \alpha \rangle \frac{\partial p}{\partial z} - \frac{g}{g_c} \rho_v \langle \alpha \rangle \sin \theta - \frac{\tau_i P_i}{A_{xs}} + \frac{\Gamma u_i}{g_c} \end{aligned} \quad (40.10a)$$

For the liquid phase,

$$\begin{aligned} & \frac{1}{g_c} \left[ \frac{\partial}{\partial t} [\rho_l (1 - \langle \alpha \rangle) \langle u_l \rangle] + \frac{1}{A_{xs}} \frac{\partial}{\partial z} [\rho_l (1 - \langle \alpha \rangle) \langle u_l \rangle^2 A_{xs}] \right] \\ &= -(1 - \langle \alpha \rangle) \frac{\partial p}{\partial z} - \frac{g}{g_c} \rho_l (1 - \langle \alpha \rangle) \sin \theta - \frac{\tau_w P_f}{A_{xs}} + \frac{\tau_i P_i}{A_{xs}} - \frac{\Gamma u_i}{g_c} \end{aligned} \quad (40.10b)$$

where the velocity of the vapor/liquid interface,  $u_i$ , the interfacial perimeter,  $P_i$ , and the interfacial shear stress,  $\tau_i$ , must be constituted to achieve closure [Lahey and Drew, 1992].

Adding Eqs. (40.10a) and (40.10b) and allowing for the possibility of  $N$  local losses, the one-dimensional mixture momentum equation can be written as

$$\begin{aligned} \frac{1}{g_c} \left[ \frac{\partial G}{\partial t} + \frac{1}{A_{xs}} \frac{\partial}{\partial z} (G^2 A_{xs} / \langle \rho' \rangle) \right] = & - \frac{\partial p}{\partial z} - \frac{g}{g_c} \langle \rho \rangle \sin \theta - \frac{\tau_w P_f}{A_{xs}} \\ & - \sum_{i=1}^N K_i \frac{G^2 \delta(z - z_i) P_f}{2g_c \rho_l A_{xs}} \Phi(z) \end{aligned} \quad (40.11)$$

where  $\langle \rho' \rangle$  is a two-phase "density" given by:

$$\langle \rho' \rangle = \left[ \frac{(1 - \langle x \rangle)^2}{\rho_l (1 - \langle \alpha \rangle)} + \frac{\langle x \rangle^2}{\rho_v \langle \alpha \rangle} \right]^{-1} \quad (40.12)$$

and  $\langle x \rangle$  is the flow quality. It should be noted that closure models for the wall shear,  $\tau_w$ , will be discussed subsequently.

It is interesting to note that one can rearrange Eq. (40.11) into drift-flux form [Lahey and Moody, 1993]:

$$\begin{aligned} \frac{1}{g_c} \left[ \frac{\partial G}{\partial t} + \frac{1}{A_{xs}} \frac{\partial}{\partial z} \left( \frac{G^2 A_{xs}}{\langle \rho \rangle} \right) \right] = & - \frac{\partial p}{\partial z} - \frac{g}{g_c} \langle \rho \rangle \sin \theta - \frac{\tau_w P_f}{A_{xs}} \\ & - \frac{1}{g_c A_{xs}} \frac{\partial}{\partial z} \left[ A_{xs} \left( \frac{\rho_l - \langle \rho \rangle}{\langle \rho \rangle - \rho_v} \right) \frac{\rho_l \rho_v}{\langle \rho \rangle} (V'_{gj})^2 \right] + \sum_{i=1}^N K_i \frac{G^2 \delta(z - z_i) P_f}{2g_c \rho_l A_{xs}} \Phi(z) \end{aligned} \quad (40.13)$$

where, as will be discussed subsequently,  $V'_{gj} = V_{gj} + (C_0 - 1)\langle j \rangle$  is the generalized drift velocity.

## Energy Conservation

For the vapor phase,

$$\begin{aligned} \frac{\partial}{\partial t} [\rho_v \langle \alpha \rangle (\langle e_v \rangle - p/J\rho_v) A_{xs}] + \frac{\partial}{\partial z} [\rho_v \langle u_v \rangle \langle \alpha \rangle A_{xs} \langle e_v \rangle] \\ = \Gamma A_{xs} e_{v_i} - p_i A_{xs} \frac{\partial \langle \alpha \rangle}{\partial t} + q_v''' \langle \alpha \rangle A_{xs} - q_{v_i}'' P_i \end{aligned} \quad (40.14a)$$

For the liquid phase,

$$\begin{aligned} & \frac{\partial}{\partial t} [\rho_l(1 - \langle \alpha \rangle) (\langle e_l \rangle - p/J\rho_l) A_{xs}] + \frac{\partial}{\partial z} [\rho_l \langle u_l \rangle (1 - \langle \alpha \rangle) A_{xs} \langle e_l \rangle] \\ &= -\Gamma A_{xs} e_{l_i} + p_i A_{xs} \frac{\partial \langle \alpha \rangle}{\partial t} + q_w'' P_H + q_l''' (1 - \langle \alpha \rangle) A_{xs} + q_{l_i}'' P_i \end{aligned} \quad (40.14b)$$

where:

$$\langle e_k \rangle \triangleq h_k + \frac{\langle u_k \rangle^2}{2g_c J} + \frac{gz \sin \theta}{g_c J} \quad (40.15)$$

As noted before, appropriate closure laws are needed for the specific phasic interfacial energy,  $e_{k_i}$ , the heat fluxes,  $q_{k_i}''$ , the interfacial perimeter,  $P_i$ , and the volumetric heating rate,  $q_k'''$  [Lahey and Drew, 1992]. Also, it is interesting to note that the interfacial jump condition gives

$$\Gamma = (q_{v_i}'' - q_{l_i}'') P_i / A_{xs} h_{fg} \quad (40.16)$$

If we add Eqs. (40.14a) and (40.14b) we obtain the mixture energy equation in the form

$$\begin{aligned} & \frac{\partial}{\partial t} [\rho_l(1 - \langle \alpha \rangle) (\langle e_l \rangle - p/J\rho_l) + \rho_v \langle \alpha \rangle (\langle e_v \rangle - p/J\rho_v)] A_{xs} \\ &+ \frac{\partial}{\partial z} [w_l \langle e_l \rangle + w_v \langle e_v \rangle] = q_w'' A_{xs} + q''' A_{xs} \end{aligned} \quad (40.17)$$

## 40.3 Closure

In order to be able to evaluate the conservation equations that describe two-phase flow we must first achieve closure by constituting all parameters in these equations in terms of their state variables.

In order to demonstrate the process, let us consider the mixture conservation equations, Eqs. (40.8), (40.13), and (40.17).

We can use the Zuber-Findlay drift-flux model to relate the void fraction to the superficial velocities:

$$\langle \alpha \rangle = \frac{\langle j_v \rangle}{C_0 \langle j \rangle + V_{gj}} \quad (40.18)$$



where  $\langle j \rangle \triangleq \langle j_v \rangle + \langle j_l \rangle$ ,  $C_0$  is the void concentration parameter, and  $V_{gj}$  is the so-called *drift velocity*. Both of these drift-flux parameters are normally given by correlations that come from appropriate data [Lahey and Moody, 1993].

The wall shear,  $\tau_w$ , is given by

$$\frac{\tau_w P_f}{A_{xs}} = \frac{f G^2 \phi_{l0}^2}{2 g_c D_H \rho_l} \quad (40.19)$$

where  $\phi_{l0}^2$  is the two-phase friction pressure drop multiplier. It is just the ratio of the single-phase (liquid) to the two-phase density,

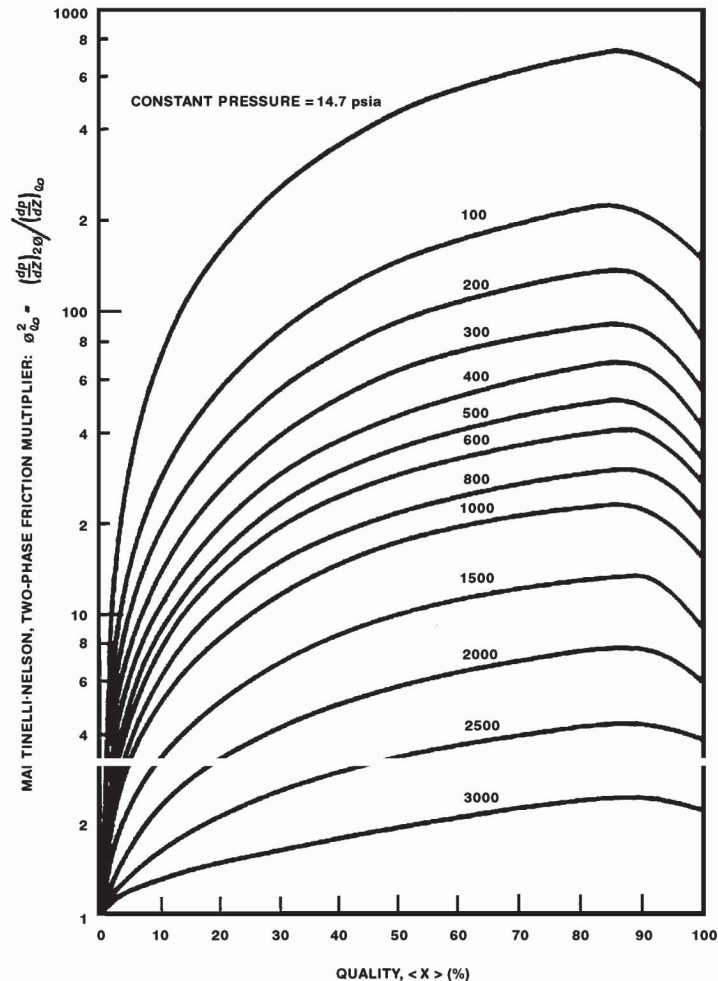
$$\phi_{l0}^2 = \rho_l / \rho_{2\phi} \quad (40.20)$$

For example, for homogeneous two-phase flows (in which the slip ratio,  $S$ , is unity) we have

$$\phi_{l0}^2 = \rho_l / \langle \rho_h \rangle = 1 + \frac{v_{fg}}{v_f} \langle x \rangle \quad (40.21)$$

For slip flows, empirical correlations, such as that of Martinelli-Nelson, shown in Fig. 40.1, are often used.

**Figure 40.1** Martinelli-Nelson two-phase friction multiplier for steam/water as a function of quality and pressure.

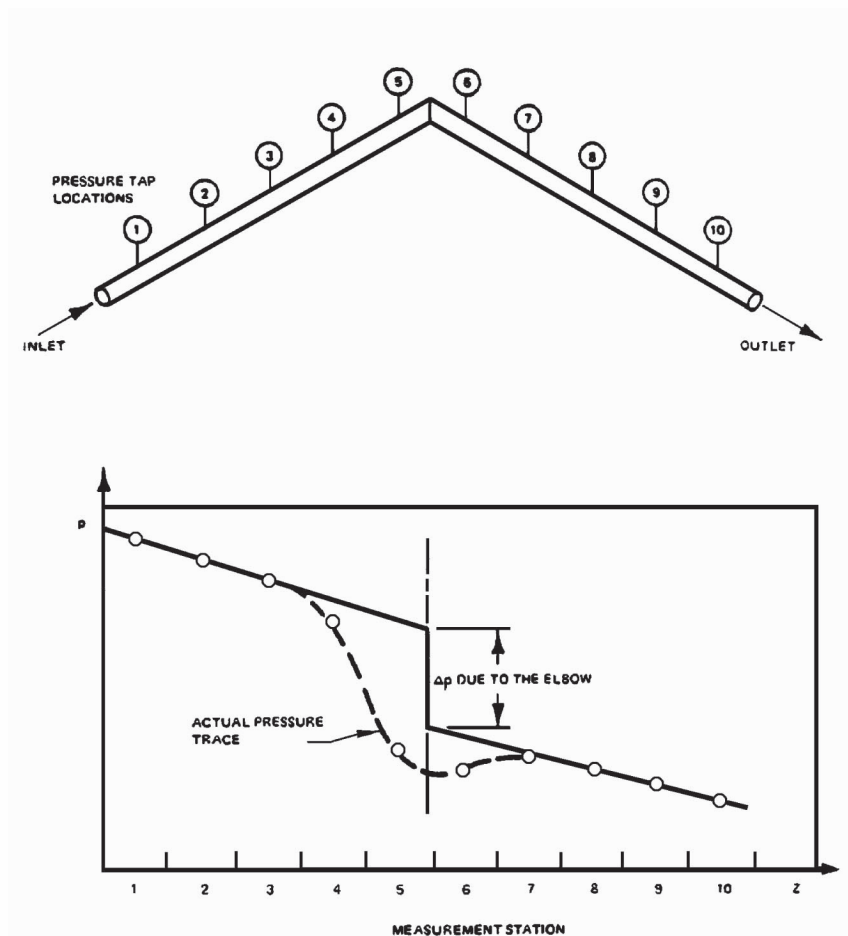


For local losses (e.g., orifices, spacers, etc.) it is normal practice to assume a homogeneous multiplier; thus, as in Eq. (40.21),

$$\Phi = 1 + \frac{v_{fg}}{v_f} \langle x \rangle \quad (40.22)$$

Figure 40.2 shows a typical pressure drop profile, which includes both local and distributed two-phase losses.

**Figure 40.2** Two-phase pressure drop in an elbow.



One of the interesting features of two-phase flows is that countercurrent flows may occur, in which the phases flow in different directions. In a vapor/liquid system this can lead to a countercurrent flow limitation (CCFL) or **flooding** condition.

When flooding occurs in a vertical conduit, in which the liquid is flowing downward and the vapor upward, the liquid downflow will be limited by excessive friction at the vapor/liquid

interface. A popular CCFL correlation is that of Wallis [1969],

$$(j_g^*)^{1/2} + (|j_f^*|)^{1/2} = C \quad (40.23)$$

where the square roots of the phasic Froude numbers are given by

$$j_k^* = \frac{\langle j_k \rangle \rho_k^{1/2}}{[g D_H (\rho_l - \rho_v)]^{1/2}} \quad (40.24)$$

This CCFL correlation is known to work well in small-diameter conduits.

For large-diameter conduits the Kutateladze CCFL correlation works much better:

$$K_v^{1/2} + |K_l|^{1/2} = 1.79 \quad (40.25)$$

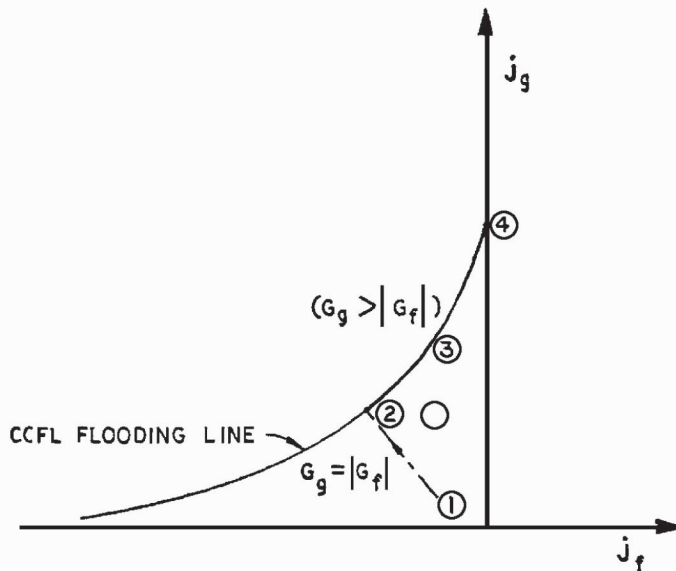
where

$$K_k = \frac{\langle j_k \rangle \rho_k^{1/2}}{[\sigma g g_c (\rho_l - \rho_v)]^{1/4}} \quad (40.26)$$

A strategy for switching from one CCFL correlation to the other has been given by Henry *et al.* [1993] for various size conduits.

Figure 40.3 shows that both Eqs. (40.23) and (40.25) imply that there will be no liquid downflow when the vapor upflow velocity ( $j_v$ ) is large enough (i.e., above point 4).

**Figure 40.3** Typical CCFL flooding locus.



Let us now use the conservation equations to analyze some two-phase flow phenomena of interest. For example, let us consider **critical flow** (i.e., the sonic discharge of a two-phase mixture).

For steady state conditions in which there are no local losses Eq. (40.11) can be expanded and rearranged to yield, for choked flow conditions,

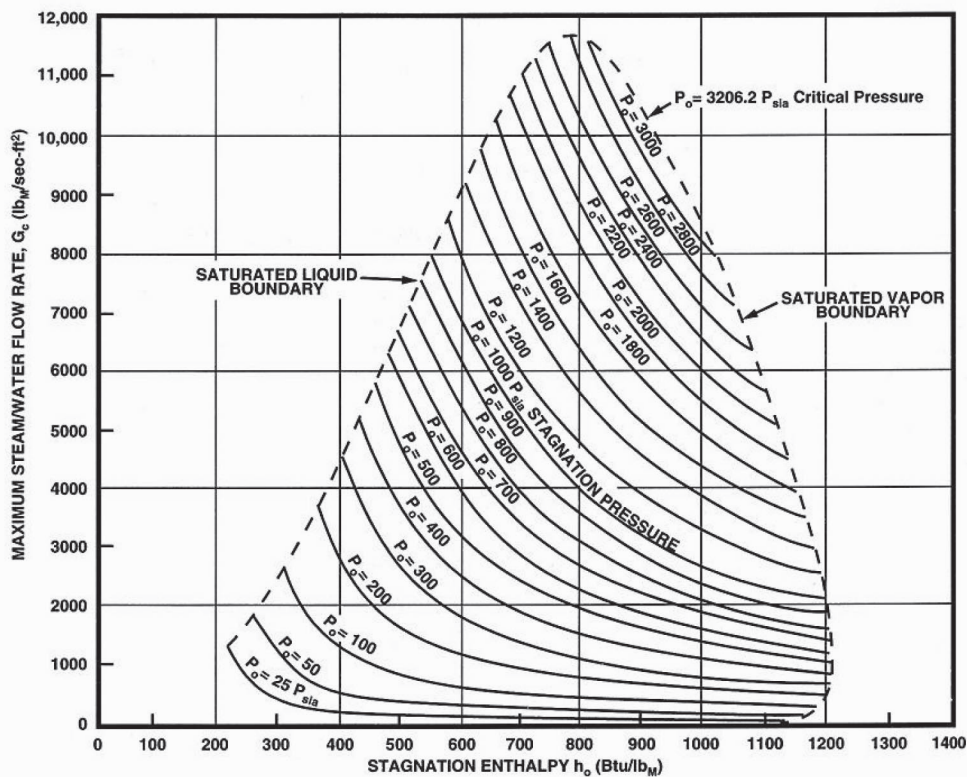
$$-\frac{dp}{dz} = \frac{\frac{-G_c^2}{g_c A_{XS} \langle \rho' \rangle} \frac{dA_{XS}}{dz} + \frac{g}{g_c} \langle \rho \rangle + \frac{\tau_w P_f}{A_{XS}}}{\left[ 1 + \frac{G_c^2}{g_c} \frac{d}{dp} \left( \frac{1}{\langle \rho' \rangle} \right) \right]} \quad (40.27)$$

Vanishing of the numerator determines where the choking plane will be (i.e., where the Mach number is unity), whereas vanishing of the denominator defines the local critical mass flux as

$$G_c = \left[ -g_c \frac{dp}{d(1/\langle \rho' \rangle)} \right]^{1/2} \quad (40.28)$$

In order to proceed, we must know the local properties and the slip ratio ( $S$ ). It can be shown [Lahey and Moody, 1993] that if the specific kinetic energy of the two-phase mixture is minimized we obtain  $S = (\rho_l/\rho_v)^{1/3}$ . Moreover, if an isotropic thermodynamic process is assumed through a converging nozzle, Fig. 40.4 is obtained. This figure is very easy to use. For example, if the stagnation pressure ( $p_0$ ) and enthalpy ( $h_0$ ) upstream of the nozzle are 1000 psia and 800 Btu/lbm, respectively, the critical mass flux will be about  $G_c = 3995 \text{ lbm/s-ft}^2$ .

**Figure 40.4** Maximum steam/water flow rate and local stagnation properties (Moody model).



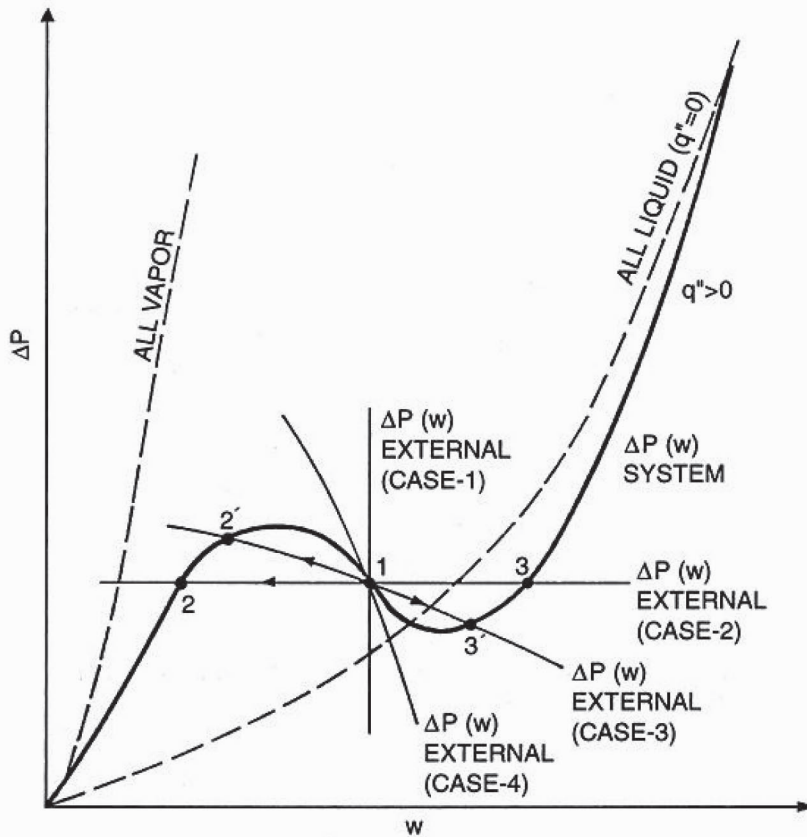
## 40.4 Two-Phase Instabilities

It is also significant to note that important static and dynamic instabilities may occur in two-phase flows. It can be shown [Lahey and Moody, 1993] that the criterion for the occurrence of an excursive instability is

$$\frac{\partial (\Delta p_{\text{system}})}{\partial w} > \frac{\partial (\Delta p_{\text{ext}})}{\partial w} \quad (40.29)$$

where the external (ext) pressure increase ( $\Delta p_{\text{ext}}$ ) is normally due to a pump. Figure 40.5 shows that a two-phase system having a positive displacement pump (case 1) and a centrifugal pump with a steep pump/head curve (case 4) will be stable, whereas the case of parallel channels (case 2) and a centrifugal pump with a relatively flat pump/head curve (case 3) are unstable and have multiple operating points.

**Figure 40.5** Excursive instability.



If the state variables in the mixture conservation equations are perturbed (i.e., linearized) about a steady state,

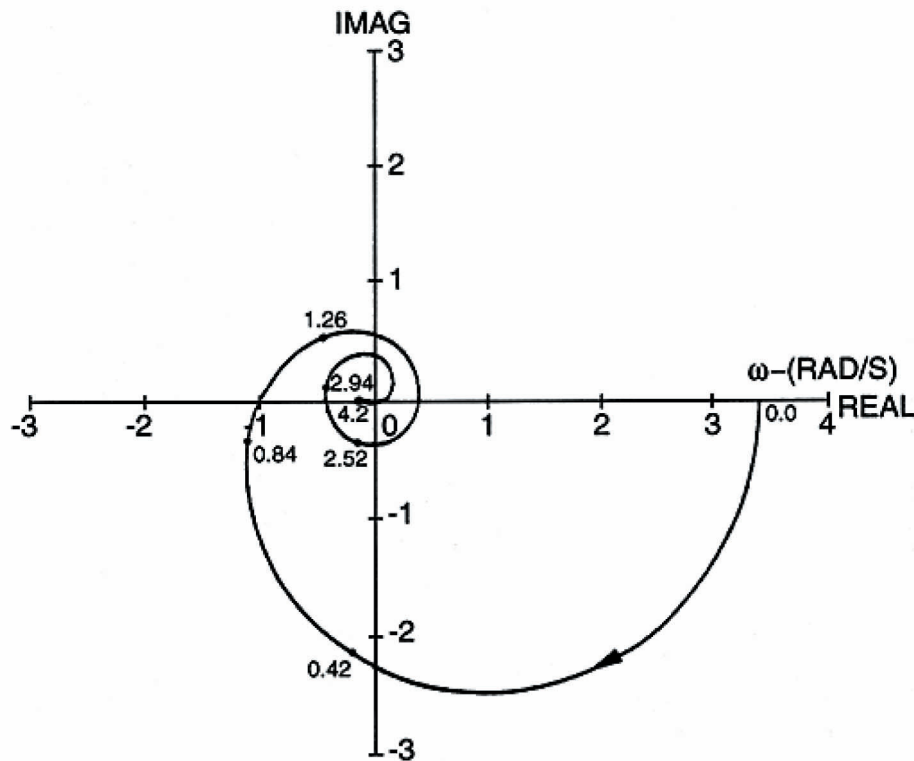
$$\delta \varphi(\eta, t) \triangleq \varphi(\eta, t) - \varphi_0 = \left. \frac{\partial \varphi}{\partial \eta} \right|_0 \delta \eta \quad (40.30)$$

and the resultant linear equations are combined and integrated in the axial ( $z$ ) direction, then the so-called *characteristics equation* of a boiling (or condensing) system becomes [Lahey and Podowski, 1989]:

$$\delta(\Delta p_{1\varphi}) + \delta(\Delta p_{2\varphi}) = 0 \quad (40.31)$$

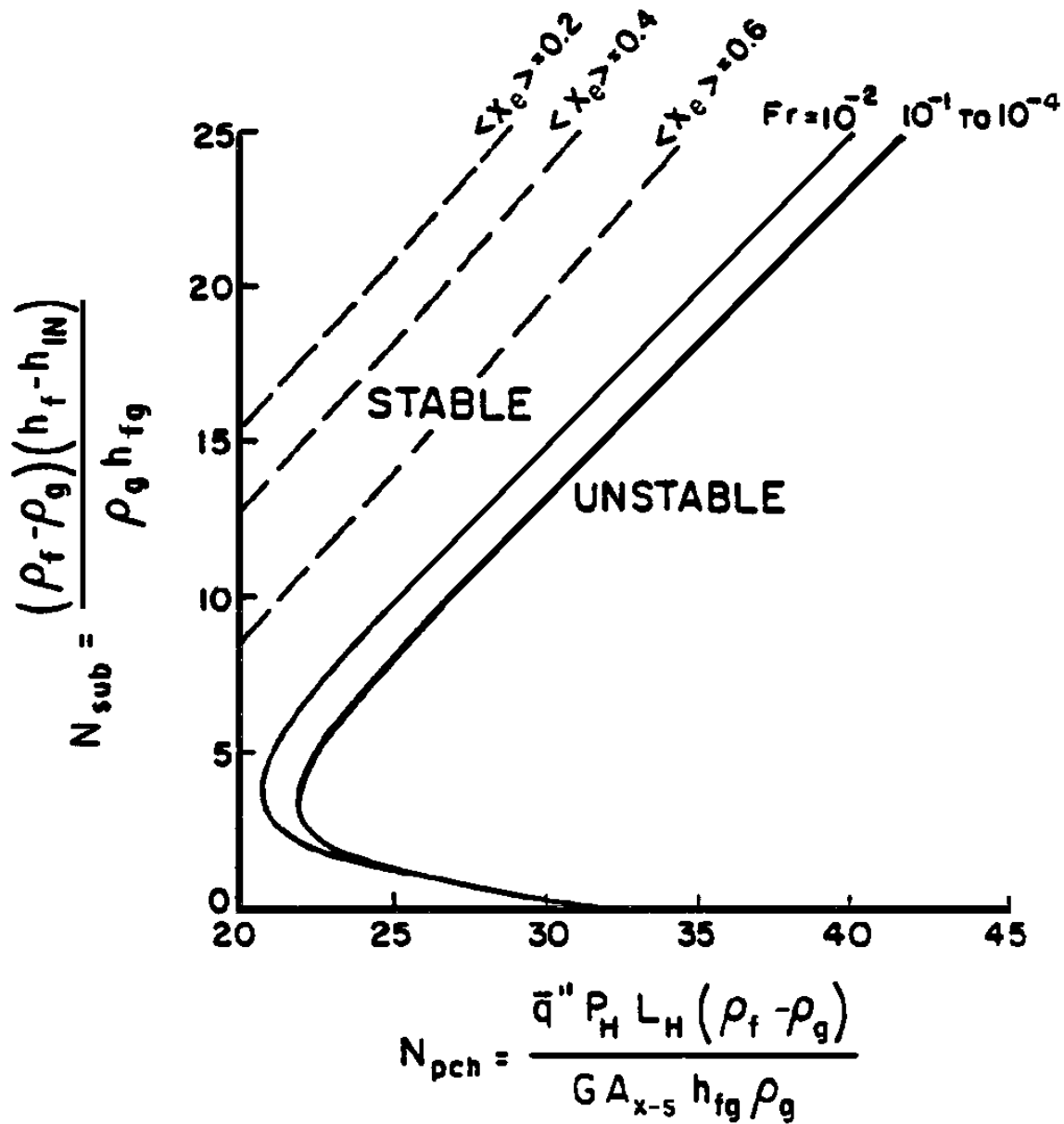
After Laplace-transforming Eq. (40.31)—to convert it from the time domain to the frequency domain—we can apply Nyquist's stability technique to determine whether or not density-wave oscillations (DWO) are expected. Significantly, DWO are the most important and prevalent dynamic oscillations that may occur in two-phase flows. For the parallel channel case shown schematically in Fig. 40.6, we see that the system is marginally stable (i.e., the Nyquist locus of  $\delta(\Delta \hat{p}_{2\varphi}) / \delta(\Delta \hat{p}_{1\varphi})$  goes through the  $-1$  point) and DWO are thus anticipated.

**Figure 40.6** Nyquist plot for abnormal BWR/4 operating conditions (parallel channels).



A stability map of boiling/condensing systems is often given in terms of the subcooling number ( $N_{\text{sub}}$ ) and the phase change number ( $N_{\text{pch}}$ ). A typical plot for a typical boiling water nuclear reactor (BWR/4) is given in Fig. 40.7. It can be seen that, as the inlet subcooling is increased at a given power-to-flow ratio (i.e., for  $N_{\text{pch}}$  fixed), the system may go in and out of DWO.

**Figure 40.7** Typical BWR/4 stability map ( $K_{in} = 27:8$ ;  $K_{exit} = 0:14$ ).



## 40.5 Conclusion

This chapter summarizes some of the methods used in the analysis of two-phase flows. The emphasis has been on two-phase fluid mechanics, although phase change heat transfer is often also very important.

The literature associated with two-phase thermal-hydraulics is vast; nevertheless, it is hoped that this chapter will provide a useful road map to this literature.

## Defining Terms

**Critical flow:** A condition in which the two-phase mixture is flowing at the local sonic velocity.

**Density:** The mass per unit volume of the material in question ( $\rho$ ).

**Enthalpy:** The internal energy per unit volume plus the related flow work ( $h$ ).

**Flooding:** A countercurrent flow limitation due to the resistance induced by cocurrent vapor/liquid flow streams.

**Flow area:** The cross-sectional area through which the fluid flows ( $A_{xs}$ ).

**Flow rate:** The mass flowing per unit time ( $w$ ).

**Phase:** The various phases of a fluid are solid, liquid, and gas.

**Void fraction:** The local volume fraction of the dispersed vapor phase.

**Volumetric flow rate:** The volume flowing per unit time of phase  $k$  ( $Q_k$ ).

## Nomenclature

$\alpha$	= void function
$\delta(z)$	= Dirac delta function
$\Gamma$	= mass of vapor generated per unit volume per unit time
$\theta$	= angle from the horizontal plane
$\rho$	= density
$\tau$	= shear stress
$A_{xs}$	= cross sectional area
$D_H$	= hydraulic diameter
$\langle e_k \rangle$	= specific total convected energy
$f$	= Moody friction factor
$g_c$	= gravitational constant
$G$	= mass flux
$j_k$	= superficial velocity of phase $k$
$J$	= mechanical equivalent of heat
$K_k$	= Kutateladze number of phase $k$
$p$	= static pressure
$P$	= perimeter
$q''$	= heat flux
$q'''$	= volumetric heat generation rate
$Q$	= volumetric flow rate
$S$	= slip ratio
$u_k$	= velocity of phase $k$
$w$	= flow rate
$x$	= flow quality



$\tilde{z}$	= axial position
<b>Subscripts</b>	
$c$	= critical
$f$	= friction
$i$	= interface
$k$	= phase identification
$l$	= liquid
$v$	= vapor
$w$	= wall
$2\phi$	= two-phase

## References

- Henry, C., Henry, R., Bankoff, S. G., and Lahey, R. T., Jr. 1993. Buoyantly-driven two-phase countercurrent flow in liquid discharge from a vessel with an unvented gas space. *J. Nucl. Eng. Design*. 141(1 & 2):237–248.
- Lahey, R. T., Jr., ed. 1992. *Boiling Heat Transfer—Modern Development and Advances*, Elsevier, New York.
- Lahey, R. T., Jr., and Drew, D. A. 1992. On the development of multidimensional two-fluid models for vapor/liquid two-phase flows. *J. Chem. Eng. Commun.* 118:125–140.
- Lahey, R. T., Jr., and Moody, F. J. 1993. *The Thermal-Hydraulics of a Boiling Water Nuclear Reactor*. ANS Monograph. American Nuclear Society, La Grange Park, IL.
- Lahey, R. T., Jr., and Podowski, M. Z. 1989. On the analysis of instabilities in two-phase flows. *Multiphase Sci. Tech.* 4:183–340.
- Roco, M., ed. 1993. *Particulate Two-Phase Flow*. Hemisphere, New York.
- Wallis, G. B. 1969. *One Dimensional Two-Phase Flow*. McGraw-Hill, New York.

## Further Information

- Ishii, M. 1975. *Thermo-Fluid Dynamic Theory of Two-Phase Flow*. Eyrolles, Paris, France.
- Butterworth, D. and Hewitt, G. F. 1977. *Two-Phase Flow and Heat Transfer*. Harwell Series, Oxford University Press.
- Collier, J. G. 1972. *Convective Boiling and Condensation*. McGraw-Hill, New York.

Oldshue, J. Y. “Basic Mixing Principles for Various Types of Fluid...”  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Basic Mixing Principles for Various Types of Fluid Mixing Applications

---

## 41.1 Scaleup/Scaledown

## 41.2 Effect of Circulation Time Spectrum and the Spectrum of Shear Rates on Ten Different Mixing Technologies

Gas-Liquid Dispersion • Gas-Liquid Mass Transfer • Solids Suspension and Dispersion • Solid-Liquid Mass Transfer • Liquid-Liquid Emulsions • Liquid-Liquid Extraction • Blending • Chemical Reactions • Fluid Motion • Heat Transfer

## 41.3 Computational Fluid Dynamics

### James Y. Oldshue

*Oldshue Technologies International, Inc.*

The fluid mixing process involves three different areas of viscosity which affect flow patterns and scaleup, and two different scales within the fluid itself: **macro scale** and **micro scale**. Design questions come up with regard to the performance of mixing processes in a given volume.

Consideration must be given to proper impeller and tank geometry as well as to the proper speed and power for the impeller. Similar considerations arise when it is desired to scale up or scale down, and this involves another set of mixing considerations.

If the fluid discharge from an impeller is measured with a device that has a high frequency response, one can track the velocity of the fluid as a function of time. The velocity at a given point in time can then be expressed as an average velocity ( $\bar{v}$ ) plus a fluctuating component ( $v'$ ). Average velocities can be integrated across the discharge of the impeller, and the pumping capacity normal to an arbitrary discharge plane can be calculated. This arbitrary discharge plane is often defined by the boundaries of the impeller blade diameter and height. Because there is no casing, however, an additional 10 to 20% of flow typically can be considered as the primary flow of an impeller.

The velocity gradients between the average velocities operate only on larger particles. Typically, these particles are greater than 1000  $\mu\text{m}$  in size. This is not a precise definition, but it does give a feel for the magnitudes involved. This defines macro-scale mixing. In the turbulent region, these macro-scale fluctuations can also arise from the finite number of impeller blades passing a finite number of baffles. These set up velocity fluctuations that can also operate on the macro scale.

Smaller particles primarily see only the fluctuating velocity component. When the particle size is much less than 100  $\mu\text{m}$ , the turbulent properties of the fluid become important. This is the definition of the boundary size for micro-scale mixing.

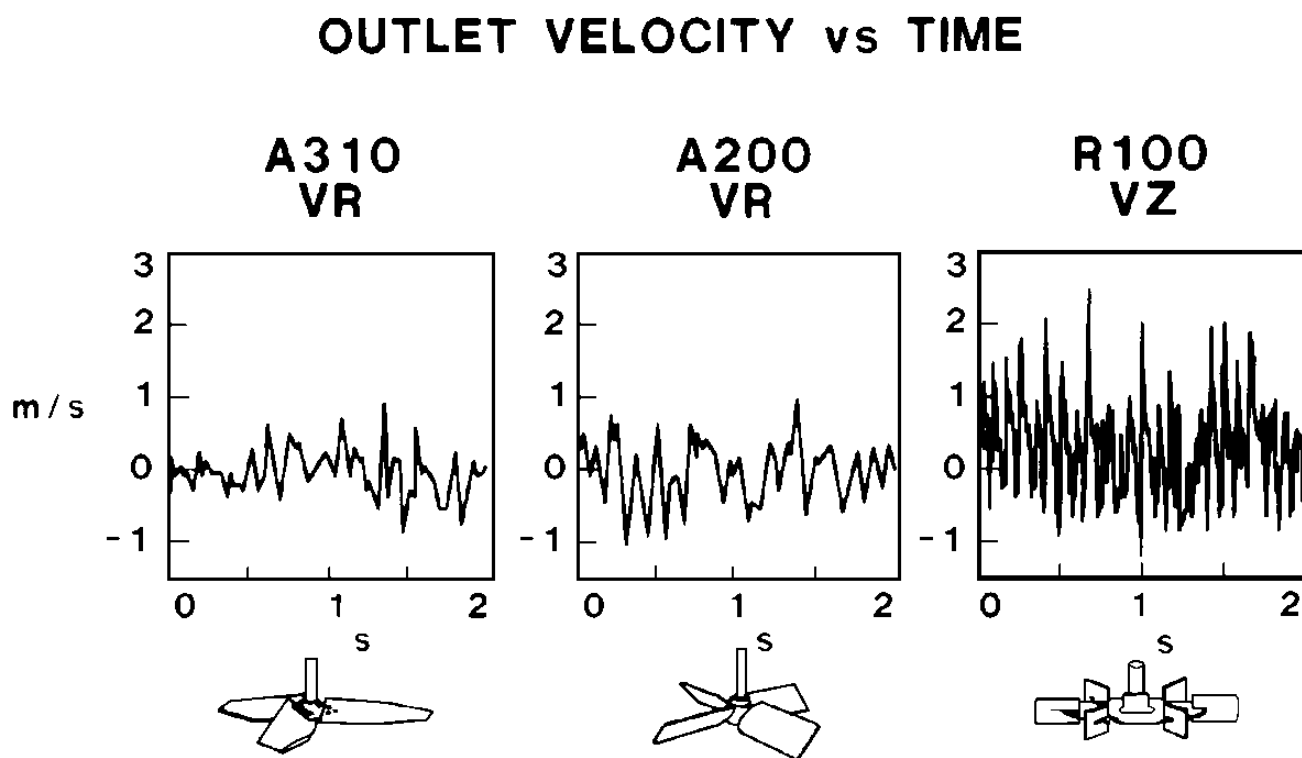
All of the power applied by a mixer to a fluid through the impeller appears as heat. The conversion of power to heat is through viscous shear and is 2542 Btu/h/hp. Viscous shear is present in turbulent flow only at the micro-scale level. As a result, the power per unit volume is a major component of the phenomenon of micro-scale mixing. At a  $1\ \mu\text{m}$  level, in fact, it doesn't matter what specific impeller design is used to apply the power.

Numerous experiments show that the power per unit volume in the zone of the impeller (which is about 5% of the total tank volume) is about 100 times higher than the power per unit volume in the rest of the vessel. Based on some reasonable assumptions about the fluid mechanics parameters, the root-mean-square (rms) velocity fluctuation in the zone of the impeller appears to be approximately five to ten times higher than in the rest of the vessel. This conclusion has been verified by experimental measurements.

The ratio of the rms velocity fluctuation to the average velocity in the impeller zone is about 50% for many open impellers. If the rms velocity fluctuation is divided by the average velocity in the rest of the vessel, however, the ratio is on the order of 5 to 10%. This is also the ratio of rms velocity fluctuation to the mean velocity in pipeline flow. In micro-scale mixing, phenomena can occur in mixing tanks that do not occur in pipeline reactors. Whether this is good or bad depends upon the process requirements.

Figure 41.1 shows velocity versus time for three different impellers. The differences between the impellers are quite significant and can be important for mixing processes. All three impeller velocities are calculated for the same impeller flow,  $Q$ , and same diameter. The A310 (Fig. 41.2) draws the least power and has the lowest velocity fluctuations. This gives the lowest micro-scale turbulence and shear rate. The A200 (Fig. 41.3) displays increased velocity fluctuations and draws more power. The R100 (Fig. 41.4) draws the most power and has the highest micro-scale shear rate. The proper impeller should be used for each individual process requirement.

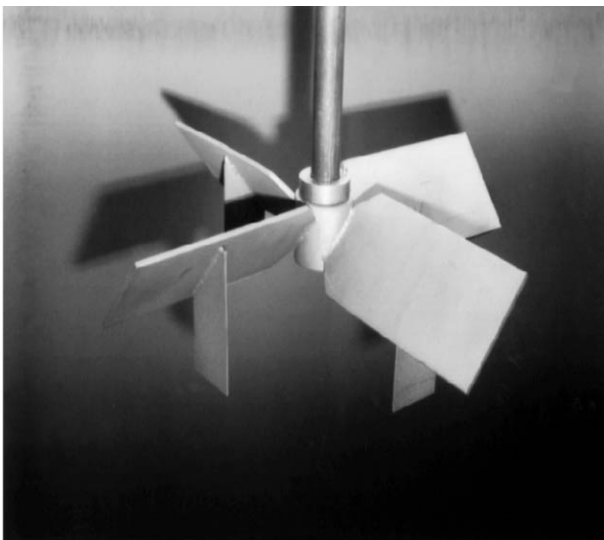
**Figure 41.1** Typical velocity as a function of time for three different impellers, all at the same total pumping capacity. (Courtesy of LIGHTNIN.)



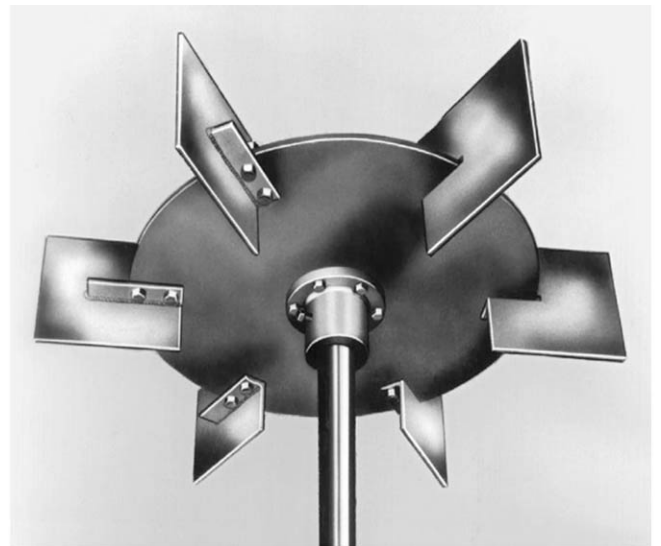
**Figure 41.2** Fluidfoil impeller (A310). (Courtesy of LIGHTNIN.)



**Figure 41.3** Typical axial-flow turbine (A200). (Courtesy of LIGHTNIN.)

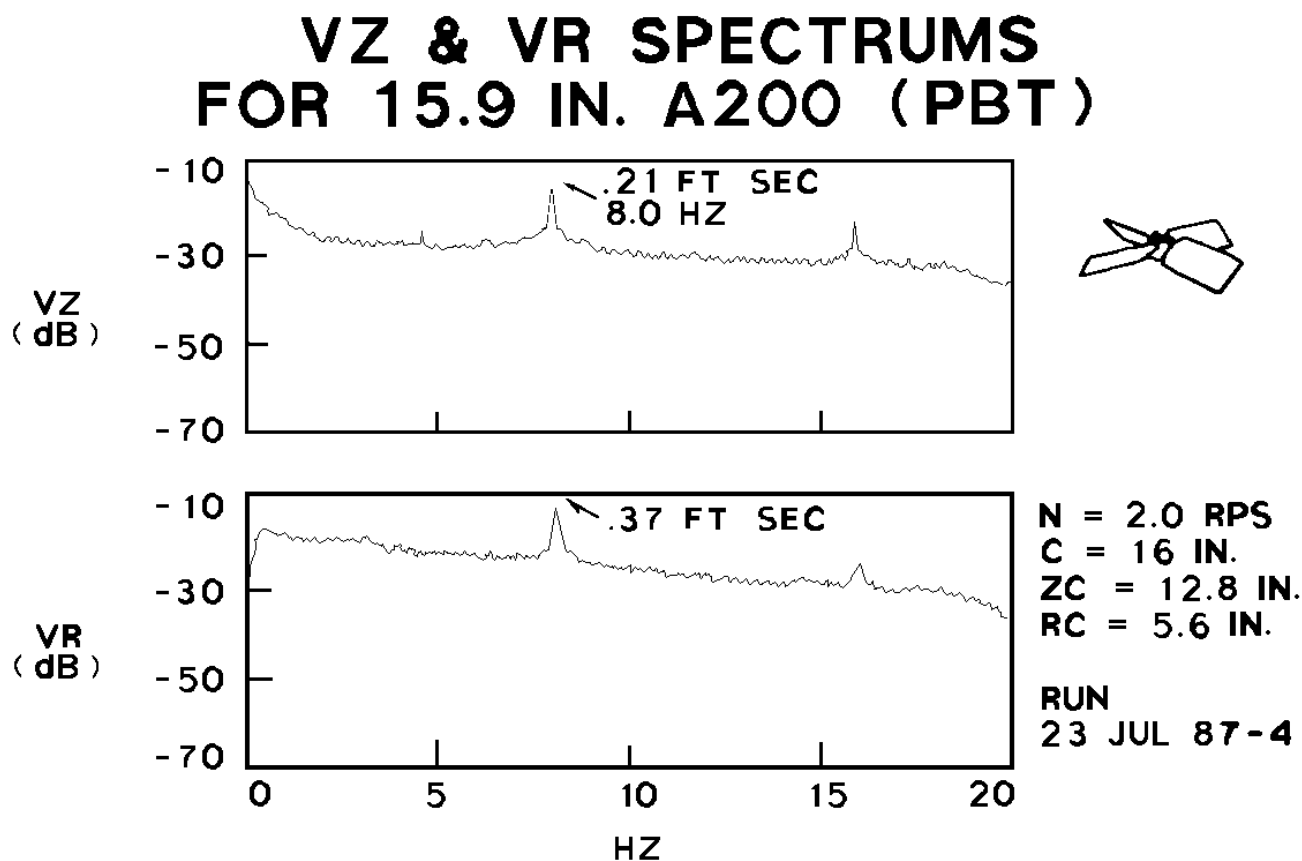


**Figure 41.4** Radial-flow Rushton turbine (R100). (Courtesy of LIGHTNIN.)



The velocity spectra in the axial direction for the axial-flow impeller A200 are shown in [Fig. 41.5](#). A decibel correlation has been used in this figure because of its well-known applicability in mathematical modeling as well as the practicality of putting many orders of magnitude of data in a reasonably sized chart. Other spectra of importance are the power spectra (the square of the velocity) and the Reynolds stress (the product of the  $R$  and  $Z$  velocity components), which is a measure of the momentum at a point.

**Figure 41.5** Typical velocity spectrum as a function of fluctuation frequency. (Courtesy of LIGHTNIN.)



The ultimate question is this: How do all of these phenomena apply to process design in mixing vessels? No one today is specifying mixers for industrial processes based on meeting criteria of this type. This is largely because processes are so complex that it is not possible to define the process requirements in terms of these fluid mechanics parameters. If the process results could be defined in terms of these parameters, sufficient information probably exists to permit the calculation of an approximate mixer design. It is important to continue studying fluid mechanics parameters in both mixing and pipeline reactors to establish what is required by different processes in fundamental terms.

One of the most practical recent results of these studies has been the ability to design pilot plant experiments (and, in many cases, plant-scale experiments) that can establish the sensitivity of a process to macro-scale mixing variables (as a function of power, pumping capacity, impeller diameter, impeller tip speeds, and macro-scale shear rates) in contrast to micro-scale mixing variables (which are relative to power per unit volume, rms velocity fluctuations, and some estimation of the size of the micro-scale eddies).

Another useful and interesting concept is the size of the eddies at which the power of an impeller is eventually dissipated. This concept utilizes the principles of **isotropic turbulence** developed by Komolgoroff. The calculations assume some reasonable approach to the degree of isotropic turbulence, and the estimates do give some idea as to how far down in the micro scale the power per unit volume can effectively reach. The equation is

$$L = (v^3/\varepsilon)^{1/4}$$

## 41.1 Scaleup/Scaledown

---

Two applications of scaleup frequently arise. One is building a model for pilot plant studies to develop an understanding of the process variables for an existing full-scale mixing installation. The other is taking a new process and studying it in the pilot plant to work out pertinent scaleup variables for a new mixing installation.

Because there are thousands of specific processes each year that involve mixing, there will be at least hundreds of different situations requiring a somewhat different pilot plant approach. Unfortunately, no set of rules states how to carry out studies for any specific program, but here are a few guidelines that can help one carry out a pilot plant program.

- For any given process, take a qualitative look at the possible role of fluid shear stresses. Try to consider pathways related to fluid shear stress that may affect the process. If there are none, then this extremely complex phenomenon can be dismissed and the process design can be based on such things as uniformity, circulation time, blend time, or velocity specifications. This is often the case in the blending of miscible fluids and the suspension of solids.
- If fluid shear stresses are likely to be involved in obtaining a process result, then one must qualitatively look at the scale at which the shear stresses influence the result. If the particles, bubbles, droplets, or fluid clumps are on the order of 1000  $\mu\text{m}$  or larger, the variables are macro-scale, and average velocity at a point is the predominant variable.

When macro-scale variables are involved, every geometric design variable can affect the role of shear stresses. These variables can include power, impeller speed, impeller diameter, impeller blade shape, impeller blade width or height, thickness of the material used to make the impeller, number of blades, impeller location, baffle location, and number of impellers.

Micro-scale variables are involved when the particles, droplets, baffles, or fluid clumps are on the order of 100  $\mu\text{m}$  or less. In this case, the critical parameters usually are power per unit volume, distribution of power per unit volume between the impeller and the rest of the tank, rms velocity fluctuation, energy spectra, dissipation length, the smallest micro-scale eddy size for the particular power level, and viscosity of the fluid.

- The overall circulating pattern, including the circulation time and the deviation of the circulation times, can never be neglected. No matter what else a mixer does, it must be able to circulate fluid throughout an entire vessel appropriately. If it cannot, then that mixer is not suited for the tank being considered.

## 41.2 Effect of the Circulation Time Spectrum and the Spectrum of Shear Rates on Ten Different Mixing Technologies

---

### Gas-Liquid Dispersion

The macro-scale shear rate change affects the bubble size distribution in tanks of various sizes. As processes are scaled up, the linear, superficial gas velocity tends to be higher in the larger tank. This is the major contributor to the energy input of the gas stream. If the power per unit volume put in by the mixer remains relatively constant, then small tanks have a different ratio of mixing energy to gas expansion energy, which affects the flow pattern and a variety of other fluid mechanics parameters. The large tank will tend to have a larger variation of the size distribution of bubbles than will the small tank.

This phenomenon is affected by the fact that the surface tension and viscosity varies all the way from a relatively pure liquid phase to all types of situations with dissolved chemicals, either electrolytes or nonelectrolytes, and other types of surface-active agents.

### Gas-Liquid Mass Transfer

If we are concerned only with the total volumetric mass transfer rate, then we can achieve very similar  $K_G a$  values in large tanks and in small tanks.

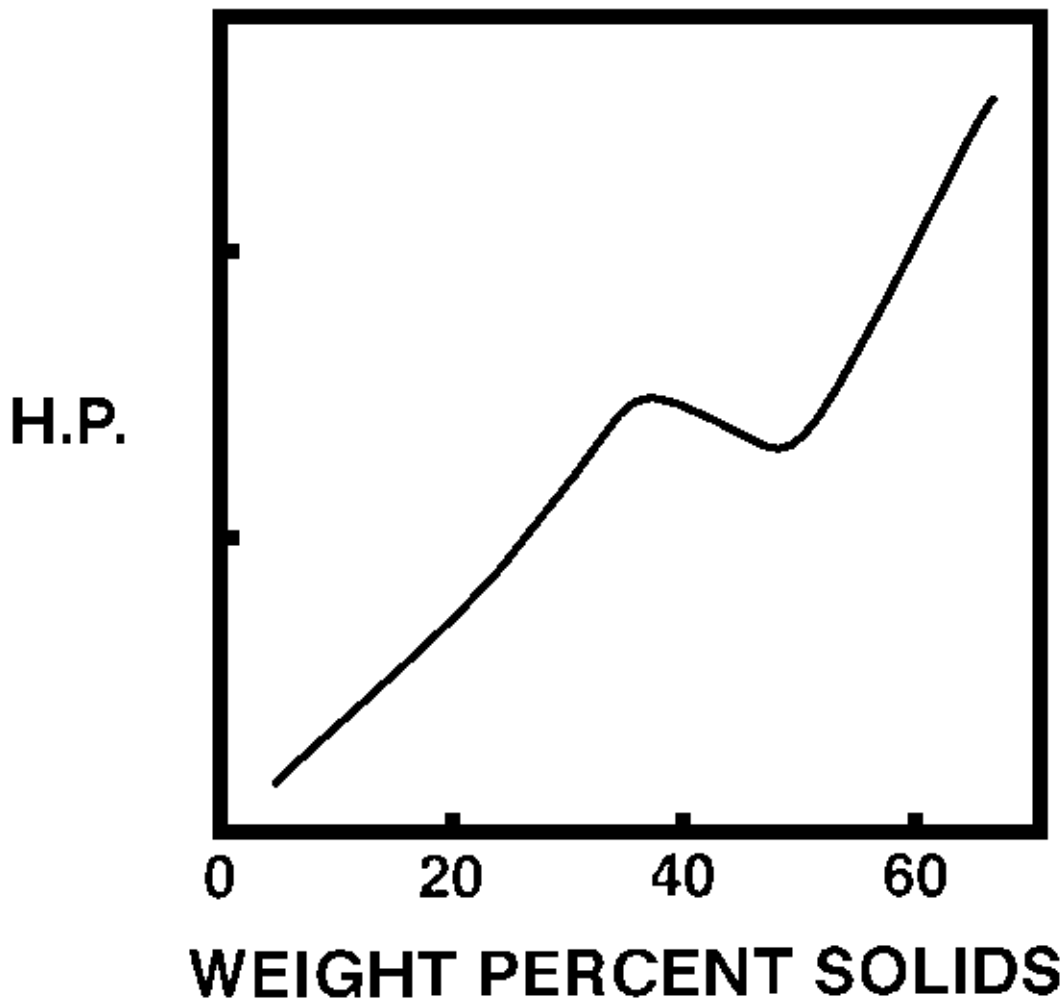
Blend time enters the picture primarily for other process steps immediately preceding or following the gas-liquid mass transfer step. Blending can play an important role in the total process, of which gas-liquid mass transfer is only one component.

### Solids Suspension and Dispersion

Solids suspension is not usually affected by blend time or shear rate changes in the relatively low to medium solids concentration in the range from zero to 40% by weight. However, as solids become more concentrated, the effect of solids concentration on the power required changes the criteria from the settling velocity of the individual particles in the mixture to the apparent viscosity of the more concentrated slurry. This means that we enter an area where the blending of non-Newtonian fluid regions, the shear rates, and circulation patterns play a marked role (see [Fig. 41.6](#)).



**Figure 41.6** Effect of percent solids by weight and power required for uniformity and fluid motion. (Courtesy of LIGHTNIN.)



The suspension of a single solid particle should depend primarily on the upward velocity at a given point and also should be affected by the uniformity of this velocity profile across the entire tank cross section. There are upward velocities in the tank, and there also must be corresponding downward velocities. In addition to the effect of the upward velocity on a settling particle, there is also the random motion of the micro-scale environment, which does not affect large particles very much but is a major factor in the concentration and uniformity of particles in the transition and micro-scale range.

Using a draft tube in the tank for solids suspension introduces another, different set of variables. There are other relationships that are very much affected by scaleup in this type of process. Different scaleup problems exist depending on whether the impeller is pumping up or down within the draft tube.

If the process involves the dispersion of solids in a liquid, then we may either be concerned with breaking up agglomerates or possibly physically breaking or shattering particles that have a low cohesive force between their components. Normally, we do not think of breaking up ionic bonds with the shear rates available in mixing machinery.

If we know the shear stress required to break up a particle, we can determine the shear rate

required of the machinery by various viscosities with the equation

$$\text{Shear stress} = \text{Viscosity} \times \text{Shear rate}$$

The shear rate available from various types of mixing and dispersion devices is known approximately, as is the range of viscosities in which they can operate. This makes the selection of the mixing equipment subject to calculation of the shear stress required for the viscosity to be used.

In the equation above, it is assumed that there is 100% transmission of the shear rate in the shear stress. However, with the slurry viscosity determined essentially by the properties of the slurry, at high slurry concentrations there is a slippage factor in which internal motion of particles in the fluids over and around each other can reduce the effective transmission of viscosity efficiencies from 100% to as low as 30%.

Animal cells in biotechnology do not normally have a tough skin like fungal cells and are very sensitive to mixing effects. Many approaches have been tried to minimize the effect of increased shear rates on scaleup, and these include encapsulating the organism in or on micro particles and conditioning cells selectively to shear rates. In addition, traditional fermentation processes have maximum shear rate requirements in which cells become progressively more and more damaged until they become motile.

## **Solid-Liquid Mass Transfer**

There is potentially a major effect of both shear rate and circulation time in these processes. The solids may be inorganic, in which case we are looking at the slip velocity of the particle and also whether we can break up agglomerates of particles, which may enhance the mass transfer. When the particles become small enough, they tend to follow the flow pattern, so the slip velocity necessary to affect the mass transfer becomes less and less available.

This shows that from the definition of off-bottom motion to complete uniformity, the effect of mixer power is much less than from going to on-bottom motion to off-bottom suspension. The initial increase in power causes more and more solids to become in active communication with the liquid and has a much greater mass transfer rate than that occurring above the power level for off-bottom suspension, in which slip velocity between the particles of fluid is the major contributor.

Since there may well be chemical or biological reactions happening on or in the solid phase, depending upon the size of the process participants, it may or may not be appropriate to consider macro- or micro-scale effects.

In the case of living organisms, their access to dissolved oxygen throughout the tank is of great concern. Large tanks in the fermentation industry often have a  $Z/T$  ratio of 2:1 to 4:1; thus, top to bottom blending can be a major factor. Some biological particles are facultative and can adapt and reestablish their metabolism at different dissolved oxygen levels. Other organisms are irreversibly destroyed by sufficient exposure to low dissolved oxygen levels.

## Liquid-Liquid Emulsions

Almost every shear rate parameter we have affects liquid-liquid emulsion formation. Some of the effects are dependent upon whether the emulsion is both dispersing and coalescing in the tank, or whether there are sufficient stabilizers present to maintain the smallest droplet size produced for long periods of time. Blend time and the standard deviation of circulation times affect the length of time it takes for a particle to be exposed to the various levels of shear work and thus the time it takes to achieve the ultimate small particle size desired.

As an aside, when a large liquid droplet is broken up by shear stress, it tends to initially elongate into a dumbbell type of shape, which determines the particle size of the two large droplets formed. Then the neck in the center of the "dumbbell" may explode or shatter. This would give a debris of particle sizes which can be quite different from the two major particles produced.

## Liquid-Liquid Extraction

If our main interest is in the total volumetric mass transfer between the liquids, the role of shear rate and blend time is relatively minor. However, if we are interested in the bubble size distribution—and we often are because that affects the settling time of an emulsion in a multistage cocurrent or countercurrent extraction process—then the change in macro and micro rates on scaleup is a major factor. Blend time and circulation time are usually not major factors on scaleup.

## Blending

If the blending process occurs between two or more fluids with relatively low viscosity such that the blending is not affected by fluid shear rates, then the difference in blend time and circulation between small and large tanks is the only factor involved. However, if the blending involves wide disparities in the density of viscosity and surface tension between the various phases, a certain level of shear rate may be required before blending can proceed to its ultimate degree of uniformity.

The role of viscosity is a major factor in going from the turbulent regime, through the transition region, into the viscous regime, and there is a change in the rate of energy dissipation discussed previously. The role of non-Newtonian viscosity is very strong since that tends to markedly change the influence of impellers and determines the appropriate geometry that is involved.

Another factor here is the relative increase in Reynolds number on scaleup. This means that we could have pilot plants running in the turbulent region as well as the plant. We could have the pilot plant running in the transition region and the plant in the turbulent, or the pilot plant could be in the viscous region while the plant is in the transition region. There is no apparent way to prevent this Reynolds number change upon scaleup. In reviewing the qualitative flow pattern in a pilot scale system, it should be realized that the flow pattern in the large tank will be at an apparently much lower viscosity and therefore at a much higher Reynolds number than is observed in the pilot plant. This means that the roles of tank shape,  $D/T$  ratio, baffles, and impeller locations can be based on different criteria in the plant size unit than in the pilot size unit under observation.

## Chemical Reactions

Chemical reactions are influenced by the uniformity of concentration both at the feed point and in the rest of the tank and can be markedly affected by changes in overall blend time and circulation time as well as the micro-scale environment. It is possible to keep the ratio between the power per unit volume at the impeller and that in the rest of the tank relatively similar on scaleup, but much detail needs to be considered regarding the reaction conditions, particularly where selectivity is involved. This means that reactions can take different paths depending upon chemistry and fluid mechanics, and this is a major consideration in what should be examined. The method of introducing the reagent stream can be projected in several different ways depending upon the geometry of the impeller and the feed system.

## Fluid Motion

Sometimes the specification is purely in terms of pumping capacity. Obviously, the change in volume and velocity relationships depends upon the size of the two- and three-dimensional area or volume involved. The impeller flow is treated in a head/flow concept, and the head required for various types of mixing systems can be calculated or estimated.

## Heat Transfer

In general, the fluid mechanics of the film on the mixer side of the heat transfer surface is a function of what happens at that surface rather than the fluid mechanics around the impeller zone. The impeller provides flow largely across and adjacent to the heat transfer surface, and that is the major consideration of the heat transfer result obtained. Many of the correlations are in terms of traditional dimensionless groups in heat transfer, while the impeller performance is often expressed as the impeller Reynolds number.

## 41.3 Computational Fluid Dynamics

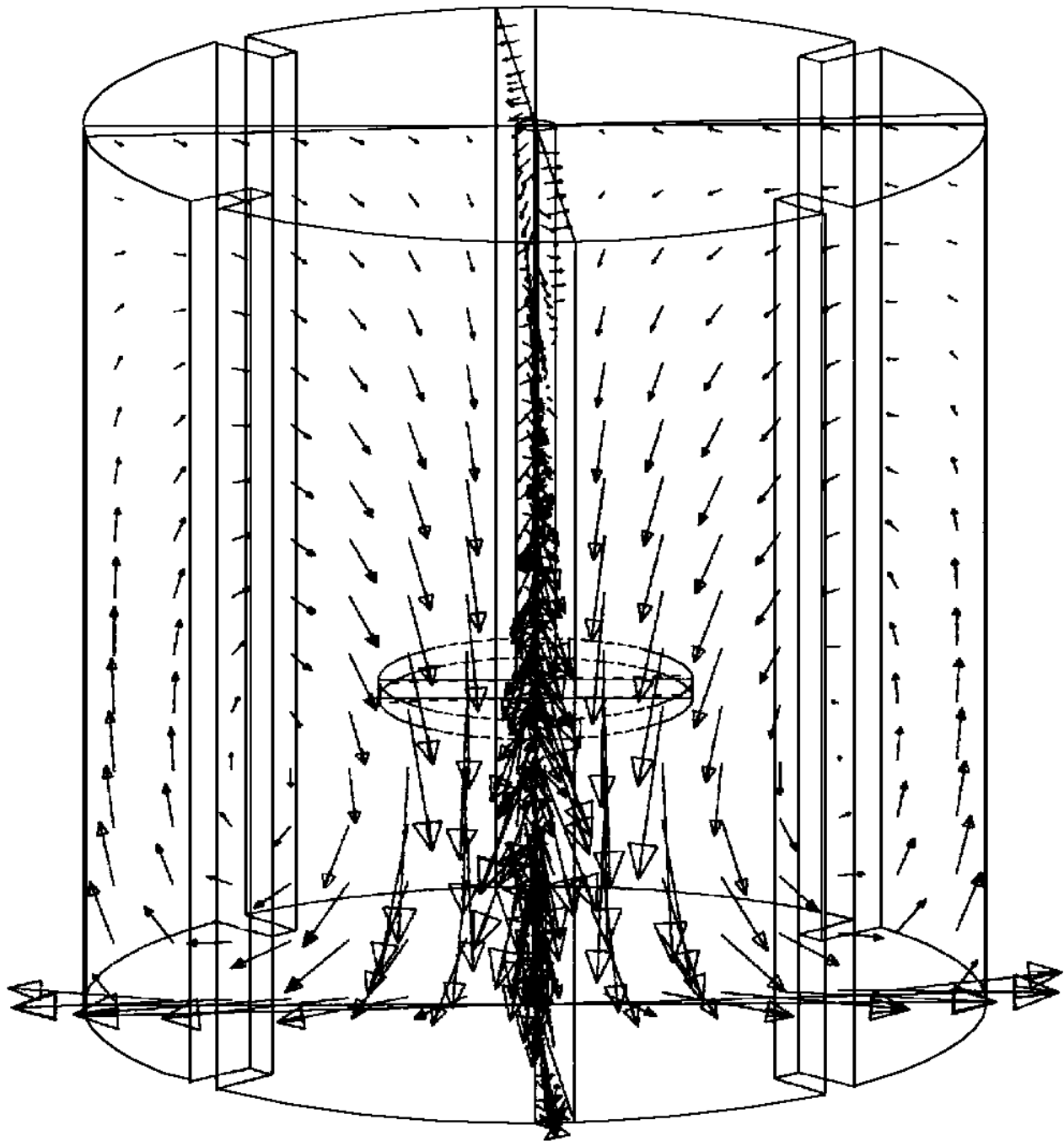
---

There are several software programs available to model flow patterns of mixing tanks. They allow the prediction of flow patterns based on certain boundary conditions. The most reliable models use accurate fluid mechanics data generated for the impellers in question and a reasonable number of modeling cells to give the overall tank flow pattern. These flow patterns can give velocities, streamlines, and localized kinetic energy values for the systems. Their main use at the present time is in examining the effect of changes in mixing variables based on adjustments to the mixing process. These programs can model velocity, shear rates, and kinetic energy but probably cannot adapt to the chemistry of diffusion or mass transfer kinetics of actual industrial processes at the present time.

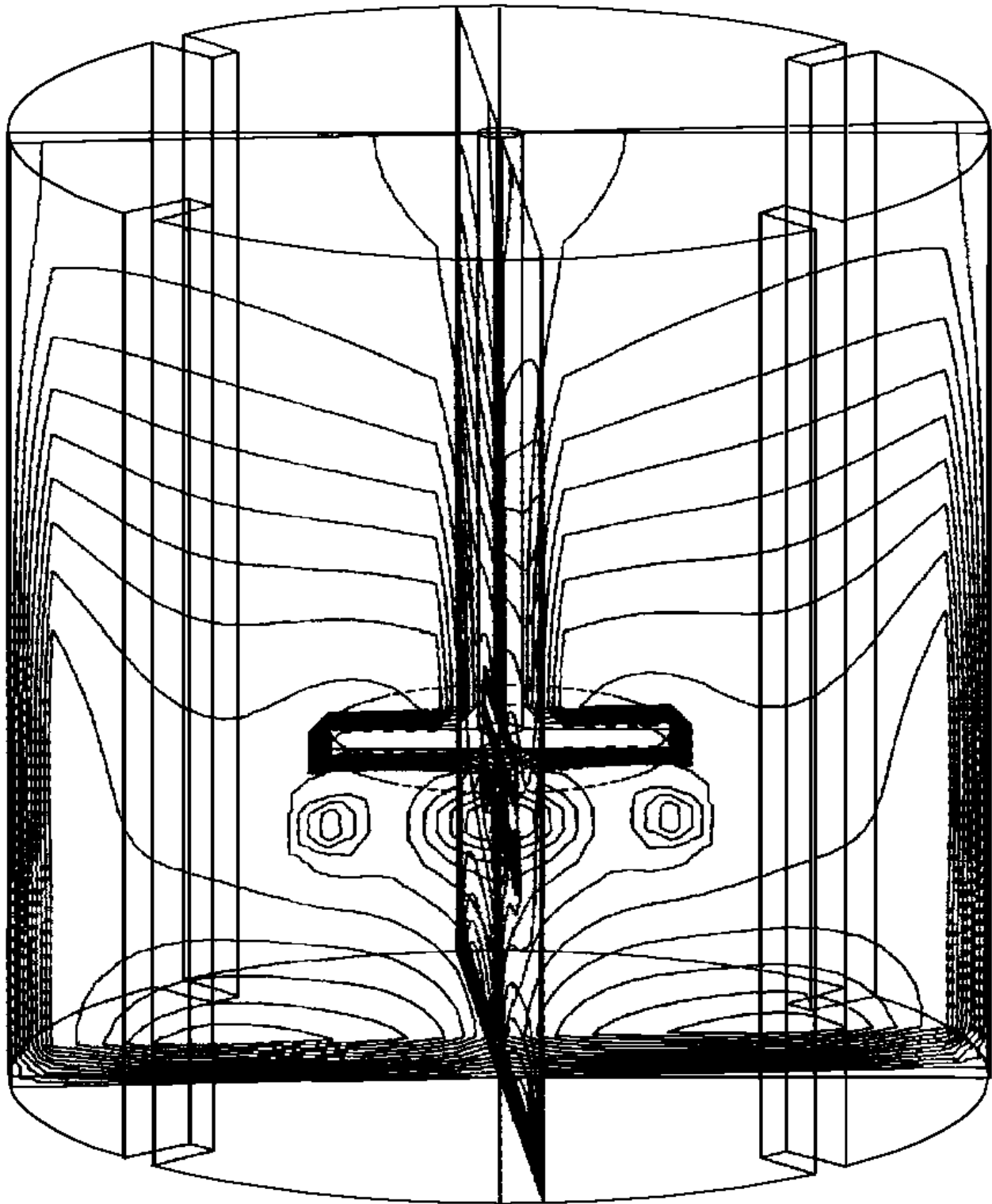
Relatively uncomplicated transparent tank studies using tracer fluids or particles can also give a feel for the overall flow pattern. The time and expense of calculating these flow patterns with computational fluid dynamics should be considered in relation to their applicability to an actual industrial process. The future of computational fluid dynamics looks very encouraging, and a reasonable amount of time and effort placed in this regard can yield immediate results as well as the potential for future process evaluation.

Figures 41.7–41.9 show some approaches. Figure 41.7 shows velocity vectors for an A310 impeller. Figure 41.8 shows contours of kinetic energy of turbulence. Figure 41.9 uses a particle trajectory approach with neutral buoyancy particles.

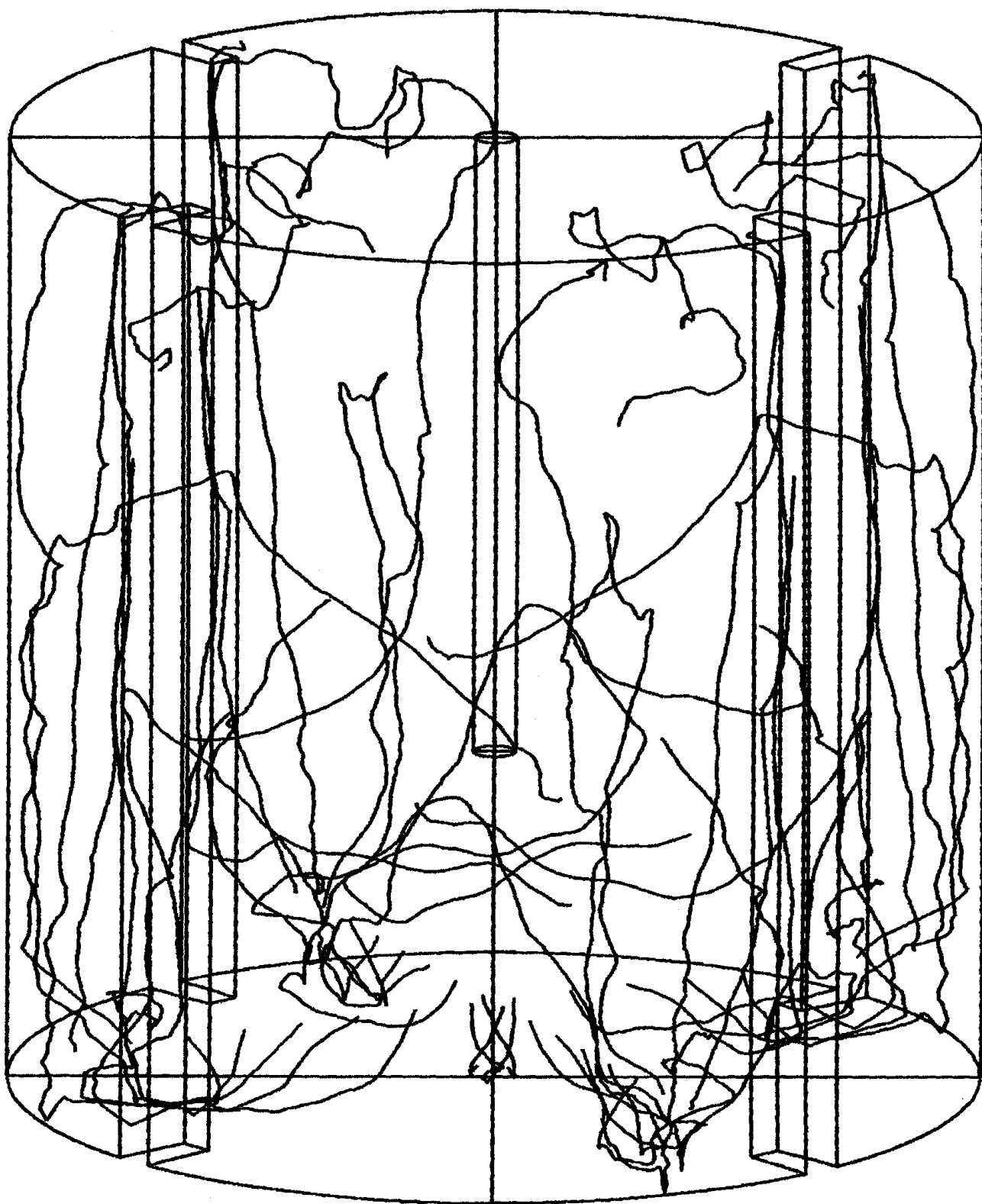
**Figure 41.7** Typical velocity pattern for a three-dimensional model using computational fluid dynamics for an axial flow impeller (A310). (Courtesy of LIGHTNIN.)



**Figure 41.8** Typical contours of kinetic energy of turbulence using a three-dimensional model with computational fluid dynamics for an axial flow impeller (A310). (Courtesy of LIGHTNIN.)



**Figure 41.9** Typical particle trajectory using an axial flow impeller (A310) with a 100-micron particle using computational fluid dynamics. (Courtesy of LIGHTNIN.)



Numerical fluid mechanics can define many of the fluid mechanics parameters for an overall reactor system. Many of the models break the mixing tank up into small microcells. Suitable material and mass transfer balances between these cells throughout the reactor are then made. This can involve massive computational requirements. Programs are available that can give reasonably acceptable models of experimental data in mixing vessels. Modeling the three-dimensional aspect of a flow pattern in a mixing tank can require a large amount of computing power.

## Defining Terms

**Isotropic turbulence:** Fluid shear rate is a velocity gradient that results in shear stress which can break up, disperse, or otherwise affect particles.

**Macro scale:** Any process governed by large particles on the order of 1000  $\mu\text{m}$  or more.

**Micro scale:** Any process governed by small particles on the order of less than 100  $\mu\text{m}$ .

## References

Levich, V. 1962. *Physio-Chemical Hydrodynamics*. Prentice-Hall, Englewood Cliffs, NJ.

Middleton, J. C. 1989. *Proceedings of the Third European Conference on Mixing*, p. 15–36. BHRA, Cranfield, England.

Neinow, A. W., Buckland, B., and Weetman, R. J. 1989. *Mixing XII Research Conference*. Potosi, MO.

Oldshue, J. Y. 1989. Mixing '89. *Chem. Eng. Prog.* p. 33–42.

Oldshue, J. Y., Post, T. A., and Weetman, R. J. 1988. Comparison of mass transfer characteristics of radial and axial flow impellers. *Proceedings of the Sixth European Conference on Mixing*. BHRA, Cranfield, England.

## Nomenclature

$N$	impeller speed
$D$	impeller diameter
$T$	tank diameter
$Z$	liquid level
$P/V$	power per unit volume
$SR$	Solidity ratio, obtained by dividing the projected area of the impeller blades by the area of a disk circumscribing the impeller blades
$N_P$	power number
$H$	velocity head, $v^2/2g$
$P$	power
$L$	length scale
$v$	fluid velocity
$v'$	fluid velocity fluctuation
$\bar{v}$	average fluid velocity
$K_G a$	gas-liquid mass transfer coefficient



$K_L a$	liquid-liquid mass transfer coefficient
$k_s$	liquid-solid mass transfer coefficient
$L$	size of microscale eddy
$\varepsilon$	energy dissipation rate
$\dot{\nu}$	kinematic viscosity
$\nu$	dynamic viscosity

## Further Information

- Harnby, N., Edwards, M. F., and Neinow, A. W., eds. 1986. *Mixing in the Process Industries*. Butterworth, Stoneham, MA.
- Lo, T. C., Baird, M. H. I., and Hanson, C. 1983. *Handbook of Solvent Extraction*. John Wiley & Sons, New York.
- McDonough, R. J. 1992. *Mixing for the Process Industries*. Van Nostrand Reinhold, New York.
- Nagata, S. 1975. *Mixing: Principles and Applications*. Kodansha Ltd., Tokyo; John Wiley & Sons, New York.
- Oldshue, J. Y. 1983. *Fluid Mixing Technology*. McGraw-Hill, New York.
- Tatterson, G. B. 1991. *Fluid Mixing and Gas Dispersion in Agitated Tanks*. McGraw-Hill, New York.
- Uhl, V. W. and Gray, J. B., eds. 1966, 1986. *Mixing*. Vols. I and II, Academic Press, New York; vol. III, Academic Press, Orlando, FL.
- Ulbrecht, J. J. and Paterson, G. K., eds. 1985. *Mixing of Liquids by Mechanical Agitation*. Gordon & Breach Science Publishers, New York.

## Proceedings

- Fluid Mechanics of Mixing*, ed. R. King. Kluwer Academic Publishers, Dordrecht, Netherlands, 1992.
- Fluid Mixing*, vol. I. Inst. Chem. Eng. Symp., Ser. No. 64 (Bradford, England). Institute of Chemical Engineers, Rugby, England, 1984.
- Mixing—Theory Related to Practice*. AIChE, Inst. Chem. Eng. Symp. Ser. No. 10 (London). AIChE and Institute of Chemical Engineers, London, 1965.
- Proceedings of the First European Conf. on Mixing*, ed. N. G. Coles. BHRA Fluid Eng., Cranfield, England, 1974.
- Proceedings of the Second European Conference on Mixing*, ed. H. S. Stephens and J. A. Clarke. BHRA Fluid Eng., Cranfield, England, 1977.
- Proceedings of the Third European Conference on Mixing*, ed. H. S. Stephens and C. A. Stapleton. BHRA Fluid Eng., Cranfield, England, 1979.
- Proceedings of the Fourth European Conference on Mixing*, ed. H. S. Stephens and D. H. Goodes. BHRA Fluid Eng., Cranfield, England, 1982.
- Proceedings of the Fifth European Conference on Mixing*, ed. S. Stanbury. BHRA Fluid Eng., Cranfield, England, 1985.
- Proceedings of the Sixth European Conference on Mixing*. BHRA Fluid Eng., Cranfield, England, 1988.
- Process Mixing: Chemical and Biochemical Applications*, ed. G. B. Tatterson and R. V. Calabrese. AIChE Symp. Ser. No. 286, 1992.

Sherif, S. A. "Fluid Measurements"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Fluid Measurements

---

## 42.1 Fundamental Principles

## 42.2 Basic Equations

Differential Pressure Meters • Thermal Anemometers • Laser Doppler Anemometers • Volume Flow Measurements

**S. A. Sherif**

*University of Florida, Gainesville*

The subject of fluid measurements covers a broad spectrum of applications. Examples include wind-tunnel experiments, turbomachinery studies, water tunnel and flume investigations, erosion studies, and meteorological research, to name a few. Many wind-tunnel investigations call for determining forces on scale models of large systems. **Similarity analysis** is then performed to permit generalization of the experimental results obtained on the laboratory model. Techniques for measuring fluid flow parameters are extremely diverse and include thermal anemometers, laser velocimeters, volume flow devices, flow visualization by direction injection, and optical diagnostics. Measurements can be carried out in liquids and gases, incompressible and compressible fluids, two-phase and single-phase flows, and Newtonian and **non-Newtonian fluids**. Types of quantities measured vary greatly and include velocity, temperature, turbulence, vorticity, Reynolds stresses, turbulent heat flux, higher-order turbulence moments, volume and mass flow rates, and differential pressure. Topics covered in this chapter will be limited to differential pressure measurements, thermal anemometry, laser velocimetry, and volume flow measurements. Coverage of these topics will take the form of providing a summary of fundamental principles followed by a summary of the basic equations used.

## 42.1 Fundamental Principles

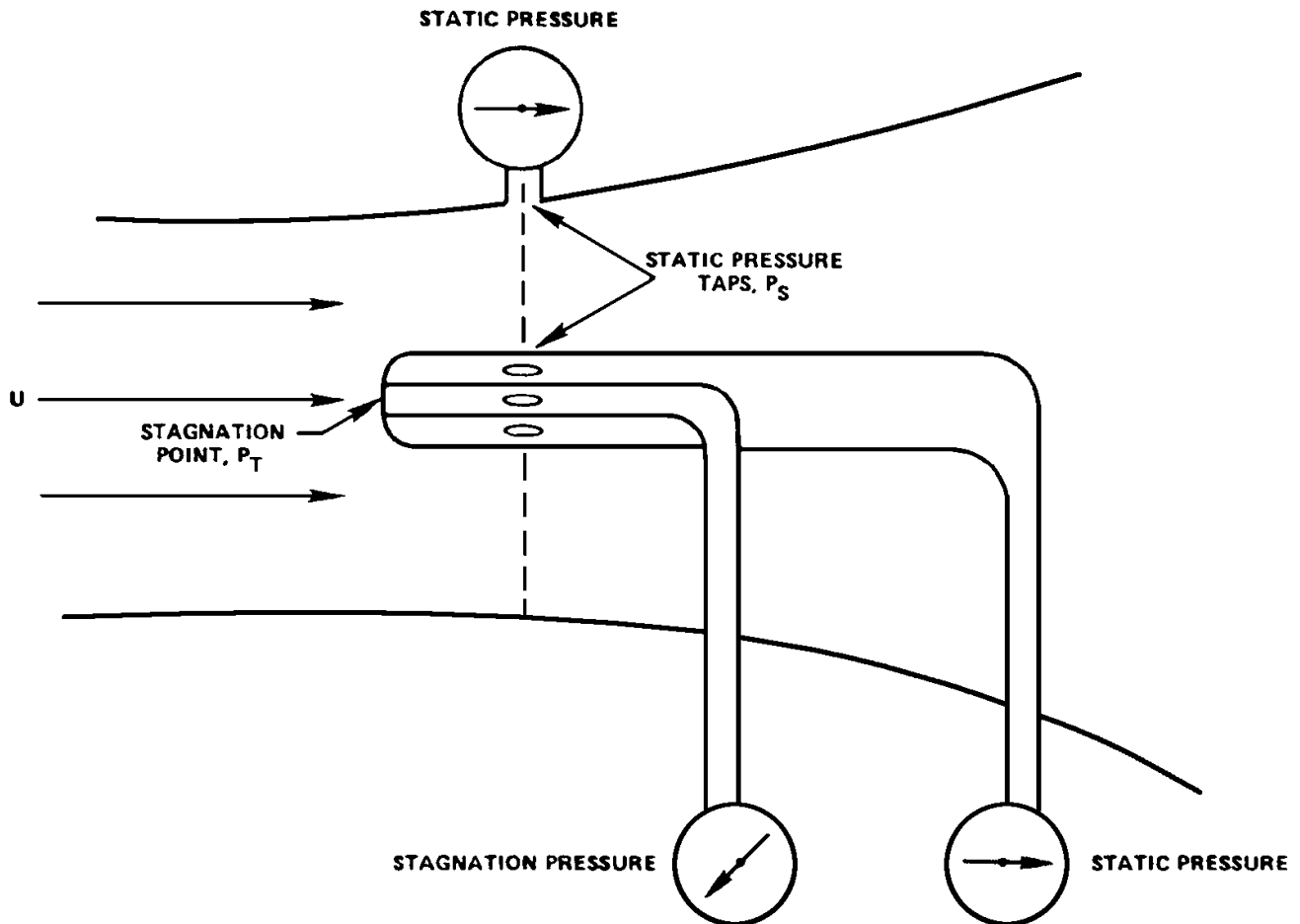
---

The purpose of this section is to provide an overview of the fundamental principles involved in fluid measurements employing the most common and widely used techniques. These techniques include using differential pressure-based instruments, hot-wire and hot-film anemometers, laser Doppler velocimeters, and volume flow devices. These topics will be discussed in some detail in this section.

Differential pressures may be thought of as differences between time-averaged pressures at two points in a fluid flow or between a time-averaged and an instantaneous value of pressure evaluated at a point in the flow [Blake, 1983]. This type of measurement provides an alternative to **thermal anemometry** for determining velocity magnitude and direction as well as turbulence intensity of a

fluid flow. The Pitot-static tube is one of the most reliable differential pressure probes for flow measurement (see Fig. 42.1). Special forms of Preston tubes have been used for measurement of wall shear stress in boundary layers of smooth walls. Preston tubes are hypodermic needles that respond to the mean velocity profile in the vicinity of the wall.

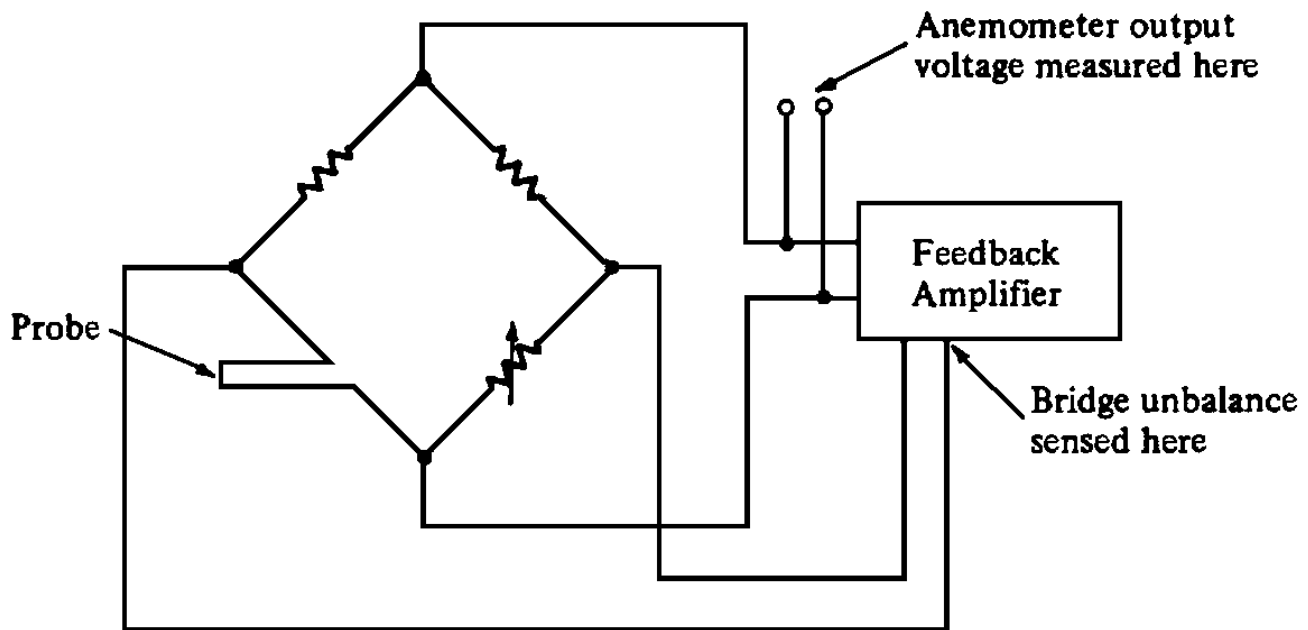
**Figure 42.1** Typical installation of a Pitot-static tube in a duct. (Source: Blake, W. K. 1983. Differential pressure measurement. In *Fluid Mechanics Measurements*, ed. R. J. Goldstein, pp. 61–97. Hemisphere, Washington, DC. Reproduced with permission. All rights reserved.)



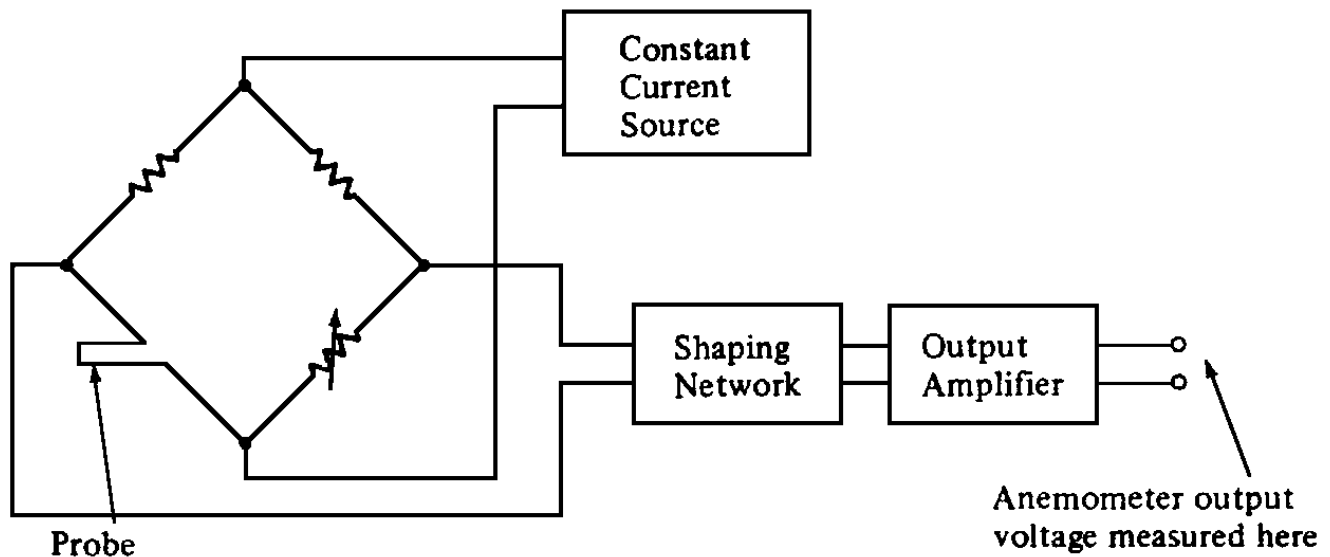
*Thermal anemometry* refers to the use of a wire or film sensor for velocity, temperature, concentration, and turbulence measurements. This technique started in the late 1800s in the form of employing homemade constant-current anemometers (CCA) for velocity measurement. Three categories of anemometers now exist. The constant-temperature anemometer (CTA) supplies a sensor with heating current that changes with the flow velocity to maintain constant sensor resistance. This type is primarily used for velocity, turbulence, Reynolds stress, and vorticity measurements. The constant-current anemometer, on the other hand, supplies a constant heating current to the sensor. This can either be used for velocity measurements or for temperature and

temperature fluctuation measurements. This choice is dictated by the magnitude of the probe sensor current, where the probe sensitivity to velocity fluctuations diminishes for low values of current. The third type of anemometer is the pulsed wire anemometer, which is capable of measuring the fluid velocity by momentarily heating the sensor. This causes heating of the fluid around the sensor and a subsequent convection of that fluid segment downstream to a second wire that acts as a temperature sensor. The time of flight of the heated fluid segment is inversely proportional to the fluid velocity. Typical block diagrams representing the constant-temperature and constant-current anemometers are shown in Figs. 42.2 and 42.3, respectively. Thermal anemometers are always connected to a probe via a probe cable. The sensor of a typical probe can either be made of wire or film. Wire probes are made of tungsten or platinum and are about 1 mm long and  $5\text{ }\mu\text{m}$  in diameter. Film probes, on the other hand, are made of nickel or platinum deposited in a thin layer onto a backing material (such as quartz) and connected to the anemometer employing leads attached to the ends of the film. A thin protective coating is usually deposited over the film to prevent damage by chemical reaction or abrasion. Typical thicknesses of film probes are  $70\text{ }\mu\text{m}$ .

**Figure 42.2** Block diagram of a constant-temperature anemometer.

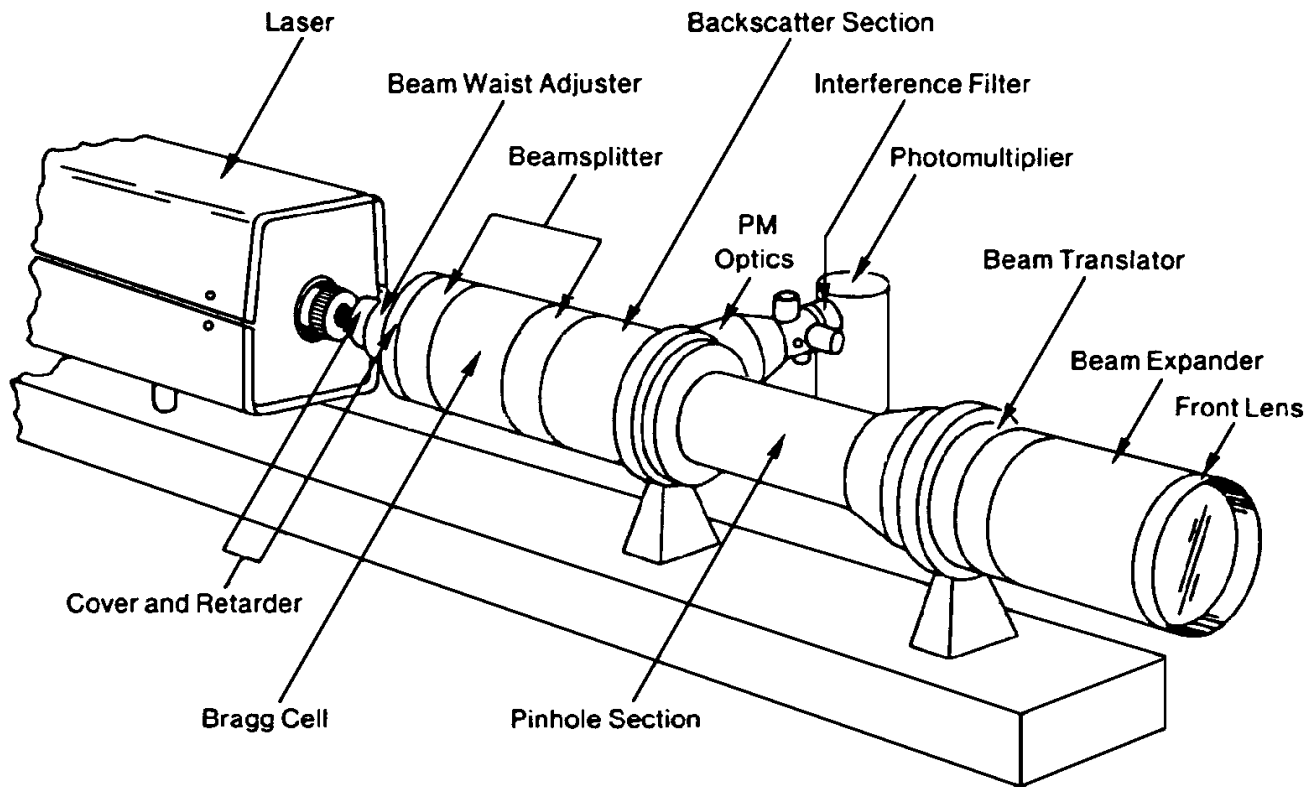


**Figure 42.3** Block diagram of a constant-current anemometer.



Laser Doppler anemometry (LDA), or laser Doppler velocimetry (LDV), deals with measuring fluid velocities and higher-order turbulence quantities by detecting the Doppler frequency shift of laser light that has been scattered by small particles moving with the fluid [Adrian, 1983]. Three different types of LDA optical systems exist. These are the reference-beam system, the dual-beam system, and the dual-scatter system. The dual-beam LDA produces two types of signals—coherent and incoherent. The coherent signal occurs when at least two particles simultaneously reside in the measurement volume. The incoherent signal occurs when a single particle scatters two light waves, one from each illuminating beam. The reference-beam LDA mixes light scattered from an illuminating beam with a reference beam to detect the frequency difference. Both the reference-beam LDA and the dual-beam LDA have the disadvantage of having to satisfy a coherent aperture condition. The amplitude of the reference-beam Doppler signal is proportional to the square root of the reference-beam power, thus allowing an unlimited increase of the signal amplitude by the simple expedient of increasing the power. This feature is particularly useful when the scattered light flux is small compared to a background radiation level. A typical LDA system consists of a laser, a **beam waist adjuster**, a beamsplitter, a **bragg cell**, a backscatter section, a **photomultiplier tube**, photomultiplier optics, an interference filter, a **pinhole section**, a **beam translator**, a beam expander, and a front lens (see Fig. 42.4). Other components include a **signal processor** and a **frequency tracker** or a **frequency counter**.

**Figure 42.4** Typical laser Doppler anemometer system. (Courtesy of Dantec Measurement Technology A/S.)

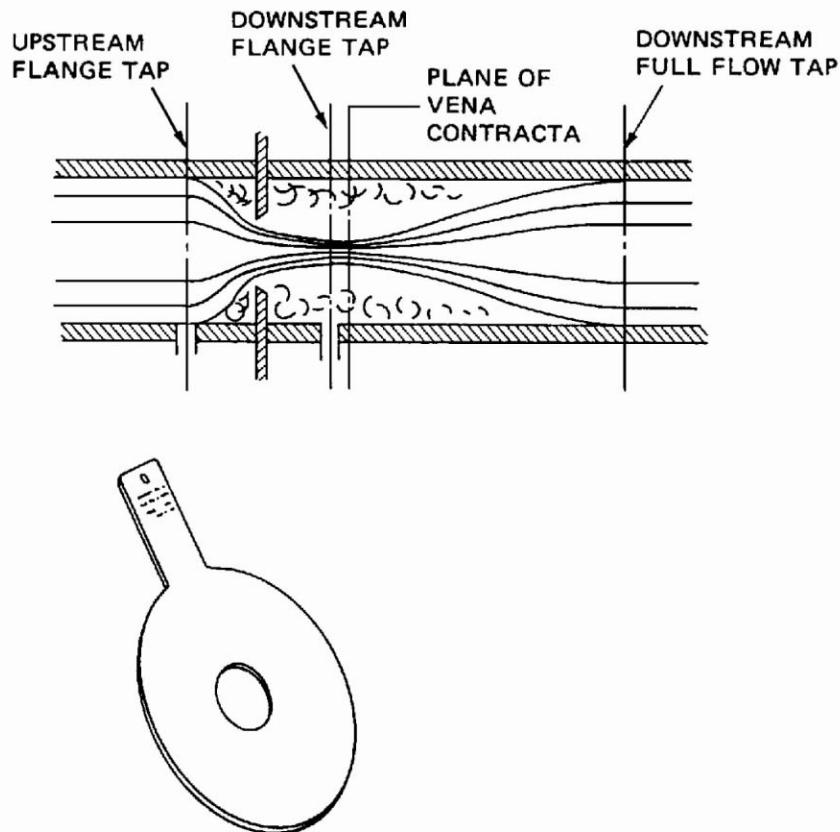


Fluid meters may be classified into those that determine fluid quantity and those that determine fluid flow rate. Quantity meters may be classified as weighing meters (such as weighers or tilting traps) or volumetric meters (such as calibrated tanks, reciprocating pistons, rotating disks and pistons, sliding and rotating vanes, gear and lobed impellers, bellows, and liquid sealed drums). Rate meters, on the other hand, can be classified as differential pressure meters (such as orifice, venturi, nozzle, centrifugal, Pitot-tube, or linear resistance meters), momentum meters (such as turbine, propeller, or cup anemometers), variable-area meters (such as gate, cones, floats-in-tubes, or slotted cylinder and piston meters), force meters (such as target or hydrometric pendulum meters), thermal meters (such as hot-wire and hot-film anemometers), fluid surface height or head meters (such as weirs and flumes), and miscellaneous meters (such as electromagnetic, tracers, acoustic, vortex-shedding, laser, and Coriolis meters) [Mattingly, 1983].

A typical orifice meter (see Fig. 42.5) consists of four parts: (1) the upstream section including flow-conditioning devices; (2) the orifice fixture and plate assembly; (3) the downstream meter-tube section; and (4) the secondary element (not shown). A venturi tube typically has a shape that closely approximates the streamline pattern of the flow through a reduced cross-sectional area (see Fig. 42.6). Elbow meters are considered nonintrusive since they do not interfere with the flow pattern in the pipe and do not introduce structural elements in the flow [Mattingly, 1983]. Laminar flow meters are based on establishing laminar flow between the pressure taps and using the laminar flow rate relationships through a tube of a known area and with a specified pressure drop across its length. Turbine meters enable the fluid flow to spin a propeller wheel whose angular speed is related to the average flow rate in the duct. The angular speed is

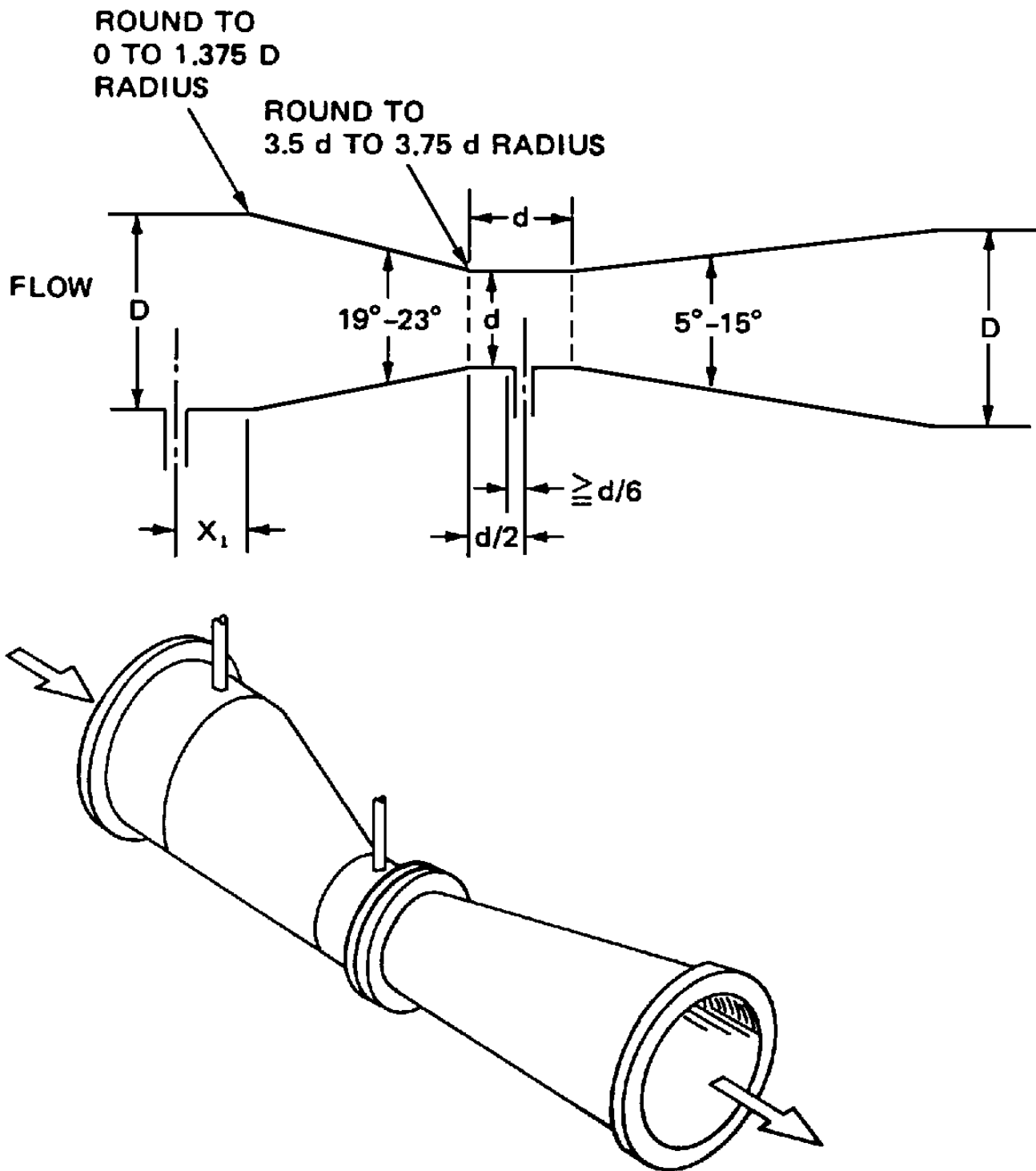
typically detected by the passage of the blade tips past a coil pickup in the pipe. Rotameters are vertically installed devices that operate by balancing the upward fluid drag on the float with the weight of the float in the upwardly diverging tube (see Fig. 42.7). Appropriate choice of the configuration of the metering tube can allow the position of the float to be linearly proportional to the flow rate. Target meters (see Fig. 42.8) operate on the principle that the average flow rate in a pipe flow is related to the fluid drag on a disk supported in the pipe. The fluid drag can be measured using secondary devices such as strain gauges and fluid-activated bellows. Target meters are particularly useful in flow metering applications involving dirty fluids so long as the suspended particles do not alter the critical geometrical arrangement. Thermal flow meters operate on the principle of sensing the increase in fluid temperature between two thermometers placed in the flow when heat is added between the thermometers. Thermal flow meters may be made to be nonintrusive and are also capable of operating on the basis of cooling rather than heating.

**Figure 42.5** Orifice meter. (Source: Mattingly, G. E. 1983. Volume flow measurements. In *Fluid Mechanics Measurements*, ed. R. J. Goldstein, pp. 245–306. Hemisphere, Washington, DC. Reproduced with permission. All rights reserved.)

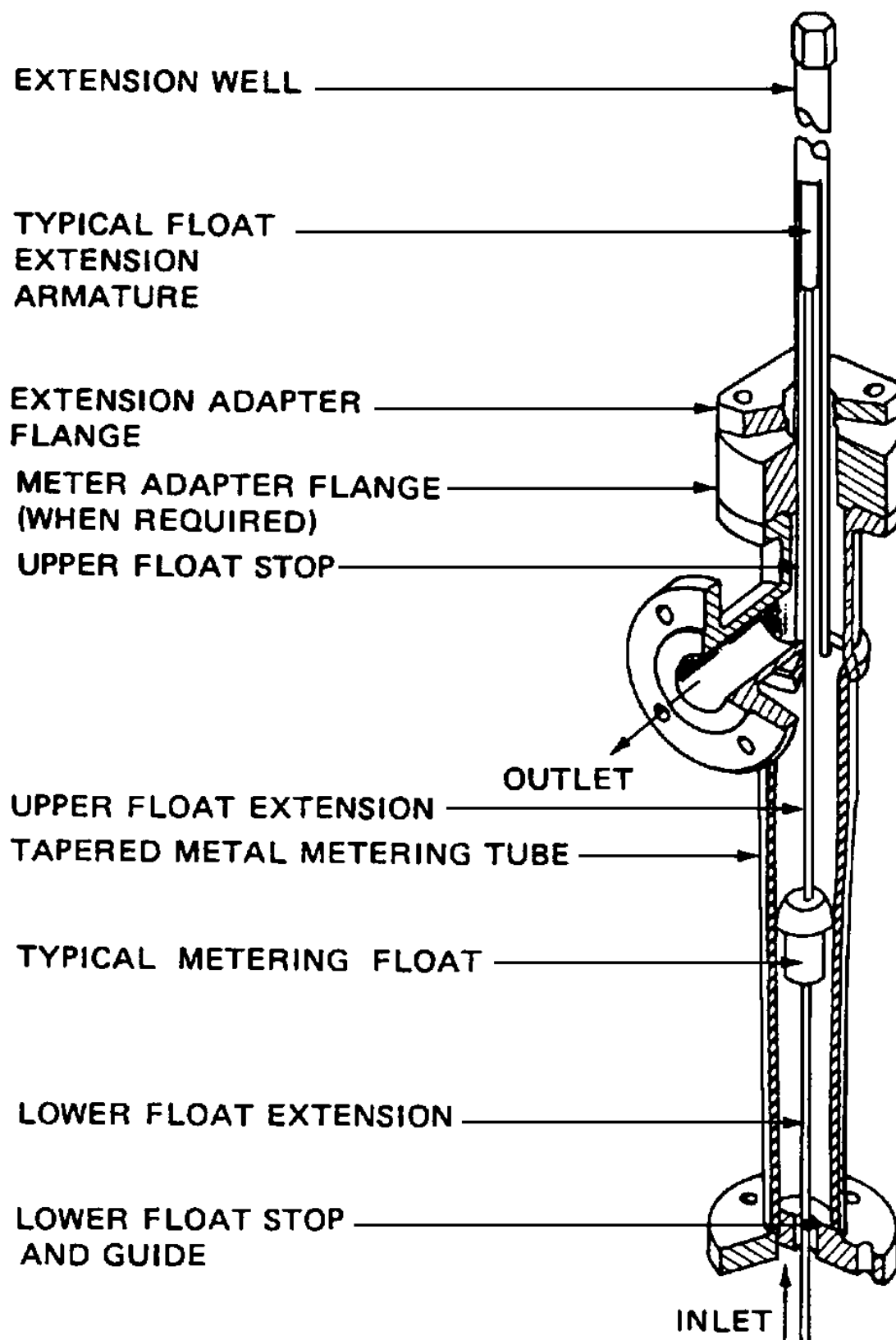




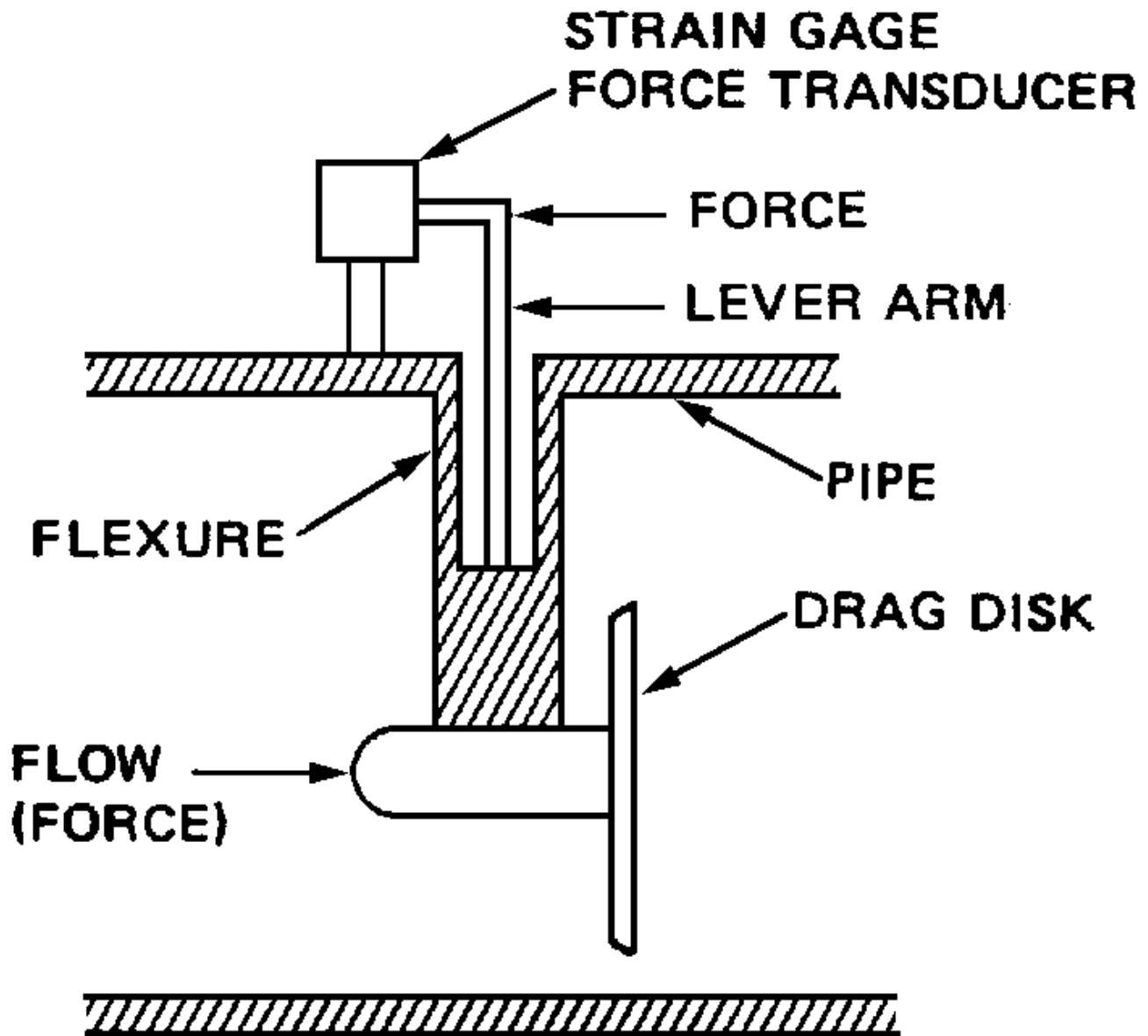
**Figure 42.6** Venturi tube. (Source: Mattingly, G. E. 1983. Volume flow measurements. In *Fluid Mechanics Measurements*, ed. R. J. Goldstein, pp. 245–306. Hemisphere, Washington, DC. Reproduced with permission. All rights reserved.)



**Figure 42.7** Rotameter. (Source: Mattingly, G. E. 1983. Volume flow measurements. In *Fluid Mechanics Measurements*, ed. R. J. Goldstein, pp. 245–306. Hemisphere, Washington, DC. Reproduced with permission. All rights reserved.)



**Figure 42.8** Target meter. (Source: Mattingly, G. E. 1983. Volume flow measurements. In *Fluid Mechanics Measurements*, ed. R. J. Goldstein, pp. 245–306. Hemisphere, Washington, DC. Reproduced with permission. All rights reserved.)



## 42.2 Basic Equations

### Differential Pressure Meters

The velocity in a one-dimensional laminar flow using a Pitot tube may be expressed as

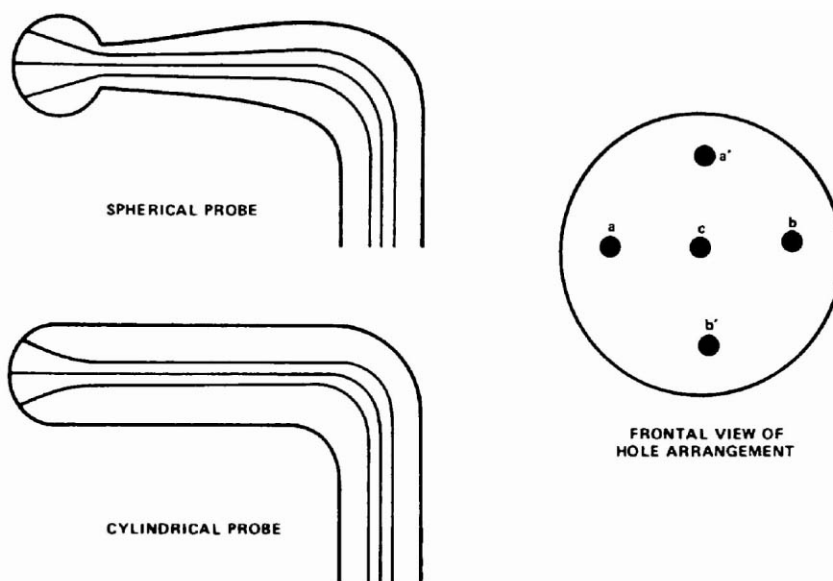
$$U = \sqrt{\frac{2(p_T - p_s)}{\rho}} \quad (42.1)$$

The most widely used method for measuring mean velocity in multidimensional flows is a five-hole pressure probe, which is a streamlined axisymmetric body that points into the flow (see Fig. 42.9). The vector decomposition of the flow velocity  $U^*$ , which is incident on the five pressure taps  $a$ ,  $b$ ,  $c$ ,  $a'$ , and  $b'$  of a spherical probe, is shown in Fig. 42.10. Pien [1958] has shown the following:

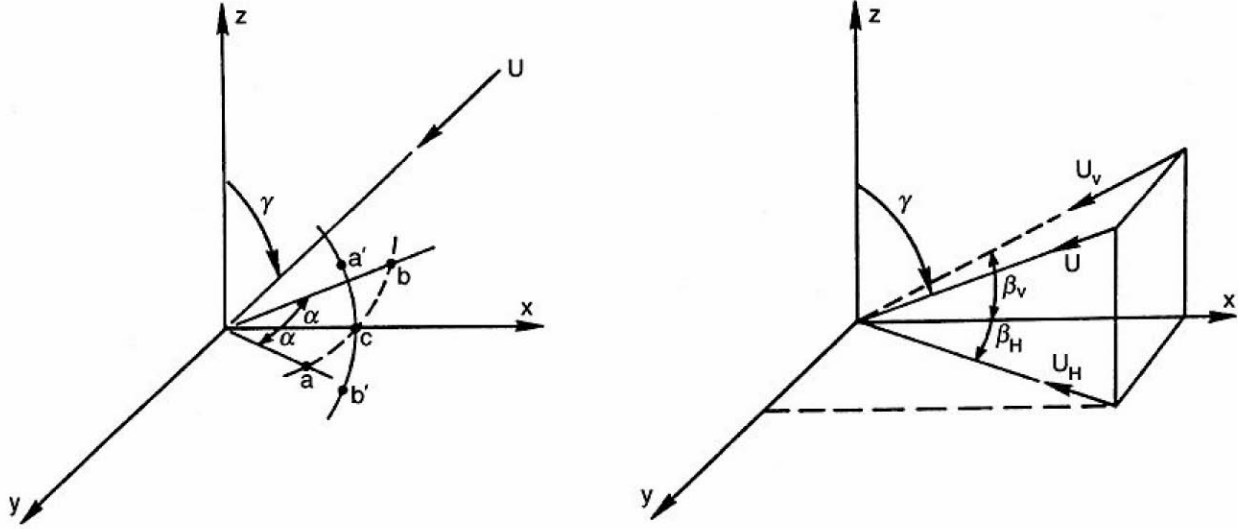
$$\frac{p_a - p_b}{2p_c - p_a - p_b} = \frac{\sin 2\alpha}{1 - \cos 2\alpha} \tan 2\beta_h \quad (42.2)$$

$$\frac{p_a - p_b}{\frac{1}{2}\rho V_h^2} = \frac{9}{4} \sin 2\alpha \sin 2\beta_h \quad (42.3)$$

**Figure 42.9** Some five-hole Pitot-tube geometries. (Source: Blake, W. K. 1983. Differential pressure measurement. In *Fluid Mechanics Measurements*, ed. R. J. Goldstein, pp. 61–97. Hemisphere, Washington, DC. Reproduced with permission. All rights reserved.)



**Figure 42.10** Vector decomposition of hole geometry and flow direction for spherical-head Pitot tube. (Source: Blake, W. K. 1983. Differential pressure measurement. In *Fluid Mechanics Measurements*, ed. R. J. Goldstein, pp. 61–97. Hemisphere, Washington, DC. Reproduced with permission. All rights reserved.)



## Thermal Anemometers

The overheat ratio and resistance difference ratio may be expressed as

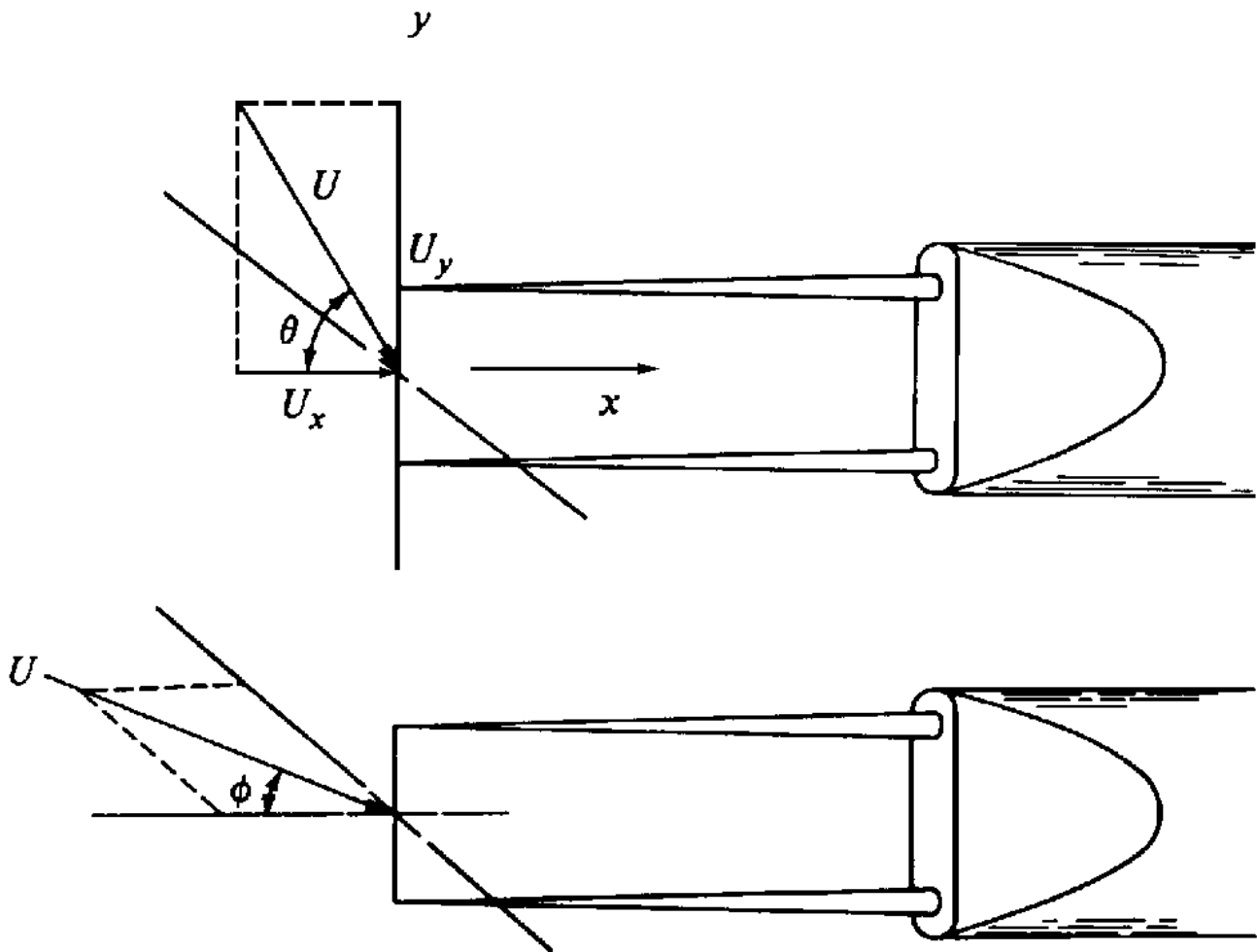
$$a_1 = \frac{R_s}{R_f} \quad \text{and} \quad a_2 = \frac{R_s - R_f}{R_f} \quad (42.4)$$

The yaw, pitch, and roll sensitivities (Fig. 42.11) may be expressed by the partial derivatives  $\partial E / \partial \theta$ ,  $\partial E / \partial \phi$ , and  $\partial E / \partial \psi$ , respectively. The effective cooling velocity, on the other hand, may be expressed as either of the following forms:

$$U_{\text{eff}}^2 = U_x^2 + k^2 U_y^2 + h^2 U_z^2 \quad (42.5)$$

$$U_{\text{eff}}^2 = U^2 (\cos^2 \theta \cos^2 \phi + k^2 \sin^2 \theta \cos^2 \phi + h^2 \sin^2 \phi) \quad (42.6)$$

**Figure 42.11** Yaw and pitch angles of a standard hot-wire probe.



The resistance-temperature relationships may be expressed by

$$R = \rho_r \frac{l}{A} \quad \text{and} \quad (42.7)$$

$$R_s = R_o[1 + \alpha(T_s - T_o) + \alpha_1(T_s - T_o)^2 + \dots]$$

where  $\alpha = 3.5 \cdot 10^{-3} \text{ } ^\circ\text{C}^{-1}$  and  $\alpha_1 = -5.5 \cdot 10^{-7} \text{ } ^\circ\text{C}^{-2}$  for platinum and  $\alpha = 5.2 \cdot 10^{-3} \text{ } ^\circ\text{C}^{-1}$  and  $\alpha_1 = 7.0 \cdot 10^{-7} \text{ } ^\circ\text{C}^{-2}$  for tungsten.

There are a number of cooling laws that are commonly used in thermal anemometry. The most common of these cooling laws is King's [1914]:

$$\frac{I^2 R_s}{T_s - T_f} = A_o + B_o \sqrt{\text{Re}} \quad (42.8)$$

Siddall and Davies [1972] expressed King's law in a modified form given by

$$E_b^2 = A + BU^{0.5} + CU \quad (42.9)$$

where  $A = 1.273$  ,  $B = 0.860$  , and  $C = -0.017$  . Collis and Williams [1959], on the other hand, derived the following cooling law:

$$\text{Nu} \left( \frac{T_m}{T_f} \right)^{-0.17} = A_1 + B_1 \text{Re}^{n_1} \quad (42.10)$$

where  $A_1$  ,  $B_1$  , and  $n_1$  take the values 0.24, 0.56, and 0.45, respectively, when the Reynolds number is between 0.02 and 44, while their values become 0, 0.48, and 0.51, respectively, for a Reynolds number larger than 44 and smaller than 140. The quantity  $T_m$  is the arithmetic average of  $\bar{T}_s$  and  $T_f$ .

Kramer's cooling law [Hinze, 1959] can be expressed by

$$\text{Nu} = 0.42\text{Pr}^{0.20} + 0.57\text{Pr}^{0.33} \text{Re}^{0.50} \quad (42.11)$$

This equation is valid over the Reynolds number range of  $0.1 < \text{Re} < 10\,000$  . Other cooling laws include Van der Hegge Zijnen's [1956], who derived the following equation:

$$\text{Nu} = 0.38\text{Pr}^{0.2} + (0.56\text{Re}^{0.5} + 0.001\text{Re})\text{Pr}^{0.333} \quad (42.12)$$

The effect of ambient temperature changes may be expressed in terms of the velocity and temperature sensitivities as follows:

$$e_s = S_{\text{vel}}U + S_{\text{temp}}t_f \quad (42.13)$$

$$S_{\text{vel}} = \left\{ \frac{n_1 \pi l k_o R_s}{2^{1.97} \bar{E}_s \bar{U} T_o^{0.80}} \left[ \frac{\rho_o \bar{U} d}{\mu_o} (2T_o)^{1.76} \right]^{n_1} B_1 \right\} \cdot \frac{(T_s + \bar{T}_f)^{0.97-1.76n_1} (T_s - \bar{T}_f)}{\bar{T}_f^{0.17}} \quad (42.14)$$

$$\begin{aligned}
S_{\text{temp}} = & -\frac{1}{2} \left\{ \frac{1.76n_1 \pi l k_o R_s}{2^{0.97} \bar{E}_s T_o^{0.80}} \left[ \frac{\rho_o \bar{U} d}{\mu_o} (2T_o)^{1.76} \right]^{n_1} \right. \\
& \cdot B_1 \frac{T_s - \bar{T}_f}{\bar{T}_f^{0.17} (T_s + \bar{T}_f)^{0.03+1.76n_1}} \\
& \cdot \bar{E}_s \left( \frac{0.17}{\bar{T}_f} + \frac{1}{T_s - \bar{T}_f} - \frac{0.97}{T_s + \bar{T}_f} \right) \left. \right\} \quad (42.15)
\end{aligned}$$

The corrected bridge voltage is:

$$E_{bc} \simeq E_b \left[ 1 - \frac{T_{01} - T_{02}}{2(T_s - T_{01})} \right] \quad (42.16)$$

The frequency response of a constant-current anemometer may be expressed by the following differential equation:

$$\begin{aligned}
\frac{dr_s}{dt} + \frac{\alpha R_o}{\rho c A l} \left\{ \frac{\pi l k_f}{\alpha R_o} \left[ 0.42 \text{Pr}^{0.20} + 0.57 \text{Pr}^{0.33} \left( \frac{\bar{U} d}{\nu} \right)^{0.5} \right] - \bar{I}^2 \right\} r_s \\
= \left( \frac{\pi k_f}{2 \rho c A \bar{U}} \right) 0.57 \text{Pr}^{0.33} \left( \frac{\bar{U} d}{\nu} \right)^{0.5} (\bar{R}_s - R_f) u \quad (42.17)
\end{aligned}$$

whose time constant is

$$\tau_{\text{cca}} = \frac{\rho C A l (\bar{R}_s - R_f)}{\bar{I}^2 \alpha R_f R_o} \quad (42.18)$$

while the frequency response of a constant-temperature anemometer may be expressed by the following equation:

$$\begin{aligned}
\frac{di_s}{dt} + \frac{\alpha R_o \bar{I}_s^2 [2g \bar{R}_s (\bar{R}_s - R_f) + R_f^2]}{\rho c A l (\bar{R}_s - R_f)} i_s \\
= 0.57 \text{Pr}^{0.33} \sqrt{\frac{\bar{U} d}{\nu}} \frac{\pi l k_f (\bar{R}_s - R_f)}{2 \alpha R_o \bar{U}} u \quad (42.19)
\end{aligned}$$

whose time constant is



$$\tau_{\text{cta}} = \frac{\rho c A l (\overline{R}_s - R_f)}{\alpha R_o \overline{I}_s^2 [2g R_s (\overline{R}_s - R_f) + R_f^2]} \quad (42.20)$$

The mean velocity vector may be computed using the following set of equations:

$$U_{\text{eff}_x}^2 = U^2 (\cos^2 \alpha + k^2 \sin^2 \alpha) \quad (42.21)$$

$$U_{\text{eff}_y}^2 = U^2 (\cos^2 \beta + k^2 \sin^2 \beta) \quad (42.22)$$

$$U_{\text{eff}_z}^2 = U^2 (\cos^2 \gamma + k^2 \sin^2 \gamma) \quad (42.23)$$

$$\cos \alpha = \frac{1}{U} \sqrt{\frac{U_{\text{eff}_x}^2 - k^2 U^2}{1 - k^2}} \quad (42.24)$$

$$\cos \beta = \frac{1}{U} \sqrt{\frac{U_{\text{eff}_y}^2 - k^2 U^2}{1 - k^2}} \quad (42.25)$$

$$\cos \gamma = \frac{1}{U} \sqrt{\frac{U_{\text{eff}_z}^2 - k^2 U^2}{1 - k^2}} \quad (42.26)$$

$$\cos^2 \alpha + \cos^2 \beta + \cos^2 \gamma = 1 \quad (42.27)$$

$$U = \sqrt{\frac{U_{\text{eff}_x}^2 + U_{\text{eff}_y}^2 + U_{\text{eff}_z}^2}{1 + 2k^2}} \quad (42.28)$$

## Laser Doppler Anemometers

The Doppler frequency may be expressed in terms of half the angle between two identical laser beams with frequency  $f$  (wavelength  $\lambda$ ) and the flow velocity component  $V_\theta$  perpendicular to the two-beam bisector, as follows:

$$f_D = \frac{2 \sin \theta}{\lambda} V_\theta \quad (42.29)$$

A major advantage of this equation is the fact that it is linear and does not contain any undetermined constants, thus eliminating the need for calibration. The width of the measuring

volume may be expressed in terms of the transmitting length focal distance,  $f_T$ , and the diameter of the beam waist before the transmitting lens,  $D$ :

$$d = \frac{4f_T \lambda}{\pi D} \quad (42.30)$$

The length and height, on the other hand, may be expressed by

$$l = \frac{d}{\sin \theta} \quad \text{and} \quad h = \frac{d}{\cos \theta} \quad (42.31)$$

The spacing of the nearly parallel fringes (produced by the interference of the two light beams in the measuring volume) as well as the number of fringes can be expressed by

$$\delta = \frac{\lambda}{2 \sin \theta} \quad \text{and} \quad n = \frac{4\Delta}{\pi D} \quad (42.32)$$

The Doppler frequency can be determined using frequency counters that time a fixed number,  $N$ , of zero crossings. This allows computing of the particle velocity by simply using the following relationship:

$$V_\theta = \frac{N\delta}{\Delta t} \quad (42.33)$$

Frequency counters are extremely accurate and have a wide dynamic range; however, their output is provided at irregular intervals. This necessitates using special statistical procedures to perform proper counting. Furthermore, the performance of frequency counters may be compromised at higher particle concentrations, when multiple particles are likely to coexist in the measuring volume.

## Volume Flow Measurements

The equations describing the ideal performance characteristics of differential pressure meters for incompressible fluids can be expressed by

$$V_2 = \left[ \frac{2g(p_1 - p_2)}{\xi(1 - \beta^4)} \right]^{1/2} \quad \text{and} \quad \dot{M}_I = A_2 \left[ \frac{2\rho(p_1 - p_2)}{(1 - \beta^4)} \right]^{1/2} \quad (42.34)$$

The corresponding equations for compressible fluids are:

$$V_2 = \left[ \frac{2\gamma p_1 (1 - r^{1-1/\gamma})}{(\gamma - 1)\rho_1 (1 - \beta^4 r^{2/\gamma})} \right]^{1/2} \quad \text{and} \quad (42.35)$$

$$\dot{M}_I = \frac{A_2 p_1}{T_1^{1/2}} \left[ \frac{r^{2/\gamma} (r^{2/\gamma} - r^{1+1/\gamma})}{1 - \beta^4 r^{2/\gamma}} \right]^{1/2} \left( \frac{g}{R} \frac{2\gamma}{\gamma - 1} \right)^{1/2}$$

The equations describing the performance of a real compressible orifice flow, on the other hand, are:

$$\dot{Q}_h = C' \sqrt{h_w p_f} \quad \text{and} \quad C' = F_b F_r Y F_{pg} F_{tb} F_{tf} F_g F_{pv} F_m F_a F_\ell \quad (42.36)$$

The differential pressure across a paddle-type orifice plate is related to the fluid flow rate by

$$\dot{Q} = A_2 C_D \left[ \frac{2\Delta p}{\rho(1 - \beta^4)} \right]^{1/4} \quad (42.37)$$

## Defining Terms

**Beam translator:** Used in LDV systems to adjust the intersection angle by reducing the standard beam distance.

**Beam waist adjuster:** Used in LDV systems with long focal lengths to optimize the fringe pattern quality in the measuring volume.

**Bragg cell:** A module in LDV systems capable of providing a positive or negative optical frequency shift of the laser light.

**Frequency tracker:** A device capable of measuring the instantaneous frequency of the LDV signal. There are two types of trackers used in LDV: the phase-locked loop (PLL) and the frequency-locked loop (FLL).

**Frequency counter:** A device used in LDV systems capable of measuring the frequency of a signal by accurately timing the duration of an integral number of cycles of the signal.

**Non-Newtonian fluids:** Fluids in which the coefficient of viscosity is not independent of the velocity gradient.

**Photomultiplier tube (PMT):** A device in LDV systems capable of using the photoelectric effect, wherein photons striking a coating of photoemissive material on the photocathode cause electrons to be emitted from the material.

**Pinhole translator:** Used in LDV backscatter measurements. The device can image the measuring volume on a pinhole, thereby constituting an efficient additional spatial filter, thus eliminating undesirable reflections from window surfaces and walls in the vicinity of the measuring volume.

**Signal processor:** A device in LDV systems designed to measure the Doppler frequency in addition to any other relevant data coming from the photomultiplier tube (PMT) signal.

# FLUID MEASUREMENTS

## List of Symbols

$\alpha$	overheat ratio	$\alpha$	temperature coefficient of resistivity, thermal diffusivity, angle of inclination of the velocity vector
$A$	area	$\beta$	volume coefficient of expansion, angle of inclination of the velocity vector, ratio of orifice hole to pipe diameter
$b$	yaw parameter	$\gamma$	angle of inclination of the velocity vector, specific heat ratio
$c$	specific heat	$\Delta$	parallel beam separation
$C'$	orifice flow constant	$\theta$	yaw angle
$C_D, C_d$	dimensionless discharge coefficients	$\mu$	absolute viscosity
$c_p$	specific heat at constant pressure	$\nu$	kinematic viscosity
$c_v$	specific heat at constant volume	$\rho$	density
$d$	diameter	$\rho_r$	resistivity
$D$	inside pipe diameter	$\tau$	time constant
$e$	fluctuating component of voltage	$\phi$	phase angle, pitch angle
$E$	voltage	$\psi$	roll angle
$f$	frequency	$\xi$	specific weight
$F$	meter factor		
$F_b$	basic orifice factor	Subscripts	
$F_r$	Reynolds-number factor	1, 2	positions along conduit, specific meters in same pipe
$F_{pb}$	pressure-base factor	$a$	air, orifice thermal expansion
$F_{tb}$	temperature-base factor	$b$	bridge, basic
$F_{tf}$	flowing-temperature factor	$c$	cable, corrected, convection, collected
$F_g$	specific-gravity factor	cca	constant-current anemometer
$F_{pv}$	supercompressibility factor	cta	constant-temperature anemometer
$F_m$	mercury-manometer factor	eff	effective cooling
$F_a$	orifice thermal-expansion factor	$f$	fluid
$F_\ell$	gauge-location factor	$F$	facility in which meter is tested
$g$	acceleration of gravity	$g$	gas, specific gravity
$h$	coefficient of convective heat transfer, height, differential pressure	$I$	ideal
$i$	fluctuating component of current	$i$	sample number
$I$	electrical current	$\ell$	gage location
$k$	thermal conductivity, yaw factor	$m$	mean, measured, mixture, manometer
$K$	flow coefficient, flowmeter constant	$M$	specific flowmeter
$l$	length	$n$	arbitrary orthogonal coordinate
$\dot{m}, \dot{M}$	mass rate of flow	$o$	reference or stagnation conditions
$n$	exponent used in King's law	$p$	probe, constant pressure
Nu	Nusselt number	$s$	systematic, static, sensor, streamwise
$p$	steady or time-averaged pressure	$t$	arbitrary orthogonal coordinate, total
Pr	Prandtl number	$T$	total, temperature
$q$	heat transfer rate	temp	temperature
$\dot{Q}$	volume flow rate	$v$	constant volume
$r$	fluctuating component of electrical resistance, radius, recovery factor, correlation coefficient, pressure ratio	$w$	wall
$R$	electrical resistance, gas constant	$\infty$	free stream conditions
Re	Reynolds number		
$S$	sensitivity		
$t$	fluctuating component of temperature, time		
$T$	temperature, absolute temperature		
$u$	fluctuating component of velocity		
$U$	$x$ -component of velocity, fluid velocity		
$V$	average fluid velocity in conduit, point fluid velocity		
$\dot{W}$	weight flow rate		
$x$	characteristic length, horizontal distance		
$y$	vertical distance		
$Y$	expansion factor		

**Similarity analysis:** One of the most powerful tools in fluid mechanics, which permits a wide generalization of experimental results.

**Thermal anemometer:** A device that measures fluid velocity by sensing the changes in heat transfer from a small electrically heated sensor exposed to the fluid flow.

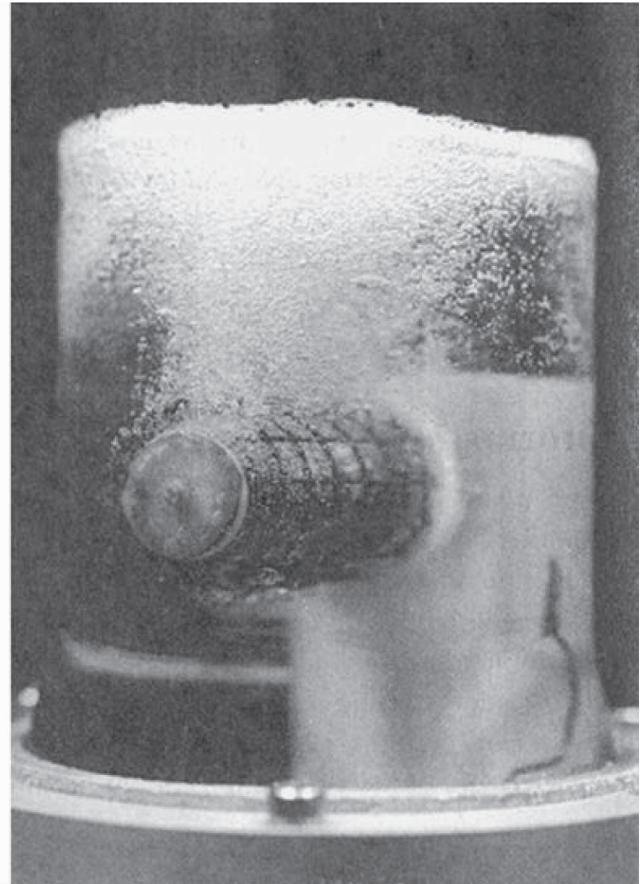
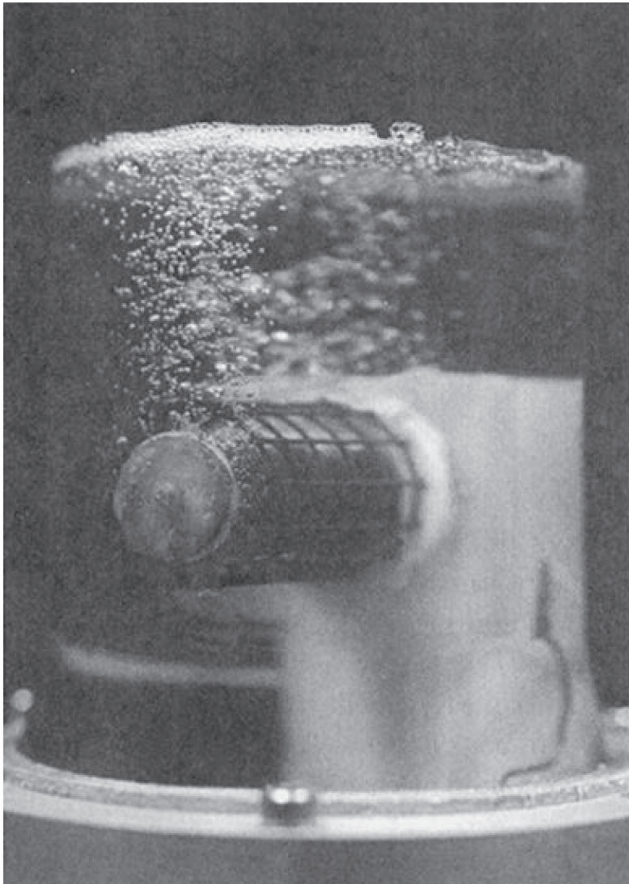
## References

- Adrian, R. J. 1983. Laser velocimetry. In *Fluid Mechanics Measurements*, ed. R. J. Goldstein, pp. 155–244. Hemisphere, Washington, DC.
- Blake, W. K. 1983. Differential pressure measurement. In *Fluid Mechanics Measurements*, ed. R. J. Goldstein, pp. 61–97. Hemisphere, Washington, DC.
- Collis, D. C. and Williams, M. J. 1959. Two-dimensional convection from heated wires at low Reynolds numbers. *J. Fluid Mech.* 6:357–384.
- Hinze, J. O. 1959. *Turbulence*. McGraw-Hill, New York.
- King, L. V. 1914. On the convection of heat from small cylinders in a stream of fluid: Determination of the convection constants of small platinum wires with applications to hot-wire anemometry. *Phil. Trans. Royal Soc. (London) A.* 214:373–432.
- Lomas, C. G. 1986. *Fundamentals of Hot Wire Anemometry*. Cambridge University Press, Cambridge, United Kingdom.
- Mattingly, G. E. 1983. Volume flow measurements. In *Fluid Mechanics Measurements*, ed. R. J. Goldstein, pp. 245–306. Hemisphere, Washington, DC.
- Pien, P. C. 1958. *Five-Hole Spherical Pitot Tube*. DTMB Report 1229.
- Siddall, R. G. and Davies, T. W. 1972. An improved response equation for hot-wire anemometry. *Int. J. Heat Mass Transfer.* 15:367–368.
- Van der Hegge Zijnen, B. G. 1959. Modified correlation formulae for the heat transfer by natural and by forced convection from horizontal cylinders. *Appl. Sci. Res. A.* 6:129–140.

## Further Information

A good discussion of the principles of hot-wire/film anemometry may be found in *Hot-Wire Anemometry* by A. E. Perry. The author covers all aspects of this measuring technique. A good introduction to the principles of laser Doppler velocimetry can be found in *Laser Doppler Measurements* by B. M. Watrasiewicz and M. J. Rudd. Volume flow measurements include several diverse topics, but the book by H. S. Bean on *Fluid Meters—Their Theory and Applications* may provide a good starting point for the interested reader. Other measurement techniques may be found in *Fluid Mechanics Measurements* by R. J. Goldstein. This book is an excellent reference for a broad spectrum of measurement techniques commonly used in fluid mechanics.

Kreith, F. "Thermodynamics and Heat Transfer"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



Heat exchangers are essential components of almost all engineering systems. More efficient heat exchangers can lead to substantial energy savings in a broad range of industrial applications. High performance heat exchangers can deliver a specified heat transfer duty with substantial savings in the materials involved and the space required to accommodate the heat exchanger. A major step in the development of high-performance heat exchangers was commercialization and wide use of *enhanced heat transfer surfaces* beginning in the mid-1980s. Such surfaces utilize various fins and macro/microscale alteration on the heat transfer surface that leads to heat transfer coefficients as much as ten times higher than their traditional smooth (or plain) versions.

More recently, researchers in Japan, the U.S., and the U.K. have successfully demonstrated a new technique that promises to revolutionize the way heat exchangers are designed, built, and integrated into systems. The technique, called electrohydrodynamics (or EHD), applies an electrostatic field to the working fluid(s) in the heat exchanger. Coupling of the electric and flow fields takes place with simple wire electrodes that run parallel to the heat transfer surface. When EHD is applied over the best-performing enhanced surfaces, an additional increase of two to ten times the heat transfer coefficients is realized. Moreover, because in an EHD-enhanced surface the magnitude of the heat transfer coefficient is directly proportional to the applied voltage, the EHD technique provides an instantaneous variable-capacity feature for the heat exchanger.

Photographs above show an evaporator tube in refrigerant 123 (an environmentally friendly substitute for refrigerant 11). The left photo shows the tube with no EHD. The refrigerant is boiling at a certain heat flux. The right photo shows the same tube under the same conditions but with the EHD field applied. The boiling is now much more vigorous, with heat transfer coefficients nearly seven to ten times higher than those for the case depicted on the left. (Courtesy of the Heat Transfer Enhancement Laboratory, University of Maryland, College Park.)

# VII

## Thermodynamics and Heat Transfer

---

**Frank Kreith**

*University of Colorado*

- 43    [The First Law of Thermodynamics](#) *R. E. Sonntag*  
System Analysis • Control Volume Analysis
- 44    [Second Law of Thermodynamics and Entropy](#) *N. Lior*  
Reversibility • Entropy • The Second Law for Bulk Flow • Applications
- 45    [The Thermodynamics of Solutions](#) *S. I. Sandler and H. Orbey*  
Fundamentals • Applications
- 46    [Thermodynamics of Surfaces](#) *W. B. Krantz*  
Basic Concepts • First Law of Thermodynamics • Effects of Curved Interfaces • Adsorption at Interfaces • Wettability and Adhesion
- 47    [Phase Equilibrium](#) *B. G. Kyle*  
Pure-Component Phase Equilibrium • Phase Equilibrium in Mixtures • Perspective
- 48    [Thermodynamic Cycles](#) *W. J. Cook*  
Power Cycles • Refrigeration Cycles
- 49    [Heat Transfer](#) *Y. Bayazitoglu and U. B. Sathuvalli*  
Conduction • Convection • Radiation • Phase Change
- 50    [Heat Exchangers](#) *M. M. Ohadi*  
Heat Exchanger Types • Shell-and-Tube Heat Exchangers • Compact Heat Exchangers • Design of Heat Exchangers
- 51    [Combustion](#) *R. S. Hecklinger*  
Fundamentals of Combustion • Combustion Calculations
- 52    [Air Conditioning](#) *V. W. Goldschmidt and C. J. Wahlberg*  
Historical Sketch • Comfort • Air Conditioning Process • Representative Cycles
- 53    [Refrigeration and Cryogenics](#) *R. F. Barron*  
Desiccant Cooling • Heat Pumps • Cryogenics
- 54    [Heat Transfer to Non-Newtonian Fluids](#) *E. F. Matthys*  
The Fluids • Friction • Heat Transfer • Instrumentation and Equipment
- 55    [Heat Pipes](#) *L. W. Swanson*  
Heat Pipe Container, Working Fluid, and Wick Structures • Heat Transfer Limitations • Effective Thermal Conductivity and Heat Pipe Temperature Difference • Application of Heat Pipes

THE BRANCH OF SCIENCE THAT DEALS WITH the relation between heat and other forms of energy is called thermodynamics. Processes by which energy transport takes place as a result of a temperature gradient are known as heat transfer. All heat transfer processes involve the transfer and conversion of energy. They must, therefore, obey the first as well as the second law of thermodynamics. However, the principles of heat transfer cannot be derived from the basic laws of



thermodynamics alone, because classical thermodynamics is restricted to a study of equilibrium states. Since heat flow is the result of temperature nonequilibrium, its quantitative treatment must draw on a variety of disciplines in engineering science in addition to thermodynamics.

This section combines thermodynamics and heat transfer to help the reader understand the interrelation between these two topics. But, from a thermodynamic viewpoint, the amount of heat transferred during a process simply equals the difference between the energy change of the system and the work done; it does not indicate the time required to carry out the process. The key problem in engineering analysis is to determine the *rate of heat transfer* under given thermal boundary conditions. The dimensions of boilers, heaters, refrigerators, and heat exchangers depends not only on the amount of heat to be transmitted, but primarily on the rate at which the heat is to be transferred under given conditions. In this section, after presentation of the basic laws of thermodynamics, the heat transfer mechanisms by the various modes of transport are treated. These modes are conduction, convection, and radiation. Convection depends not only on the temperature conditions, but also on the fluid mechanical mass transport involved. Hence, an understanding of fluid mechanics is integral to the study of the convection process, and the reader is referred to Section VI for a treatment of the fluid mechanical underpinnings.

Heat transfer and thermodynamics are key elements in the analysis of energy conversion and transport. The authors in this section have drawn on their vast experience to present the material in a way that makes it possible for the reader to apply the basic information in this section to practical applications in thermal engineering, including combustion, energy generation, refrigeration, and similar important applications.

Sonntag, R. E. "The First Law of Thermodynamics"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# The First Law of Thermodynamics

---

## 43.1 System Analysis

## 43.2 Control Volume Analysis

Steady State, Steady Flow Model • Uniform State, Uniform Flow Model

**Richard E. Sonntag**

*University of Michigan*

## 43.1 System Analysis

---

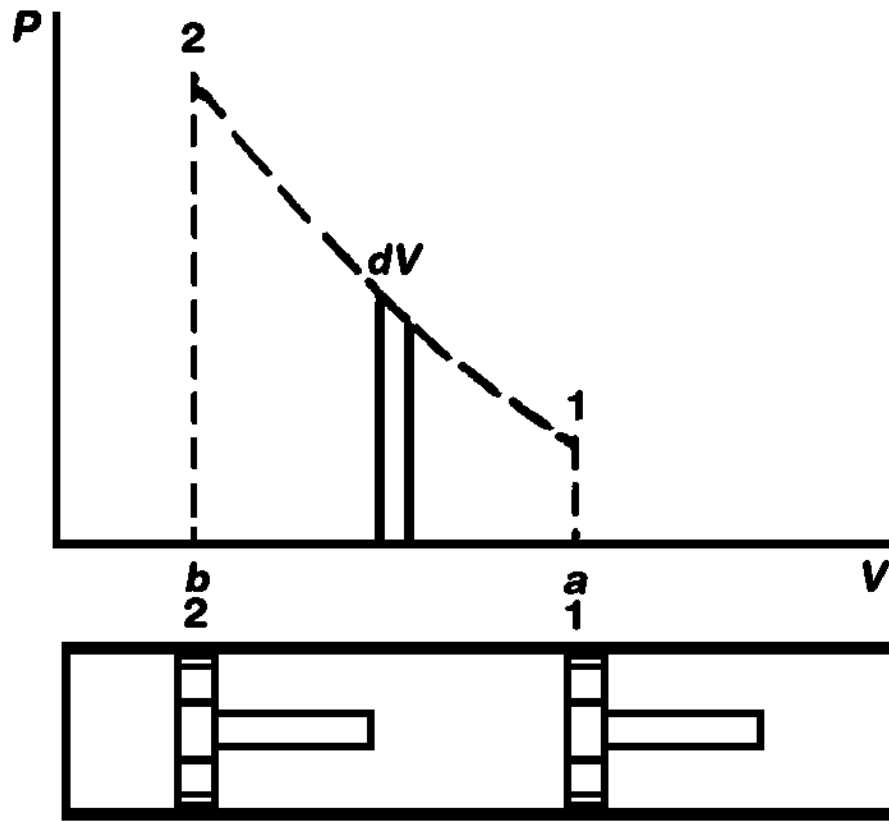
The first law of thermodynamics is frequently called the *conservation of energy* (actually, mass-energy) because it represents a compilation of all the energy transfers across the boundary of a thermodynamic system as the system undergoes a process. Energy transfers are of two forms—work and heat. Both work and heat are transient quantities—not possessed by a system—and both are boundary phenomena, that is, observed only in crossing a system boundary. In addition, both are path functions; that is, they are dependent on the path of the process followed by the change in state of the system. Work and heat differ in that work can always be identified as the equivalent of a force acting through a related displacement, whereas heat is energy transfer due to a temperature gradient or difference, from higher to lower temperature.

One common type of work is that done at a movable boundary, such as shown in [Fig. 43.1](#). The boundary movement work equals the external force times the distance through which it acts, so that for a quasi-equilibrium process in which the forces are always balanced, the work for a process between states 1 and 2 is given by the expression

$$W_{12} = \int_1^2 \delta W = \int_1^2 P \, dV \quad (43.1)$$

such that the work is equal to the area under the curve on the  $P$ - $V$  diagram shown in [Fig. 43.1](#).

**Figure 43.1** Boundary movement work. (Source: Van Wylen, G. J., Sonntag, R. E., and Borgnakke, C. 1994. *Fundamentals of Classical Thermodynamics*, 4th ed. John Wiley & Sons, New York. Used by permission.)



Systems in which other driving forces and their associated displacements are present result in corresponding quasi-equilibrium work transfers that are the product of the force and displacement, in a manner equivalent to the compressible-substance boundary movement work of Eq. (43.1). For any of these work modes the quasi-equilibrium process is an idealized model of a real process (which occurs at a finite rate because of a finite gradient in the driving force). It is nevertheless a useful model against which to compare the real process.

By convention, heat transfer to the system from its surroundings is taken as positive (heat transfer from the system is therefore negative), and work done by the system on its surroundings is positive (work done on the system is negative). For a process in which the system proceeds from initial state 1 to final state 2 the change in total energy  $E$  possessed by the system is given as

$$E_2 - E_1 = Q_{12} - W_{12} \quad (43.2)$$

The total energy can be divided into the part that depends only on the thermodynamic state (the internal energy  $U$ ) and the part that depends on the system's motion (kinetic energy  $KE$ ) and position with respect to the chosen coordinate frame (potential energy  $PE$ ), such that, at any state,

$$E = U + KE + PE \quad (43.3)$$

In many applications, changes in the system kinetic energy and potential energy are negligibly small by comparison with the other energy terms in the first law.

For a constant-pressure process with boundary movement work and no changes in kinetic and potential energies, the first law can be written as

$$\begin{aligned} Q_{12} &= U_2 - U_1 + P(V_2 - V_1) = (U_2 + P_2 V_2) - (U_1 + P_1 V_1) \\ &= H_2 - H_1 \end{aligned} \quad (43.4)$$

in which  $H$  is the property termed the *enthalpy*.

Consider a process involving a single phase (either solid, liquid, or vapor) with possible boundary movement work, as given by Eq. (43.1). Any heat transferred to the system will then be associated with a temperature change. The specific heat is defined as the amount of heat transfer required to change a unit mass by a unit temperature change. If this process occurs at constant volume, there will be no work and the heat transfer equals the internal energy change. If the process occurs at constant pressure, then the heat transfer, from Eq. (43.4), equals the enthalpy change. Thus, there are two specific heats, which become

$$C_V = \frac{1}{m} \left( \frac{\partial U}{\partial T} \right)_V = \left( \frac{\partial u}{\partial T} \right)_V ; \quad C_P = \frac{1}{m} \left( \frac{\partial H}{\partial T} \right)_P = \left( \frac{\partial h}{\partial T} \right)_P \quad (43.5)$$

The specific heat is a property that can be measured or, more precisely, can be calculated from quantities that can be measured. The specific heat can then be used to calculate internal energy or enthalpy changes. In the case of solids or liquids, the energy and enthalpy depend primarily on temperature and not very much on pressure or specific volume. For an ideal gas, a very low-density gas following the equation of state,

$$Pv = RT \quad (43.6)$$

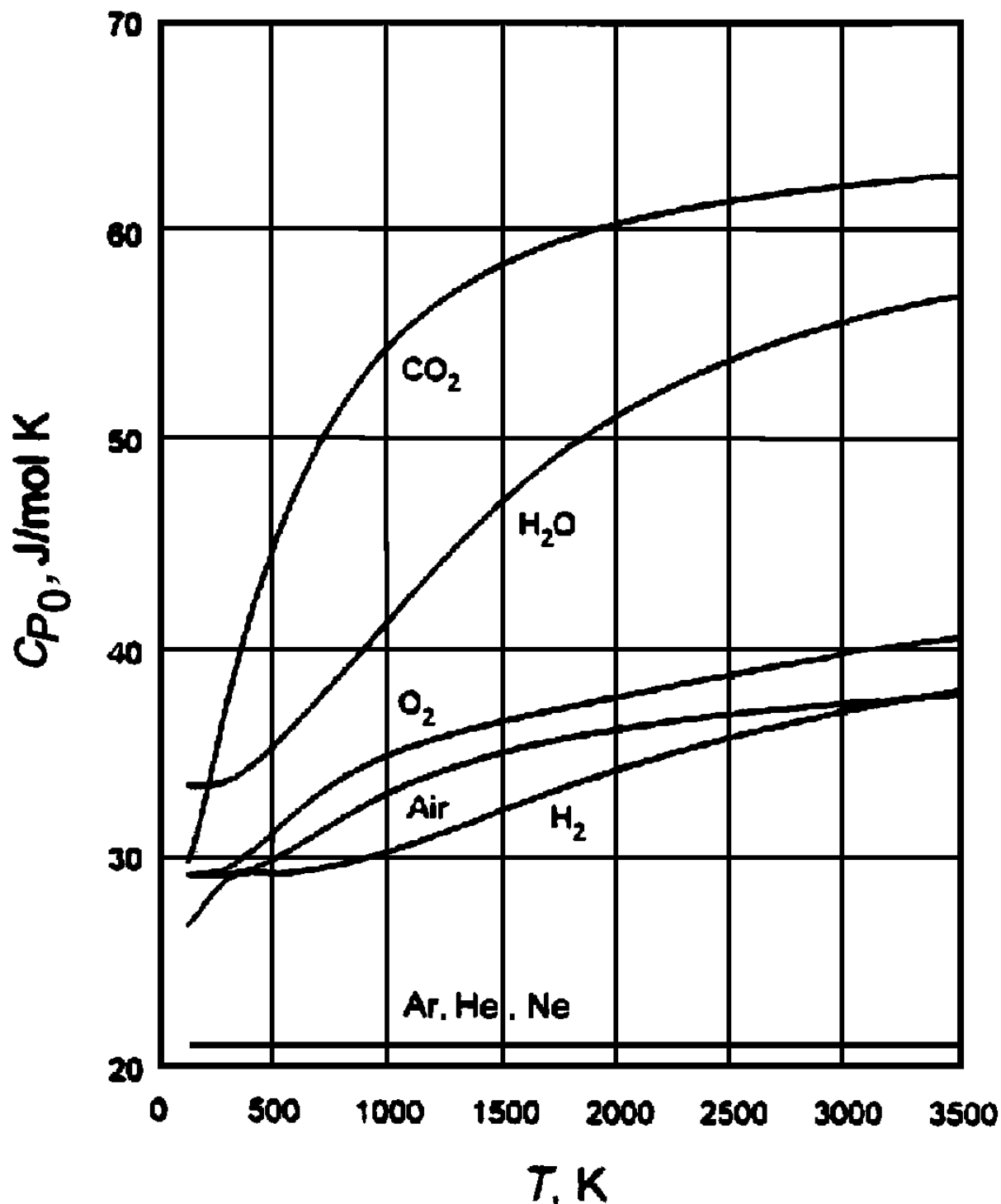
these properties depend only on temperature and not at all on  $P$  or  $v$ . Therefore, the expressions in Eq. (43.5) can be used to calculate changes in  $u$  or  $h$  as

$$u_2 - u_1 = \int_1^2 C_{V0} dT; \quad h_2 - h_1 = \int_1^2 C_{P0} dT \quad (43.7)$$

The subscript 0 is included to denote that these are the specific heats for the ideal gas model. In

order to integrate these expressions, it is necessary to know the dependence of specific heat on temperature. These are commonly calculated from statistical thermodynamics and tabulated as functions of temperature. Values for several common gases are shown in Fig. 43.2.

**Figure 43.2** Ideal gas specific heats. (Source: Van Wylen, G. J., Sonntag, R. E., and Borgnakke, C. 1994. *Fundamentals of Classical Thermodynamics*, 4th ed. John Wiley & Sons, New York. Used by permission.)



For real gases or liquids, internal energy and enthalpy dependency on pressure can be calculated using an equation of state. The changes between phases—for example, between liquid and vapor at the same temperature—can be determined using thermodynamic relations. These real properties can then all be tabulated in tables of thermodynamic properties, many of which exist in the literature.

## 43.2 Control Volume Analysis

---

In many thermodynamic applications, it is often convenient to adopt a different perspective concerning the first-law analysis. Such cases involve the analysis of a device or machine through which mass is flowing. It is then appropriate to consider a certain region in space, a control volume, and to analyze the energy being transported across its surfaces by virtue of the mass flows, in addition to the heat and work transfers. Whenever a mass  $dm_i$  flows into the control volume, it is necessarily pushed into it by the mass behind it. Similarly, a mass  $dm_e$  flowing out of the control volume has to push other mass out of the way. Both cases involve a local boundary movement work  $Pv\,dm$ , which must be included along with the other energy terms in the first law. For convenience, the  $Pv$  term is added to the  $u$  term for the mass flow terms, with their sum being the enthalpy. The complete first law for a control volume analysis, represented on a rate basis, is then

$$\dot{Q}_{cv} + \sum \dot{m}_i (h_i + KE_i + PE_i) = \frac{dE_{cv}}{dt} + \sum \dot{m}_e (h_e + KE_e + PE_e) + \dot{W}_{cv} \quad (43.8)$$

The summation signs on the flow terms entering and exiting the control volume are included to allow for the possibility of more than one flow stream. Note that the total energy contained inside the control volume at any instant of time,  $E_{cv}$ , can be expressed in terms of the internal energy as in Eq. (43.3).

The general expression of the first law for a control volume should be accompanied by the corresponding conservation of mass, which is

$$\frac{dm_{cv}}{dt} + \sum \dot{m}_e - \sum \dot{m}_i = 0 \quad (43.9)$$

### Steady State, Steady Flow Model

Two model processes are commonly utilized in control volume analysis in thermodynamics. The first is the *steady state, steady flow model*, commonly identified as the *SSSF model*. In this case, all states, flow rates, and energy transfers are steady with time. While the state inside the control volume is nonuniform, varying from place to place, it is everywhere steady with time. Therefore, for the SSSF model,

$$\frac{dm_{cv}}{dt} = 0; \quad \frac{dE_{cv}}{dt} = 0 \quad (43.10)$$

These terms drop from Eqs. (43.8) and (43.9), and everything else is steady, or independent of time. The resulting expressions are very useful in describing the steady long-term operation of a machine or other flow device, but of course would not describe the transient start-up or shutdown of such a device. The above expressions describing the SSSF model imply that the control volume remains rigid, such that the work rate (or power) term in the first law may include shaft work or electrical work but not boundary movement work.

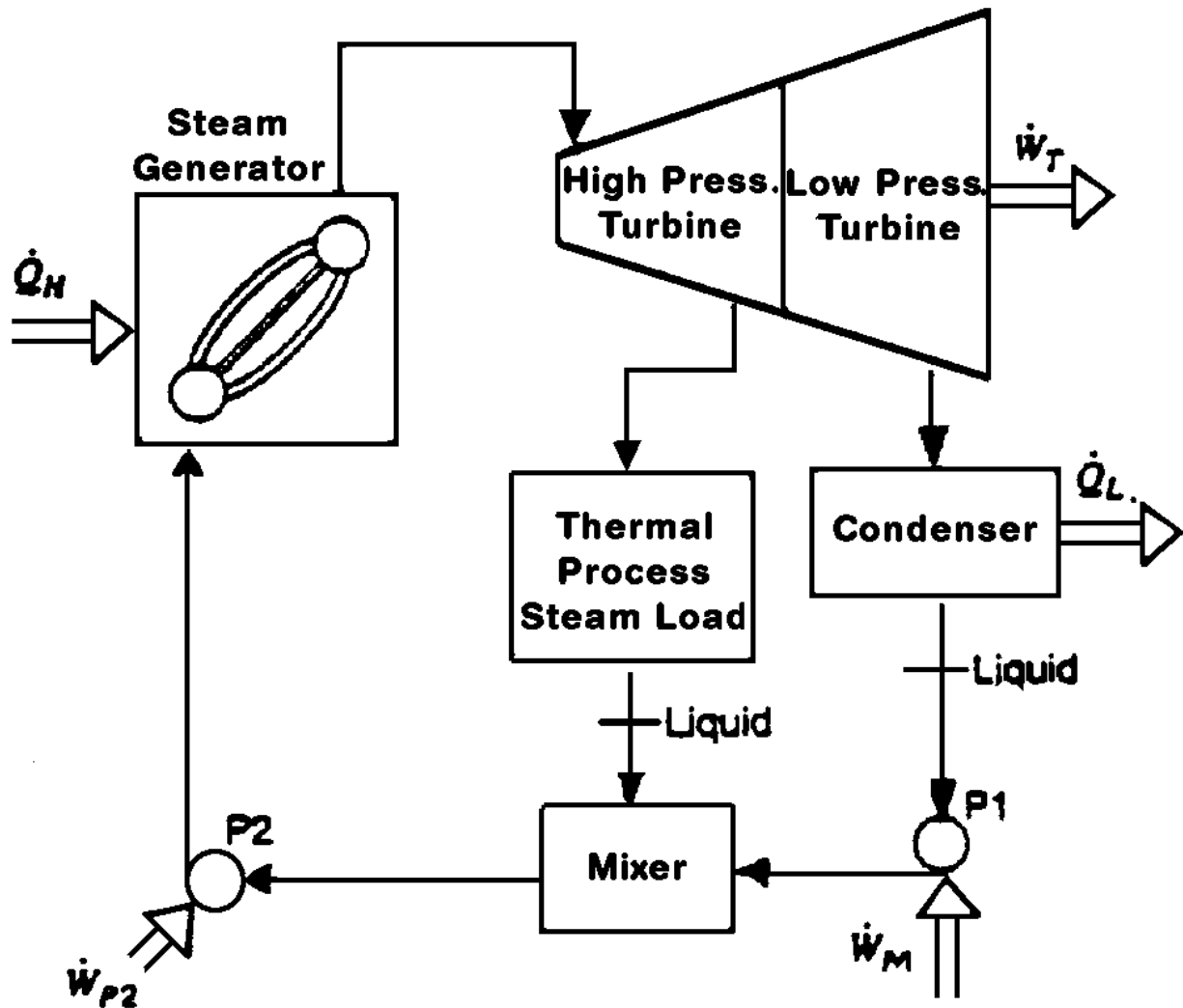
There are many common examples of SSSF model applications. One is a heat exchanger, in which a flowing fluid is heated or cooled in a process that is considered to be constant pressure (in the actual case, there will be a small pressure drop due to friction of the flowing fluid at the walls of the pipe). This process may involve a single phase—gas or liquid—or it may involve a change of phase—liquid to vapor in a boiler or vapor to liquid in a condenser. Another example of an SSSF process is a nozzle, in which a fluid is expanded in a device that is contoured such that the velocity increases as the pressure drops. The opposite flow process is a diffuser, in which the device is contoured such that the pressure increases as the velocity decreases along the flow path. Still another example is a throttle, in which the fluid flows through a restriction such that the enthalpy remains essentially constant, a conclusion reached because all the other terms in the first law are negligibly small or zero. Note that in all four of these examples of SSSF processes, the flow device includes no moving parts and there is no work associated with the process. A turbine, or other flow-expansion machine, is a device built for the purpose of producing a shaft output power—this at the expense of the pressure of the fluid. The opposite device is a compressor (gas) or pump (liquid), the purpose of which is to increase the pressure of a fluid through the input of shaft work.

Several flow devices may be coupled for a special purpose, such as the heat engine or power plant. A particular example involving **cogeneration** is shown in Fig. 43.3. High-pressure liquid water enters the steam generator, in which the water is boiled and also superheated. The vapor enters the high-pressure turbine, where it is expanded to an intermediate pressure and temperature. At this state, part of the steam is extracted and used for a specific thermal process, such as space heating. The remainder of the steam is expanded in the low-pressure turbine, producing more shaft power output. Steam exiting the turbine is condensed to liquid, pumped to the intermediate pressure, and mixed with the condensate from the thermal process steam. All the liquid is then pumped back to the high pressure, completing the cycle and returning to the steam generator. Only a small amount of shaft power is required to pump the liquid in the two pumps in comparison to that produced in the turbine. The net difference represents a large useful power output that may be used to drive other devices, such as a generator to produce electrical power.

**Figure 43.3** Cogeneration of steam and power. (Source: Van Wylen, G. J., Sonntag, R. E., and Borgnakke, C. 1994. *Fundamentals of Classical Thermodynamics*, 4th ed. John Wiley & Sons, New York. Used by permission.)

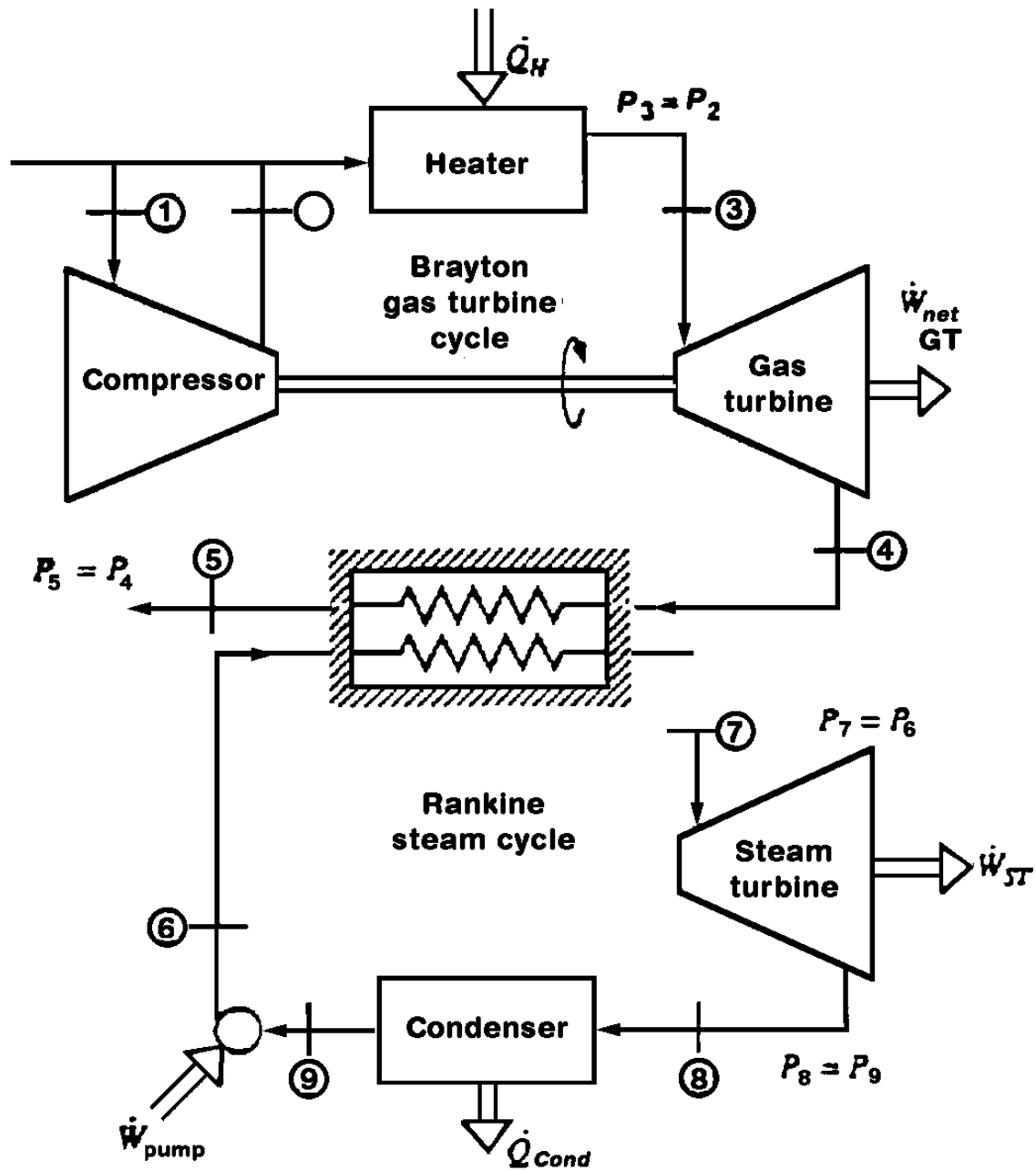


Figure 43.3



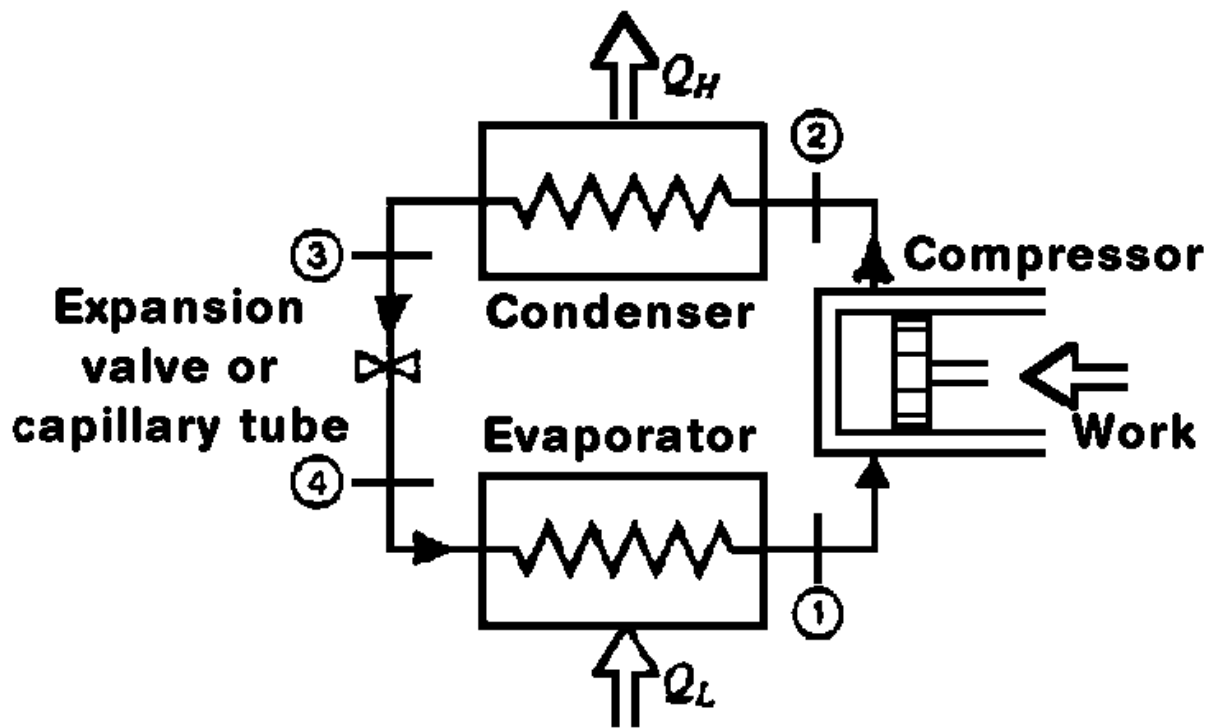
To improve overall thermal efficiency of a power plant and thereby conserve energy resources, two separate heat engines may be compounded, such as in the **combined cycle** shown in Fig. 43.4. In this application the source of heat supply to the steam turbine power plant cycle is the waste heat from a higher-operating-temperature gas turbine engine. In the latter, ambient air is compressed to a high pressure in the compressor, after which it enters a burner (represented in the figure as a heater) along with fuel. The combustion products enter the turbine at high temperature and are expanded to ambient pressure, producing a large power output, a portion of which is used to drive the compressor. The products exiting the gas turbine are still at a high enough temperature to serve as the heat source for the steam turbine cycle. The gas turbine cycle may be referred to as a *topping cycle* for the steam power plant.

**Figure 43.4** Combined cycle power system. (Source: Van Wylen, G. J., Sonntag, R. E., and Borgnakke, C. 1994. *Fundamentals of Classical Thermodynamics*, 4th ed. John Wiley & Sons, New York. Used by permission.)



Another common example of coupling several flow devices is the heat pump, or refrigerator, shown in Fig. 43.5. Low-temperature vapor enters the compressor at 1 and is compressed to a high pressure and temperature, exiting at 2. This vapor is then condensed to liquid at 3, after which the liquid is throttled to low pressure and temperature. The working fluid (part liquid and part vapor at 4) now enters the evaporator, in which the remaining liquid is boiled, completing the cycle to state 1. When the reason for building this unit is to keep the cold space at a temperature below the ambient, the quantity of interest is  $\dot{Q}_L$  and the machine is called a *refrigerator*. Likewise, when the reason for building the unit is to keep the warm space at a temperature above the ambient, the quantity of interest is  $\dot{Q}_H$  and the machine is called a *heat pump*.

**Figure 43.5** Refrigeration cycle. (Source: Van Wylen, G. J., Sonntag, R. E., and Borgnakke, C. 1994. *Fundamentals of Classical Thermodynamics*, 4th ed. John Wiley & Sons, New York. Used by permission.)



## Uniform State, Uniform Flow Model

The second model in common use in control volume analysis in thermodynamics concerns the analysis of transient processes. This model is termed the *uniform state, uniform flow model*, or *USUF model*. Equations (43.8) and (43.9) are integrated over the time of the process, during which the state inside the control volume changes, as do the mass flow rates and transfer quantities. It is,

however, necessary to assume that the state on each flow area is steady, in order that the flow terms can be integrated without detailed knowledge of the rates of change of state and flow. The integrated expressions for this model are as follows:

$$Q_{cv} + \sum m_i(h_i + KE_i + PE_i) = (E_2 - E_1)_{cv} + \sum m_e(h_e + KE_e + PE_e) + W_{cv} \quad (43.11)$$

$$(m_2 - m_1)_{cv} + \sum m_e - \sum m_i = 0 \quad (43.12)$$

The USUF model is useful in describing the overall changes in such processes as the filling of vessels with a fluid, gas, or liquid, or the opposite process, the discharge of a fluid from a vessel over a period of time.

## Defining Terms

**Cogeneration:** The generation of steam in a steam generator (boiler) for the dual purpose of producing shaft power output from a turbine and also for use in a specific thermal process load, such as space heating or for a particular industrial process.

**Combined cycle:** Combination of heat-engine power cycles for the purpose of utilizing the waste heat from the higher-operating-temperature cycle as the supply of energy to drive the other cycle. The result is an increase in thermal efficiency, as compared with the use of a single cycle, because of the increased power output from the same heat source input.

## Reference

Van Wylen, G. J., Sonntag, R. E., and Borgnakke, C. 1994. *Fundamentals of Classical Thermodynamics*, 4th ed. John Wiley & Sons, New York.

## Further Information

Advanced Energy Systems Division, American Society of Mechanical Engineers, 345 E. 47th Street, New York, NY 10017.

*International Journal of Energy Research*, published by John Wiley & Sons, New York.

Lior, N. "Second Law of Thermodynamics and Entropy"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Second Law of Thermodynamics and Entropy

---

### 44.1 Reversibility

### 44.2 Entropy

### 44.3 The Second Law for Bulk Flow

### 44.4 Applications

Direction and Feasibility of Processes • Process Efficiency • Exergy Analysis

### Noam Lior

*University of Pennsylvania*

Thermodynamics is founded on a number of axioms, laws that have not been proven in a general sense but seem to agree with all of the experimental observations of natural phenomena conducted so far. Most commonly, these axioms are formulated under the names of the first, second, third, and zeroth laws of thermodynamics, but other axiomatic definitions have been formulated and shown to be equally successful [see [Callen, 1985](#)].

The first law of thermodynamics is a statement of energy conservation, mathematically providing an energy-accounting equation useful in the analysis of **thermodynamic systems** and generally enlightening us about energy conservation and the axiom that, as far as energy is concerned, one cannot get something for nothing. It does not, however, provide guidance about the directions and limitations of the work, heat, and other energy conversion interactions in the process under consideration. For example, the assignment of an amount of heat (positive or negative) in a first law equation is not conditioned on whether that amount of heat is delivered from a low-temperature source to a high-temperature one or vice versa; heat, work, and other forms of energy are treated as having the same thermodynamic quality, and there is no restriction on how much of the heat input could be converted into work as long as overall energy conservation is maintained. The second law provides these types of guidance and thus more strongly distinguishes thermodynamics from its subset branches

of physics, such as mechanics.

Several competing statements and practical corollaries of the second law have been developed over the years. They are all consistent with each other. Those that have more rigor and generality are also typically somewhat more obscure to the practitioner, whereas those that are stated in simpler and more practical terms, though still correct, do not encompass all possible processes.

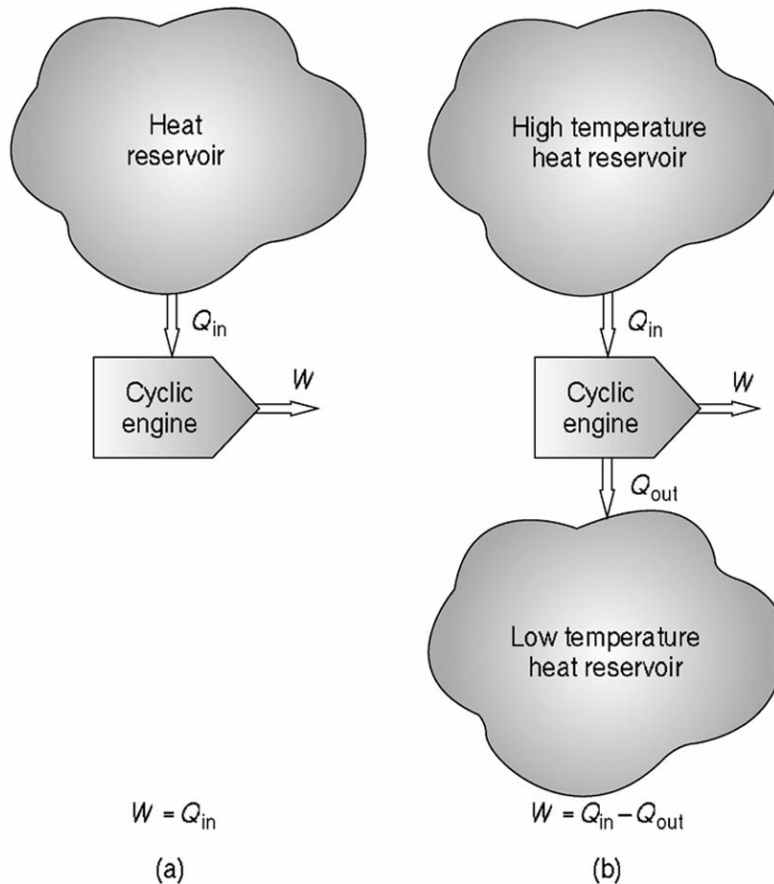
The best known statement of the second law is the Kelvin-Planck one: "It is impossible to construct an engine that, operating in a **cycle**, will produce no effect other than the extraction of heat from a single **reservoir** and the performance of an equivalent amount of work." If such an engine, depicted in [Fig. 44.1\(a\)](#), would be feasible, the produced work could then be returned to the reservoir. As a consequence of these operations, the reservoir, originally in equilibrium, is now in a new state, whereas the engine (being cyclic) and the environment haven't changed. Since a system in equilibrium (here the reservoir) can only be disequilibrated as a result of a change in its environment, the only way to satisfy the second law is by denying the possibility of a so-called *perpetual motion machine of the second kind* (PMM2) that, after having been started, produces work continuously without any net investment of energy. Rephrased, the Kelvin-Planck form of the second law thus states that a PMM2 is impossible, and, most notably, that a given quantity of heat cannot fully be converted to work: the heat-to-work conversion efficiency is always lower than 100%. Strikingly, no thermodynamic efficiency restriction exists in the opposite direction of the process: work energy can be fully converted into heat.

A work-producing cyclical engine that does not defy the second law is depicted in [Fig. 44.1\(b\)](#). Besides making work, it has an effect of rejecting a portion of the heat gained from the heat source reservoir to a heat sink reservoir. This heat rejection causes an equivalent reduction of the amount of work produced, as dictated by the first law.

Another useful statement of the second law was made by Clausius: "It is impossible to construct a device that operates in a cycle and produces no effect other than the transfer of heat from a region of lower temperature to a region of higher temperature." If such a device, depicted in [Fig. 44.2\(a\)](#), would be feasible, it would allow the transfer of heat from colder to warmer regions without any investment of work. Combined with the cycle of [Fig. 44.1\(b\)](#), it would then (1) allow the operation of a perpetual motion machine, (2) allow the operation of refrigerators that require no work investment, and (3) defy the empirical laws of heat transfer, which insist that heat is transferred only down the temperature gradient—from hot to cold regions—and not in the opposite direction.

A cooling cycle that does not defy the second law is depicted in [Fig. 44.2\(b\)](#). In addition to cooling (taking heat from) the lower temperature reservoir and delivering

**Figure 44.1** Comparison of power cycles (a) disallowed and (b) allowed by the second law: (a) a perpetual motion machine of the second kind (PMM2); (b) allowed power cycle.

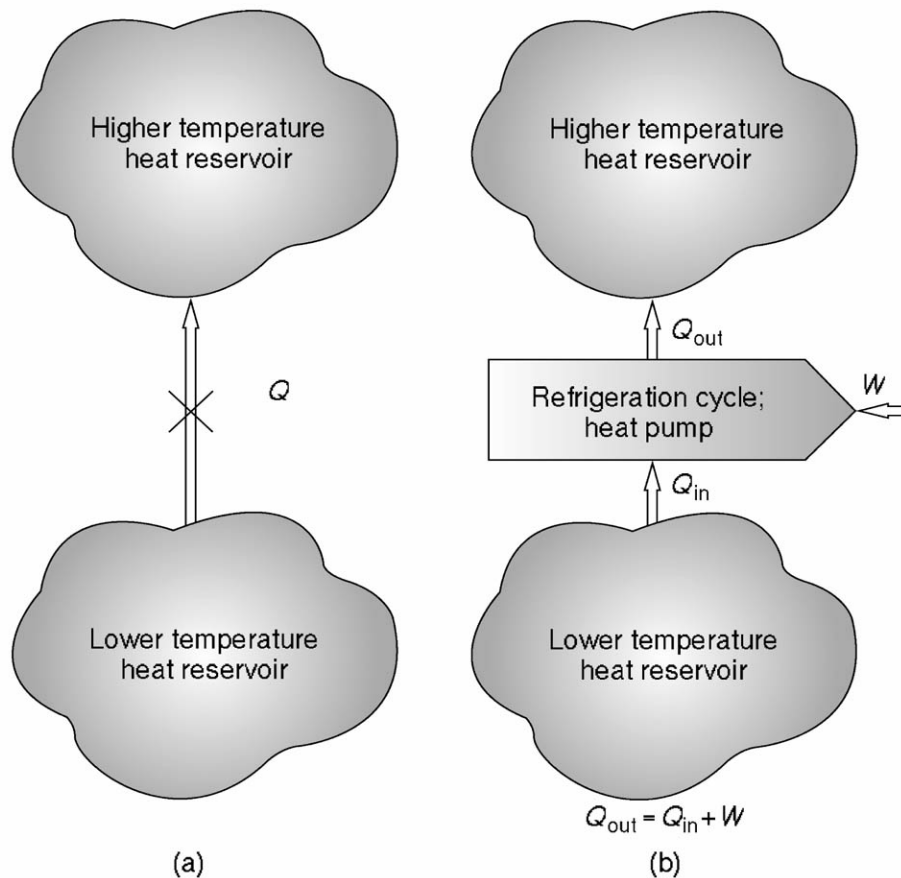


that amount of heat to the higher temperature reservoir, it has an effect of requiring some work from the outside, with an equivalent increase in the amount of heat delivered to the high-temperature reservoir, as dictated by the first law.

Two of the most rigorous statements of the second law are given as follows. Carathéodory proposed that "In the neighborhood of a **state** of a system there are states the system cannot reach through **adiabatic** processes" [Hatsopoulos and Keenan, 1965]. Furthermore, "A system having specified allowed states and an upper bound in volume can reach from any given state a **stable state** and leave no net effect on the **environment**" [Hatsopoulos and Keenan, 1965]. As explained already, it can be proven that these statements of the second law are fully consistent with the previous ones, in essence pronouncing axiomatically here that *stable equilibrium states exist*. A rudimentary exposition of the link between this statement and the impossibility of a PMM2 has been given in the preceding, and further clarification is available in the cited references.



**Figure 44.2** Comparison of conditions under which heat flow from low temperature to a higher temperature is (a) disallowed and (b) allowed by the second law: (a) heat flow from low to high temperature disallowed by the second law if the system environment does not change; (b) heat flow from low to high temperature allowed by the second law if work is supplied from the environment.



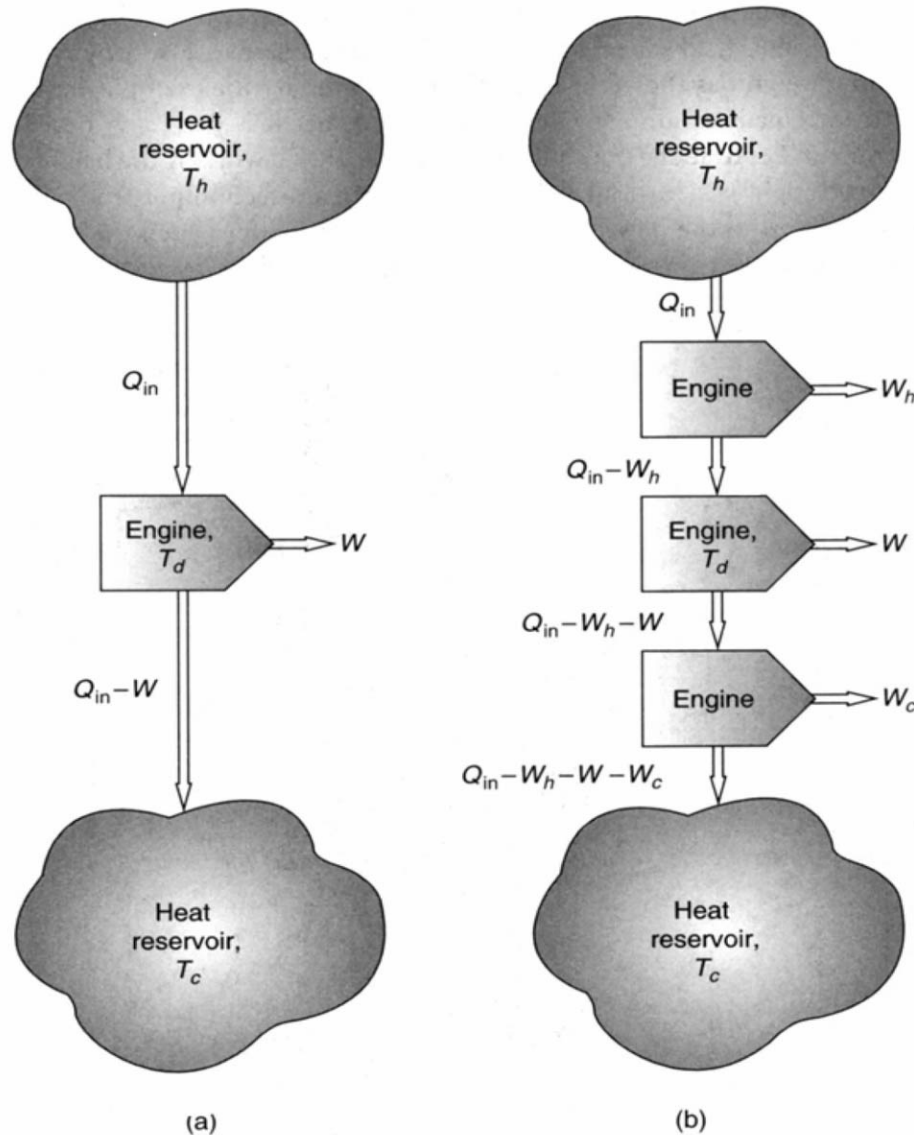
## 44.1 Reversibility

After the disturbing second law statement that heat cannot fully be converted into work, the next logical question concerns the maximal efficiency of such energy conversion. It is obvious from the first law, and even just from operational considerations, that the most efficient heat-to-work conversion process is one that incurs minimal diversions of energy into paths that do not eventually produce work. In a generalized heat-to-work conversion process such as that in Fig. 44.1(b), its components are two heat reservoirs at different temperatures, some type of device that produces work when interacting with these two heat reservoirs, and a sink for the produced work. It involves two heat transfer processes (one from the hot reservoir to the device, and one from the device to the cold reservoir), some heat-to-work conversion process inside the device, and a work transfer process from the device to the work sink. Losses in the amount of energy available for work production occur in practice in each of these processes. The most obvious from operational experience is the loss due to possible friction in the device, in the transfer from it to the work-sink, and in any fluid flow associated with the process. Friction converts work into heat and thus diminishes the net amount of work produced. From

another vantage point, the force used to overcome friction could have been used to produce work.

Many other phenomena, somewhat less operationally apparent, may also cause such losses. For example, it is easy to show that any heat transfer across a finite temperature difference causes a loss in the ability to produce work: if, as illustrated in Fig. 44.3(a), the device producing work  $W$  is at the temperature  $T_d$ , where  $T_c < T_d < T_h$ , one could have conceivably incorporated an additional heat-to-work conversion device between each of the two heat reservoirs and the original device, as shown in Fig. 44.3(b), and produced more work ( $W + W_h + W_c$ ) than the system shown in Fig. 44.3(b). These losses approach zero when the temperature differences between the heat reservoirs and the respective regions of the device with which they have a heat interaction also approach zero.

**Figure 44.3** Power production with reservoir-engine heat transfer across finite temperature differences,  $T_c < T_d < T_h$ : (a) work produced is only  $W$ ; (b) work produced is  $W + W_h + W_c$ .

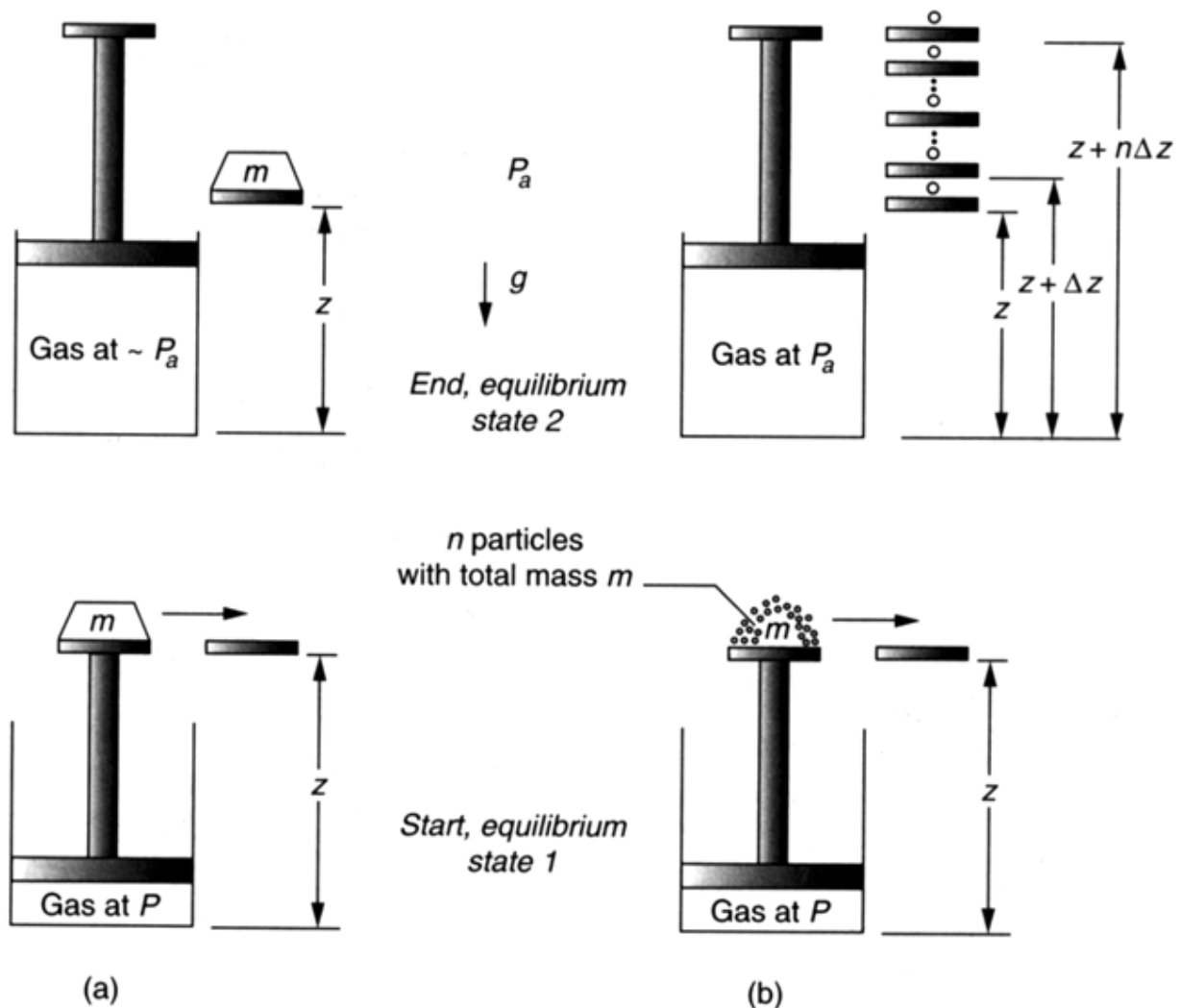


Having established that temperature difference (or more strictly, temperature gradient) is a thermodynamic driving potential for work production, one can easily recognize that there are other thermodynamic potentials for work production, such as pressure gradients (piston engines and turbines, for example), concentration gradients (batteries and fuel cells), and electric potential gradients (electric motors). Analogic to the arguments presented for losses due to finite temperature differences, it is easy to recognize that similar losses occur due to phenomena such as unrestrained expansion, the mixing of two different substances, the diffusion of a single substance in a mixture across a finite concentration difference, and the passage of current across an electric potential gradient, all when unaccompanied by commensurate production of work.

It is interesting to note that the work-production losses due to all of these processes approach zero if the respective thermodynamic driving potentials approach zero. Furthermore, processes that proceed toward the new equilibrium state due to the influence of a succession of such infinitesimally small driving potentials can in the limit also be reversed in direction without any residual change in the system and its environment. A good example is a frictionless system in which some amount of gas confined in a cylinder is held at a pressure larger than the ambient pressure by means of a piston loaded down with a weight, as shown in [Fig. 44.4\(a\)](#). To simplify the example, let us also assume that the system is originally at ambient temperature and that it is perfectly thermally insulated from the environment. Let us judge the ability of the system to do work by the height that it can lift some weight. If the weight is effortlessly slid to the side (in a direction perpendicular to gravity), the piston would pop up to the new force equilibrium state, under which the gas has expanded to a higher volume and lower pressure and temperature, but since the weight has remained at its original level the system has performed no useful work. The reversal of the process to bring the system *and the environment* to their original condition cannot be done, because work would have to be supplied for that purpose from the environment, thus changing its original state. This is an example of a process at its worst—producing no useful work and characterized by a large (finite) driving force and by irreversibility. Some engineering ingenuity can improve the situation vastly, in the extreme by replacing the single weight with a very large number of small weights having the same total mass as the original single weight, say a pile of fine sand grains [[Fig. 44.4\(b\)](#)]. Now we effortlessly slide one grain to the side, resulting in a very small rise of the piston and the production of some useful work through the consequent raising of the sand pile *sans* one grain. Removing one grain after another in the same fashion until the last grain is removed and the piston rests at the same equilibrium height as it did in the previously described worthless process will, as seems obvious, produce the most useful work given the original thermodynamic driving force. Furthermore, since the movements of the

piston are infinitesimally small, one could, without any residual effect on the environment, slide each lifted grain of sand back onto the piston slightly to recompress the gas and move the piston down a little to the position it has occupied before its incremental rise.

**Figure 44.4** Illustration of the approach to reversibility: (a) abrupt expansion with no weight raise, no useful work production (highly irreversible process); (b) gradual expansion with weight raise, some useful work production (more reversible process).



Generalizing, then, reversibility is synonymous with highest potential for producing useful work, and the degree of process inefficiency is proportional to its degree of irreversibility. The example also illustrates the practical impossibility of fully reversible

processes: the production of a finite amount of work at this maximal efficiency would either take forever or take an infinite number of such pistons, each lifting one grain at the same time (and frictionlessly at that). Practical processes are thus always a compromise between efficiency, rate, and amount of equipment needed.

## 44.2 Entropy

---

Like volume, temperature, pressure, and energy, entropy is a property characterizing thermodynamic systems. Unlike the other properties, it is not operationally and intuitively understood by novices, a situation not made simpler by the various approaches to its fundamental definition and by its increasing cavalier metaphysical application to fields as disparate as information theory, social science, economics, religion, and philosophy. Let us begin with the fundamental, prosaic definition of entropy,  $S$ , as

$$dS = \left( \frac{\delta Q}{T} \right)_{\text{rev}} \quad (44.1)$$

showing that the differential change in entropy,  $dS$ , in a process is equal to the ratio of the differential amount of heat interaction,  $\delta Q$ , that takes place if the process is reversible and the absolute temperature,  $T$ , during that process. For a finite-extent process between thermodynamic states 1 and 2, Eq. (44.1) can be integrated to give

$$S_2 - S_1 = \int_1^2 \left( \frac{\delta Q}{T} \right)_{\text{rev}} \quad (44.2)$$

It is noteworthy here that, since entropy is a property and thus uniquely defined at the states 1 and 2, the entropy change between these two states is always the same, whether the process is reversible or irreversible.

Entropy values of various materials are available in the literature: most books on thermodynamics and chemical and mechanical engineering handbooks include values of the entropy as a function of temperature and pressure; an extensive source for gas properties particularly related to high-temperature processes are the JANAF tables [Stull and Prophet, 1971]. The entropy values are given based on a common (but somewhat arbitrary) reference state, and its units are typically kJ/kg K, Btu/lb R, or kJ/kmol K if a molar basis is used. Although not used much in engineering practice, the fundamental reference state of entropy is 0 K, where according to the third law of thermodynamics

the entropy tends to 0.

If the process between states 1 and 2 is irreversible, then

$$\left(\frac{\delta Q}{T}\right)_{\text{irrev}} \neq \left(\frac{\delta Q}{T}\right)_{\text{rev}} \quad (44.3)$$

and it can be shown by using the second law that

$$S_2 - S_1 \geq \int_1^2 \left(\frac{\delta Q}{T}\right) \quad (44.4)$$

where the equality sign applies to reversible processes and the inequality to irreversible ones. Consequently, it can also be shown for isolated systems (i.e., those having no interactions with their environment) that

$$dS_{\text{isolated system}} \geq 0 \quad (44.5)$$

where again the equality sign applies to reversible processes and the inequality to irreversible ones. Since a thermodynamic system and its environment together form an isolated system by definition, Eq. (44.5) also states that the sum of the entropy changes of the system and the environment is  $\geq 0$ .

In another axiomatic approach to thermodynamics [Callen, 1985], the existence of the entropy property is introduced as an axiom and not as a derived property defined by Eq. (44.1). It is basically postulated that, out of all the new equilibrium states that the system may attain at the end of a process, the one that has the maximal entropy will be realized. Equation (44.1) and other more familiar consequences of entropy are then derived as corollaries of this maximum postulate.

Equation (44.5) and the other attributes of entropy lead to several eminent mathematical, physical, and other consequences. Mathematically, the inequality Eq. (44.5) indicates that many solutions may exist, which are bounded by the equality limit. In other words, out of all the new states 2 that an isolated system starting from state 1 can reach, only those at which the system entropy is larger are allowed; a unique (but practically unreachable) state 2 can be reached with no entropy change if the process is reversible. Also a mathematical consequence, the entropy maximum principle allows the establishment of the following equations,

$$dS = 0 \quad \text{and} \quad d^2S < 0 \quad \text{at state 2} \quad (44.6)$$

which then can be used in calculating the nature of the new equilibrium state.

Among the physical consequences, the reversible isentropic process provides a unique limit on process path, which also results in maximal efficiency. Also, the fact that the entropy of real isolated systems always rises provides guidance about which new states may be reached and which may not. Through some operational arguments, but primarily by using statistical mechanics on the molecular level, entropy was shown to be a measure of disorder. This supports the observation that isolated systems undergoing any process become less orderly; mess increases unless an external agent is employed to make order, all consistent with the entropy increase law. Profoundly, entropy is thus regarded to be the scientific indicator of the direction of time—"the arrow of time," after Eddington. The inevitable increase of disorder/entropy with time has led to deeper questions about the origin of the world and indeed of time itself, about the future of the universe in which disorder continuously grows, perhaps to a state of utter disorder (i.e., death), and has raised much philosophical discourse related to these issues. Significant attempts have been made to relate entropy and the second law to almost any human endeavor, including communications, economics, politics, social science, and religion [Bazarov (1964) and Georgescu-Roegen (1971), among many], not always in a scientifically convincing manner.

### 44.3 The Second Law for Bulk Flow

---

Based on Eq. (44.4), the entropy of a fixed amount of mass (which, by definition, does not give mass to its environment or receive any from it) receiving an amount  $\delta Q$  of heat from a heat reservoir will change as

$$dS \geq \frac{\delta Q}{T} \quad (44.7)$$

The amount of entropy produced in the mass due to reversible heat transfer from the reservoir is  $\delta Q/T$ , and, according to the second law as expressed by Eqs. (44.4) and (44.5), the amount of entropy produced due to the inevitable internal process irreversibilities,  $\sigma$ , is

$$\sigma \equiv dS - \frac{\delta Q}{T} \geq 0 \quad (44.8)$$

Addressing a transient and more general case, shown in Fig. 44.5, the rate of entropy

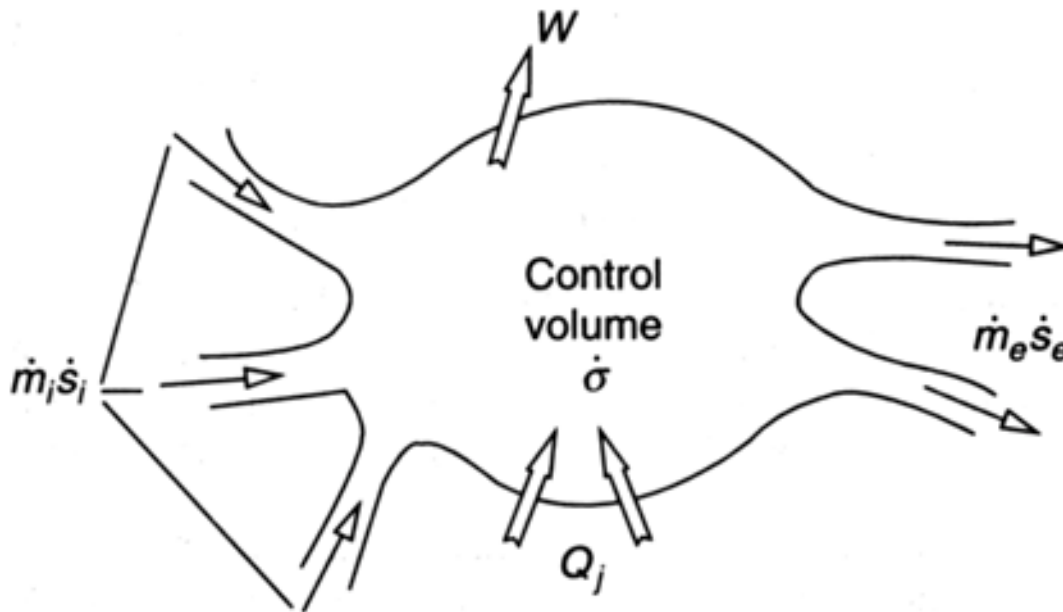


generation due to irreversibilities ( $\dot{\sigma}$ ) in a given control volume that has (1) work, heat ( $\dot{Q}$ ), and diffusion (mass transfer) interactions of  $k$  species with its environment, and (2) bulk molar flows ( $\dot{N}$ , moles/unit time) incoming (subscript  $i$ ) and exiting (subscript  $e$ ), can be expressed by simple entropy accounting as

$$\dot{\sigma} = \frac{dS_{cv}}{dt} - \sum_j \frac{\dot{Q}_j}{T_j} - \sum_i \left( \sum_k \dot{N}_k s_k \right)_i + \sum_e \left( \sum_k \dot{N}_k s_k \right)_e - \sum_k \left( \dot{N}_{k,e} - \dot{N}_{k,i} \right) s_{o,k} \geq 0 \quad (44.9)$$

where  $t$  is time,  $dS_{cv}/dt$  is the rate of entropy accumulation in the control volume, subscript  $j$  is the number of heat interactions with the environment,  $s$  is the specific molar entropy (entropy per mole), and subscript  $o$  refers to the conditions of this environment. Equation (44.9) expresses the fact that the entropy of this control volume changes due to heat interactions with its environment, due to the transport of entropy with entering and exiting bulk and diffusional mass flows, and due to internal irreversibilities (work interactions do not change the entropy).

**Figure 44.5** Entropy accounting for a control volume.





## 44.4 Applications

---

### Direction and Feasibility of Processes

Compliance with the second law is one of the criteria applied to examine the feasibility of proposed processes and patents. For example, we will examine such compliance for a steady state combustion process in which it is proposed that methane preheated to 80° C be burned with 20% excess air preheated to 200° C in a leak-tight well-insulated atmospheric-pressure combustor. Since the combustor is isolated from its environment, compliance with the second law will be inspected by calculating the entropy change in the reaction



where it may be noted that the number of moles of air was adjusted to reflect 20% excess air and that the excess oxygen and all of the nitrogen are assumed to emerge from the combustor unreacted. The entropy change can be expressed by using Eq. (44.9), which for this problem is reduced to

$$\sigma = \sum_p n_p s_p - \sum_r n_r s_r \quad (44.11)$$

where the subscripts  $p$  and  $r$  refer to the reaction products and reactants, respectively,  $n_i$  is the number of moles of species  $i$ , evident from Eq. (44.10), and  $s_i$  is the molar entropy of species  $i$ , to be found from the literature based on the temperature and partial pressure.

The calculation procedure is outlined in [Table 44.1](#), and further detail can be found in several of the references, such as Howell and Buckius [1992]. The partial pressures of the participating species are easily determined from the molar composition shown in Eq. (44.10). The temperature of the reactants is given, and that of the reaction products exiting the combustor,  $T_p$ , is calculated from the first law, where the enthalpies of the species, which are primarily dependent on the temperature, are obtained from reference tables or correlations. In this case it is found that  $T_p = 2241$  K. Based on this information, the entropies of the species are either found in the literature or calculated from available correlations or gas state equations. The results are listed in [Table 44.1](#), which contains, besides the entropy values, additional information that will be used in

another example later.

**Table 44.1** Analysis of a Methane Combustion Reaction

	Mole $n_i$	$y_i; p_i$ atm	$T$ K	$s$ kJ/ kmol · K	$s_o$ kJ/ kmol · K	$h_o = h(T_o, p_o)$ kJ/ kmol	$h(T, p)$ kJ/ kmol	$a_{ch}$ kJ/ kmol	$a$ kJ/ kmol	$n_i a$ kJ/ kmol CH <sub>4</sub>
Reactant										
CH <sub>4</sub>	1	0.08 05	35 3	213.5	186.3	−74 900	−72 854	830 745	824 685	824 685
O <sub>2</sub>	2. 4	0.19 32	47 3	232.6	205.0	0	5 309	−129	−3 036	−7 286
N <sub>2</sub>	9. 0 2 4	0.72 63	47 3	207.8	191.5	0	5 135	−102	176	1 585
Σ reactants	1 2 · 4 2 4	1.00 00								818 984
Product										
CO <sub>2</sub>	1	0.08 05	22 41	316.2	213.7	−393 800	−286 970	13 855	90 140	90 140
O <sub>2</sub>	0. 4	0.03 22	22 41	273.1	205.0	0	68 492	−4 568	43 639	17 456
N <sub>2</sub>	9. 0 2 4	0.72 63	22 41	256.1	191.5	0	64 413	−102	45 060	406 623
H <sub>2</sub> O (v)	2	0.16 10	22 41	270.3	188.7	−242 000	−157 790	4 138	64 031	128 062
Σ products	1 2 · 4 2 4	1.00 00								642 281

Dead state:  $p_o = 1$  atm,  $T_o = 298$  K; atmospheric composition (molar fractions):

$$N_2 = 0.7567, O_2 = 0.2035, H_2O = 0.0303, CO_2 = 0.0003$$

Application of these results to Eq. (44.11) gives

$$\begin{aligned}\sigma &= [1(316.2) + 0.4(273.1) + 9.024(256.1) + 2(270.3)] \\ &\quad - [1(213.5) + 2.4(232.6) + 9.024(207.8)] \\ &= 630.2 \text{ kJ}/(\text{kmol} \cdot \text{K}) > 0\end{aligned}\tag{44.12}$$

thus proving compliance of the proposed reaction with the second law and not denying its feasibility. It may be noted that such compliance is a necessary but not always sufficient condition for proving actual process feasibility; additional conditions, such as the existence of a spark in this example, may also be needed.

## Process Efficiency

Since reversible (isentropic for isolated systems) processes have maximal work-producing efficiency, they are often used as the basis for comparison in the definition of real process efficiencies, defined for work-producing processes (turbines, engines, etc.) as

$$\eta_{is} \equiv \frac{W_p}{W_{p,is}} \tag{44.13}$$

where  $\eta_{is}$  is the so-called "isentropic efficiency," subscript  $p$  refers to the work produced, and  $p,is$  to the work that would have been produced if the process was isentropic. For work-consuming processes (such as pumps, fans, etc.), the isentropic efficiency is defined as

$$\eta_{is} \equiv \frac{W_{c,is}}{W_c} \tag{44.14}$$

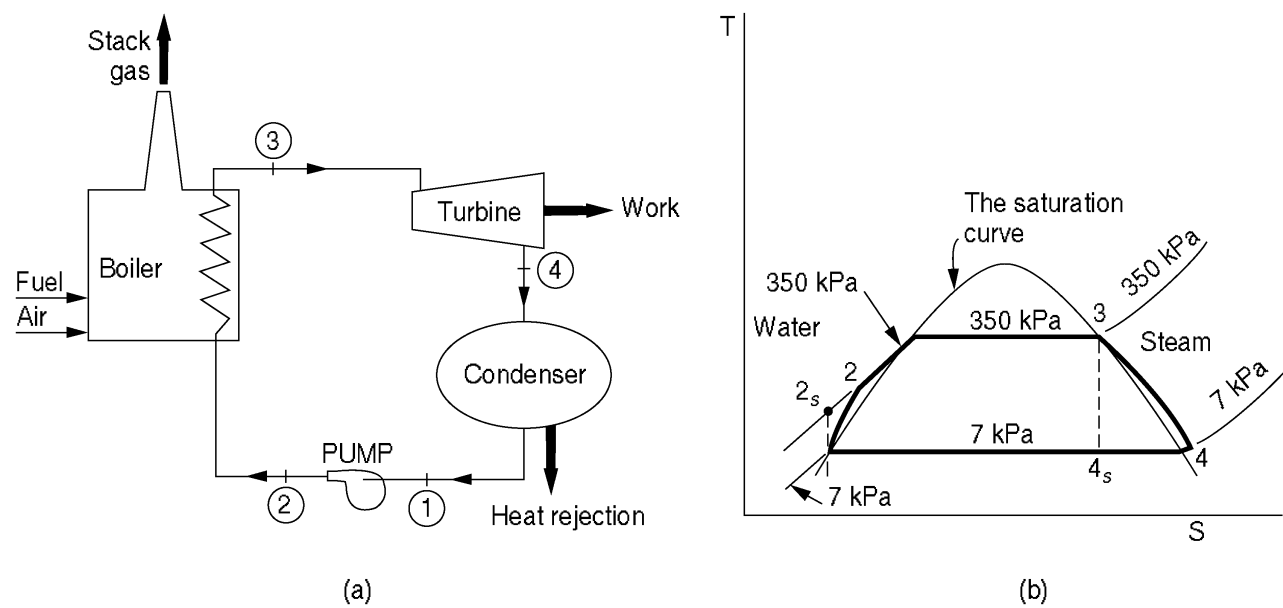
where subscript  $c$  refers to the work consumed in the process and  $c,is$  to the work that would have been consumed if the process was isentropic.

For example, the isentropic efficiency of the turbine in the simple Rankine cycle shown in Fig. 44.6(a) and qualitatively charted on the temperature-entropy ( $T$ - $S$ ) diagram of Fig. 44.6(b), in which superheated steam expands from the pressure  $p_3 = 350 \text{ kPa}$  to the condenser pressure  $p_4 = 7 \text{ kPa}$ , would be calculated using Eq. (44.11) as follows. If the expansion from  $p_3$  to  $p_4$  could have been performed

isentropically (i.e., reversibly in this case), the turbine would have produced the maximal amount of work, with the expansion of the steam terminating on the intersection point  $4_s$  of the  $p_4$  isobar and the isentrope descending vertically from point 3, with  $s_{4_s} = s_3$ . Having thus the values of  $s_{4_s}$  and  $p_4$  allows the determination of the **enthalpy**,  $h_{4_s}$ , shown in Table 44.2 (only the parameters in bold in this table would be addressed in this example). Assuming that the differences between the inlet-to-exit (states 3 and 4 in Fig. 44.6) elevations and kinetic energies of the steam are negligible compared to the respective enthalpy difference and that the turbine does not exchange heat with its environment, the isentropic work  $W_{p, is}$  can be calculated from the first law equation for the turbine control volume, reduced to

$$W_{p, is} = h_3 - h_{4_s} \qquad (44.15)$$

**Figure 44.6** The Rankine cycle example: (a) the Rankine power cycle; (b) temperature-entropy diagram of the Rankine cycle.



**Table 44.2** Data for Figure 44.6

State	$p$ , kPa	$T$ , K	$s$ , kJ/ kg · K	$h$ , kJ/ kg	$a$ , kJ/ kg	$h$ , kJ/ kg fuel	$a$ , kJ/ kg fuel
The Steam/Water Loop							

<b>1</b>	<b>7</b>	<b>312.23</b>	<b>0.5590</b>	<b>163.35</b>	1.36	2 260.4	18.8
<b>2<sub>s</sub></b>	<b>350</b>	<b>312.33</b>	<b>0.5590</b>	<b>163.70</b>	1.71	2 265.3	23.6
<b>2</b>	<b>350</b>	<b>312.38</b>	<b>0.5606</b>	<b>163.87</b>	1.40	2 267.6	19.3
<b>3</b>	<b>350</b>	<b>782.75</b>	<b>8.2798</b>	<b>3 506.00</b>	1 042.05	48 515.7	14 419.8
<b>4<sub>s</sub></b>	<b>7</b>	<b>313.15</b>	<b>8.2798</b>	<b>2 573.50</b>	109.55	35 611.9	1 515.9
<b>4</b>	<b>7</b>	<b>367.25</b>	<b>8.5821</b>	<b>2 676.10</b>	122.02	37 031.7	1 688.5
The Fuel, Air, and Stack Gas							
<b>5</b>	<b>101</b>	<b>298.15</b>		HHV = 50 019.00	$a_{\text{fuel}} = 51\,792.00$	50 019.0	51 792.0
<b>6</b>	<b>101</b>	<b>298.15</b>	<b>6.6999</b>	299.03	-0.70	6 158.1	-14.4
<b>7</b>	<b>101</b>	<b>423.15</b>	<b>7.3388</b>	459.81	79.40	9 929.0	1 714.5

Dead state:  $p_o = 1 \text{ atm}$ ,  $T_o = 298 \text{ K}$

Atmospheric composition (molar fractions):

$N_2 = 0.7567$ ,  $O_2 = 0.2035$ ,  $H_2O = 0.0303$ ,  $CO_2 = 0.0003$

In actuality the expansion is irreversible. The second law tells us that the entropy in the process will increase, with the expansion terminating on the same isobar  $p_4$ , but at a point 4 where  $s_4 > s_3 = s_{4s}$ , which is indeed true in this example, as shown in [Table 44.2](#). The work produced in this process,  $W_p$ , is again calculated from the first law:

$$W_p = h_3 - h_4 \quad (44.16)$$

Using Eq. (44.13) and the enthalpy values listed in [Table 44.2](#), the isentropic efficiency in this example is

$$\eta_{\text{is}} = \frac{h_3 - h_4}{h_3 - h_{4s}} = \frac{35\,060 - 26\,761}{35\,060 - 25\,735} = 0.89 \quad (44.17)$$

Although not shown here, the data in [Fig. 44.6](#) include calculations assuming that the isentropic efficiency of the pump is 0.7. State  $2_s$  in [Fig. 44.6\(b\)](#) would have been reached if the pumping process starting from state 1 was isentropic; using Eq. (44.14) and the given isentropic efficiency of the pump allows the determination of state 2 attained by the real process (note that states  $2_s$ , 2, and 3 are here all on the same 350 kPa isobar), where  $s_2$  is therefore greater than  $s_1$ .

Although isentropic efficiencies as defined here have a reasonable rationale and are used widely, important issue has been taken with their fundamental usefulness. Using [Fig. 44.6\(b\)](#), it can be argued that a better definition of such efficiency (to be named effectiveness,  $\varepsilon$ ) would be given if the work  $W_p$  actually produced, say in the expansion 3-4,  $(W_{p,3-4})_{\text{actual}}$  [the same as the  $W_p$  used in Eq. (44.14)], was compared with the

work that would have been produced if the expansion from 3 to 4 was reversible  $[(W_{p,3-4})_{\text{reversible}}]$ ,

$$\varepsilon = \frac{(W_{p,3-4})_{\text{actual}}}{(W_{p,3-4})_{\text{reversible}}} \quad (44.18)$$

rather than comparing it to work that would have been obtained in the isentropic expansion ending at a state  $4_s$ , which does not actually exist in the real process. An example of a deficiency of the isentropic efficiency ( $\eta_{\text{is}}$ ) compared with the effectiveness ( $\varepsilon$ ) is that it does not give credit to the fact that the steam exiting the turbine at state 4 has a higher temperature than if it had exited at state  $4_s$ , and thus has the potential to perform more work. This is especially poignant when, as usual, additional stages of reheat and expansion are present in the plant. This reasoning and further comments on effectiveness are given in the following section.

## Exergy Analysis

Based on the second law statement that only a fraction of heat energy can be converted into work, a very important application of the second law is the analysis of the potential of energy to perform useful work, that is, the examination of the "quality" of energy. To lay the grounds for such analysis, let us address for simplicity a steady flow and state open-flow system such as depicted in [Fig. 44.5](#), neglecting any potential and kinetic energy effects. The amount of work produced by the system is from the first law of thermodynamics,

$$W = \sum_i h_i m_i - \sum_e m_e h_e + \sum_j Q_j \quad (44.19)$$

where  $h$  is the specific enthalpy of each incoming or exiting stream of matter  $m$ .

As explained earlier, the maximal amount of work would be produced if the process is reversible, in which case entropy generation due to process irreversibilities ( $\sigma$ ) is zero (although entropy change of the system due to the heat interactions  $Q_j$  at the respective temperatures  $T_j$  is nonzero). Multiplying the steady state form of Eq. (44.10) by some constant temperature  $T$  (to give it energy units) and subtracting it from Eq. (44.19) yields

$$W_{\text{rev}} = \left( \sum_i m_i h_i - \sum_e m_e h_e \right) - T \left( \sum_i m_i s_i - \sum_e m_e s_e \right) + \sum_j Q_j \left( 1 - \frac{T}{T_j} \right) \quad (44.20)$$

As can be seen from this equation, the reversible work output in the process  $i \rightarrow e$  would be further maximized if the temperature at which the heat interaction occurs,  $T$ , would be the lowest practically applicable in the considered system, say  $T_o$ , yielding an expression for the maximal work output potential of the system, as

$$W_{\text{max}, i \rightarrow e} = \sum_i m_i (h_i - T_o s_i) - \sum_e m_e (h_e - T_o s_e) + \sum_j Q_j \left( 1 - \frac{T_o}{T_j} \right) \quad (44.21)$$

The term  $(h - T_o s)$  appearing at the right side of this equation is thus the measure of a system's potential to perform useful work and therefore of great thermodynamic significance. Composed of thermodynamic properties at a state and of the constant  $T_o$ , it is also a thermodynamic property at that state. This term is called the flow *exergy* or flow *availability function*,  $b$ :

$$b \equiv h - T_o s \quad (44.22)$$

Further examination of Eq. (44.22) shows that a system at state  $i$  would produce the maximal useful work when the new equilibrium state  $e$  is identical to the ambient conditions  $o$ , since at that state all the driving forces of the system—such as temperature, concentration, and pressure differences—are zero and the system cannot by itself produce any more useful work. The maximal work output, assuming mass conservation, is then

$$W_{\text{max}, i \rightarrow o} = \sum_i m_i [(h_i - h_o) - T_o (s_i - s_o)] \quad (44.23)$$

The term  $[(h - h_o) - T_o (s - s_o)]$  is also a property and is the measure of a system's potential to perform useful work between any given state and the so-called "dead state," at which the system can undergo no further spontaneous processes. This term is called *flow exergy* or *flow availability*,  $a$ :

$$a \equiv (h - h_o) - T_o(s - s_o) \quad (44.24)$$

Since enthalpy is the measure of the energy in flow systems, examination of Eqs. (44.21)–(44.24) shows clearly that the portion of the energy  $h$  that cannot be converted to useful work is the product  $T_o s$ .

The last term on the right side of Eq. (44.21) is the exergy of the heat sources  $Q_j$ , at the respective temperatures  $T_j$ , exchanging heat with the considered thermodynamic system. The form of this term is the Carnot cycle work output between  $T_j$  and the dead state temperature  $T_o$ , which indeed would produce the maximal work and is thus also the exergy of these heat sources by definition. Shaft and other mechanical power, and electric power, are pure exergy.

Turning now to real, irreversible steady state processes, their exergy accounting equation—developed using Eqs. (44.9), (44.21), and (44.24)—is

$$W = \sum_j Q_j \left(1 - \frac{T_o}{T_j}\right) + \sum_i m_i a_i - \sum_e m_e a_e - T_o \sigma \quad (44.25)$$

where the work output of the process in this control volume is, in the order of terms on the right side of the equation, produced due to (1) heat interactions  $Q_j$  at the respective temperatures  $T_j$ , with the control volume, and (2) and (3) the difference between the exergies flowing in and exiting the control volume, and is diminished by (4) the entropy generation due to process irreversibilities. This last term,  $T_o \sigma$ , amounts to the difference between the maximal work that could have been produced in a reversible process [Eqs. (44.21) and (44.24)] and the amount produced in the actual irreversible process [Eq. (44.25)]. It is called the *irreversibility (I)* or *lost work* of the process.

Beyond exergy changes due to temperature and pressure driving forces, multicomponent systems also experience exergy changes due to component mixing, phase change, and chemical reactions. It was found convenient to separate the exergy expression into its "physical" ( $a_{ph}$ ) and "chemical" ( $a_{ch}$ ) constituents,

$$a = a_{ph} + a_{ch} \quad (44.26)$$

where the *physical exergy* is the maximal work obtainable in a reversible physical process by a system initially at  $T$  and  $p$  and finally at the dead state  $T_o, p_o$ , and the *chemical exergy* is the maximal work obtainable from a system at dead state conditions  $T_o, p_o$ , initially at some species composition and finally at the dead state composition (such as the datum level composition of the environment). The total flow exergy,



showing the thermal, mechanical, and chemical flow exergy components (segregated by the double brackets) is

$$a = [(h - h_o) - T_o(s - s_o)](\text{thermal}) + \left[ \left[ \frac{\mathbf{v}^2}{2} + gz \right] \right] (\text{mechanical}) + \left[ \left[ \sum_k x_k (\mu_k^\circ - \mu_{o,k}) \right] \right] (\text{chemical}) \quad (44.27)$$

where  $\mathbf{v}$  is the flow velocity,  $g$  the gravitational acceleration,  $z$  the flow elevation above a zero reference level,  $x_k$  is the molar fraction of species  $k$  in the mixture,  $\mu_k^\circ$  is the **chemical potential** of species  $k$  at  $T_o$  and  $p_o$ , and  $\mu_{k,o}$  is the chemical potential of species  $k$  at the system dead state defined by  $T_o$ ,  $p_o$ , and the dead state composition.

The general transient exergy accounting equation that includes both physical and chemical exergy is

$$\begin{aligned} \frac{dA_{cv}}{dt}(\text{rate of exergy storage}) &= \sum_j \left( 1 - \frac{T_o}{T_j} \right) \dot{Q}_j - \left( \dot{W}_{cv} - p_o \frac{dV_{cv}}{dt} \right) + \sum_i \dot{m}_i a_i \\ &\quad - \sum_e \dot{m}_e a_e (\text{rates of exergy transfer}) \\ &\quad - \dot{I}_{cv} (\text{rate of exergy destruction}) \end{aligned} \quad (44.28)$$

where  $A_{cv}$  is the exergy of the control volume,  $\dot{W}_{cv}$  is the work performed by the control volume,  $p_o(dV_{cv}/dt)$  is the work due to the transient volume change of the control volume,  $a_i$  and  $a_e$  are the total flow exergies composed of both the physical and chemical components, and  $\dot{I}_{cv}$  ( $\equiv T_o \dot{\sigma}$ ) is the irreversibility, that is, the rate of exergy destruction. A more detailed breakdown of exergy, in differential equation form, is given in Dunbar *et al.* [1992].

For a *closed system* (i.e., one that does not exchange mass with its environment but can have work and heat interactions with it) at a state defined by the specific internal energy  $u$ , specific volume  $v$ , and specific entropy  $s$ , the expression for the exergy,  $a^\circ$ , is

$$a^\circ = \left( u + \frac{\mathbf{v}^2}{2} + gz - u_o \right) + p_o(v - v_o) - T_o(s - s_o) \quad (44.29)$$

Selection of the dead state for evaluating exergy is based on the specific application considered. Thus, for example, the dead state for the analysis of a water-cooled power plant operating at a certain geographic location should consist of the cooling water temperature and the ambient atmospheric pressure and composition, all at that location. Much work has been performed in defining "universal" dead states for processes and materials [Szargut *et al.*, 1988].

**Example—Exergy Analysis of a Power Plant.** To demonstrate the calculation procedure and the benefits of exergy analysis, we will perform such an analysis on the simple Rankine cycle described in Fig. 44.6. The fuel used in the boiler is methane, undergoing the same combustion reaction as described in Eq. (44.10). Both fuel and air enter the boiler at 25° C, 1 atm, and the combustion products exit the stack at 150° C.

Addressing the steam/water (single component) Rankine cycle loop first, the values of  $p$ ,  $T$ ,  $s$ , and  $h$  are already available from the example given and are listed in Table 44.2. Given the dead state conditions listed under the table, the flow exergy ( $a$ ) per kg of the water and steam is calculated from Eq. (44.24) and listed in the table column to the left of the vertical dividing line.

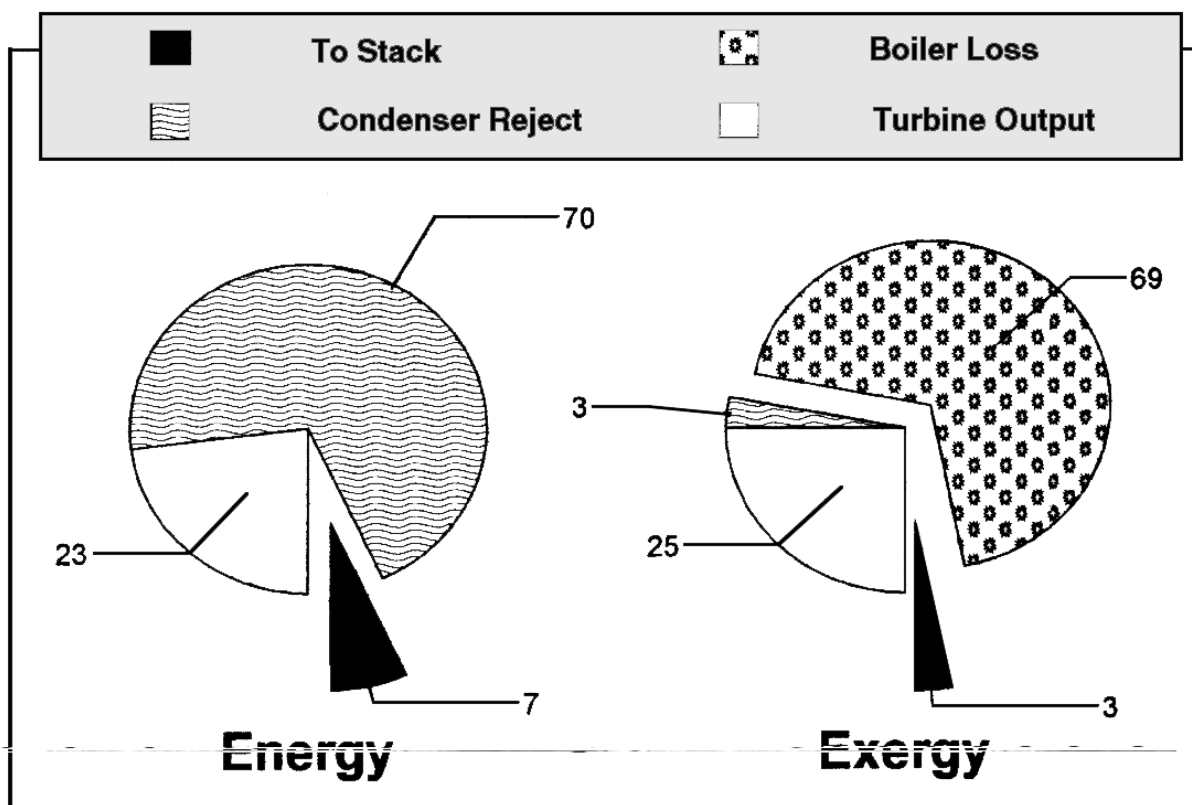
The three bottom rows of the table list the properties of the inflowing fuel and air and of the stack exhaust gas. The enthalpy of the fuel is its higher heating value (HHV), which, along with exergy, is obtained from fuel property tables [Szargut *et al.*, 1988; Howell and Buckius, 1992; Moran and Shapiro, 1992]. The specific flow exergy values of the air and stack gas are calculated from Eq. (44.28) (with negligible kinetic and potential energy components) and listed to the left of the vertical dividing line.

It is sensible to analyze the power system based on the fuel energy and exergy input, in other words, to determine how this fuel input is distributed among the different system components and processes. To that end we determine, per unit mass of fuel, the mass of cycle steam [here found to be 13.8 (kg steam)/(kg fuel) by applying the energy conservation equation to the boiler] and the mass of air and stack gas [20.6 (kg air)/(kg fuel) and 21.6 (kg exhaust gas)/(kg fuel) by applying Eq. (44.10)]. These results are listed in the last two columns of the table.

Examination of the results is encouraged, and some of the most important ones, which describe the fuel energy and exergy distribution among the different system components/processes, are shown in Fig. 44.7. Energy analysis indicates that the major energy loss, 70%, is due to the heat rejected in the condenser. Examination of the exergy analysis chart shows, however, that this large energy loss amounts to only 3% of the fuel exergy, and even complete elimination of the condenser heat rejection (if it were at all feasible) would increase the cycle efficiency by only three percentage points. Of

course, this is because the heat rejected in the condenser is at a low temperature, only slightly elevated above that of the ambient, and thus has commensurately little potential to perform work despite its large energy. Another very significant difference is the fact that the exergy analysis identifies the major losses, 69%, to be in the boiler, due to the combustion and gas-to-steam/water heat transfer processes, whereas the energy analysis associates no loss to these processes. Finally, the exergy analysis attributes much less loss to the stack gas than the energy analysis does. Notably, the turbine output has almost the same percentage in both analyses, which is because the output is shaft power (i.e., pure exergy), and the HHV and exergy of the fuel are nearly identical. In conclusion, it is only exergy analysis that can correctly identify and evaluate the losses (irreversibilities) that diminish the ability of processes to perform useful work.

**Figure 44.7** Energy and exergy breakdown for the Rankine power cycle example.



**Example—Exergy Analysis of a Combustion Process.** The methane combustor

analyzed earlier, for which the data are given in [Table 44.1](#), has a 100% energy efficiency because its enclosure is adiabatic and impermeable, as can also be ascertained by performing a first law energy conservation check using the data in [Table 44.1](#). The intuitive implication of this result is that the combustor is thus perhaps "ideal" and requires no further technological improvement. It is interesting therefore to examine the effect that it has on the potential of the invested consumables (i.e., the fuel and other reactants) on producing useful work. The flow exergy  $a$  and the chemical exergy  $a_{\text{ch}}$  are calculated by using Eq. (44.27), and property values from gas tables or correlations given in the literature, with the dead state (including the atmospheric composition) described at the bottom of [Table 44.1](#). The calculation results are summarized in [Table 44.1](#), showing, first, that the total exergy of the combustion products (642281 kJ/kmol  $\text{CH}_4$ ) is only 78.4% of the original exergy of the reactants (818984 kJ/kmol  $\text{CH}_4$ ) that they possessed prior to combustion. This combustion process, although ideal from the energy standpoint, destroyed 21.6% of the original exergy. Practical combustion processes destroy even more exergy and are typically the largest cause for the lost work (irreversibility) in current-day fossil fuel power plants. Whereas energy (first law) analysis identified this process as "ideal," exergy analysis was unique in its ability to recognize and quantify this power production efficiency loss, which we can subsequently attempt to reduce. More information on this approach and the reasons for combustion irreversibility can be found in Dunbar and Lior [[1994](#)].

Examination of [Table 44.1](#) also shows that the exergy of the fuel is dominant among the reactants and that the chemical exergy,  $a_{\text{ch}}$ , is small relative to the overall exergy.

These examples have clearly demonstrated the unique role and importance of second law analysis. Integration of such analysis with conventional first law (energy) analysis is necessary and increasingly seen in practice [*cf.* [Bejan, 1988](#); [Moran and Shapiro, 1992](#)].

## Defining Terms

**Adiabatic:** A process or surface allowing only work interactions. Since most introductory thermodynamics texts consider only work and heat interactions, *adiabatic* is most often interpreted as a process or surface allowing no heat interactions.

**Chemical potential:** A potential given by  $\mu_k \equiv h_k - T s_k$ , where  $h_k$  and  $s_k$  are, respectively, the enthalpy and entropy of the species  $k$ , and  $T$  is the temperature.

**Cycle:** A series of processes that bring a thermodynamic system back to its original state.

**Enthalpy:** Given by  $h \equiv u + pv$ , where  $u$  is the internal energy,  $p$  is the pressure, and  $v$  is the specific volume.

**Environment:** The environment of a thermodynamic system consists of everything outside of that system that could conceivably have some influence on it.

**Equilibrium:** A state in which all of the properties of a system that is not subjected to interactions do not change with time. Equilibrium states may be of various types: stable, neutral, metastable, and unstable.

**Heat reservoir:** A system in a stable equilibrium state such that, when subjected to finite heat interactions, its temperature remains constant. Although this is just a concept for simplifying the exposition of thermodynamics, the atmosphere, oceans, lakes, and rivers were often considered as practical examples of such a reservoir. It is noteworthy that the large and ever-increasing magnitudes of human-made heat interactions with these natural elements now invalidate their definition as "heat reservoirs" in this thermodynamic sense, because these interactions in fact change their temperature and cause increasingly adverse environmental effects.

**Stable state:** A system is in a stable state (or stable equilibrium state) if a finite change of state of the system cannot occur without leaving a corresponding finite change in the state of the environment. In other words a finite external influence would be needed to budge a system out of its stable state; small natural pulses and fluctuations would not suffice.

**State:** The state of a thermodynamic system is defined by all of the properties of the system, such as pressure, temperature, volume, composition, and so forth. A system returns to the same state when all the properties attain their original values.

**Thermodynamic system:** A thermodynamic system is whatever we enclose for the purpose of a study by a well-defined surface, which may be either material or imaginary and which can be isolated from everything else (i.e., from the "environment").

## References

- Bazarov, I. P. 1964. *Thermodynamics*. Pergamon Press, Oxford, England.
- Bejan, A. 1988. *Advanced Engineering Thermodynamics*. John Wiley & Sons, New York.
- Callen, H. B. 1985. *Thermodynamics and an Introduction to Thermostatistics*, 2nd ed. John Wiley & Sons, New York.
- Dunbar, W. R., Lior, N., and Gaggioli, R. 1992. The component equations of energy and exergy. *ASME J. Energy Resour. Technol.* (114): 75–83.
- Dunbar, W. R. and Lior, N. 1994. Sources of combustion irreversibility. *Comb. Sci. Technol.* 103:41–61.
- Georgescu-Roegen, N. 1971. *The Entropy Law and the Economic Process*. Harvard

University Press, Cambridge, MA.

Hatsopoulos, G. N. and Keenan, J. H. 1965. *Principles of General Thermodynamics*. John Wiley & Sons, New York.

Howell, J. R. and Buckius, R. O. 1992. *Fundamentals of Engineering Thermodynamics*. McGraw-Hill, New York.

Moran, M. J. and Shapiro, H. N. 1992. *Fundamentals of Engineering Thermodynamics*. John Wiley & Sons, New York.

Stull, D. R. and Prophet, H. 1971. *JANAF Thermochemical Tables*, 2nd ed. NSRDS-NBS 37. National Bureau of Standards, Washington, DC.

Szargut, J., Morris, D. R., and Steward, F. R. 1988. *Exergy Analysis of Thermal, Chemical and Metallurgical Processes*. Hemisphere, New York.

### **Further Information**

Anderson, E. E. *Thermodynamics*. 1994. PWS, Boston, MA. Integrates first and second law analysis.

Sonntag, R. E. and Van Wylen, G. J. *Introduction to Thermodynamics Classical and Statistical*. John Wiley & Sons, New York. Widely used textbook.

Sandler, S. I., Orbey, H. "The Thermodynamics of Solutions"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# The Thermodynamics of Solutions

---

## 45.1 Fundamentals

Real Liquid Mixtures—Excess Property Description • Real Mixtures—Equation of State Description • Solutions in the Solid Phase

## 45.2 Applications

**Stanley I. Sandler**

*University of Delaware*

**Hasan Orbey**

*University of Delaware*

It is important to know the thermodynamic properties of mixtures in order to be able to predict phase behavior such as vapor-liquid, liquid-liquid, vapor-liquid-liquid, and solid-liquid equilibria necessary to design separation and purification processes, to predict when a mixture will boil or will precipitate, and to estimate the equilibrium extents of homogeneous and heterogeneous chemical reactions. Also, mixture enthalpies are needed to design heat exchange equipment and entropies for compressor calculations.

## 45.1 Fundamentals

---

There are a collection of thermodynamic properties that describe a pure fluid such as the state variables temperature  $T$  and pressure  $P$ , and the system-intensive variables, specific volume (that is, the volume per mole of substance)  $v$ , specific internal energy  $u$ , specific enthalpy  $h$ , specific entropy  $s$ , specific Helmholtz free energy  $a$ , and specific Gibbs free energy  $g$ . The thermodynamics of mixtures would be relatively simple if, for any specific pure fluid thermodynamic variable  $\Theta$ , the corresponding mixture property  $\Theta_{\text{mix}}$  at the same temperature and pressure was

$$\Theta_{\text{mix}}(T, P, x_i) = \sum_i x_i \Theta_i(T, P) \quad (45.1)$$

where the subscripts  $i$  and mix indicate the property of pure component  $i$  and the mixture, respectively, and  $x_i$  is the mole fraction of species  $i$ . However, Eq. (45.1) is not generally valid for several reasons. First, for the entropy and properties that depend on the entropy (such as the Gibbs free energy  $g = h - Ts$  and the Helmholtz free energy  $a = u - Ts$ ) an additional term arises



because the volume available to a molecule of species  $i$  changes in the mixing process from the pure component volume  $n_i v_i$  to the total volume of the mixture  $n_t v_{\text{mix}}$ . For a mixture of ideal gases  $v_i = v_{\text{mix}} = RT/P$ , and the additional term that appears is as follows:

$$\begin{aligned} s_{\text{mix}}^{IG}(T, P, x_i) &= \sum_i x_i s_i^{IG}(T, P) - R \sum_i x_i \ln x_i \\ g_{\text{mix}}^{IG}(T, P, x_i) &= \sum_i x_i g_i^{IG}(T, P) + RT \sum_i x_i \ln x_i \\ a_{\text{mix}}^{IG}(T, P, x_i) &= \sum_i x_i a_i^{IG}(T, P) + RT \sum_i x_i \ln x_i \end{aligned} \quad (45.2)$$

The simplest liquid mixture is an ideal mixture (denoted by the superscript  $IM$ ) for which Eq. (45.1) is valid for the internal energy and volume *at all temperatures and pressures*, that is,

$$\begin{aligned} v_{\text{mix}}^{IM}(T, P, x_i) &= \sum_i x_i v_i(T, P) \\ h_{\text{mix}}^{IM}(T, P, x_i) &= \sum_i x_i h_i(T, P) \end{aligned} \quad (45.3)$$

In this case, one can show [Sandler, 1989] that

$$\begin{aligned} s_{\text{mix}}^{IM}(T, P, x_i) &= \sum_i x_i s_i(T, P) - R \sum_i x_i \ln x_i \\ g_{\text{mix}}^{IM}(T, P, x_i) &= \sum_i x_i g_i(T, P) + RT \sum_i x_i \ln x_i \\ a_{\text{mix}}^{IM}(T, P, x_i) &= \sum_i x_i a_i(T, P) + RT \sum_i x_i \ln x_i \end{aligned} \quad (45.4)$$

Although Eq. (45.4) is similar to Eq. (45.2), there is an important difference. Here, the pure component and mixture properties are those of the real fluid at the temperature and pressure of the mixture, not those of ideal gases.

Equation (45.2) may not apply to real gases because they are not ideal gases. Also, Eq. (45.4) does not apply to most liquid mixtures since Eq. (45.3) is not, in general, valid. Another difficulty that arises is that at the temperature and pressure of interest, the pure components may not exist in the same state of aggregation as the mixture. An example of this is a mixture containing a dissolved gas or a solid in a liquid.

To describe the thermodynamic behavior of a condensed phase (liquid or solid), an excess

property or equation-of-state model may be used, depending on the mixture. For a vapor mixture, an equation of state is generally used. Both these descriptions are reviewed below.

## Real Liquid Mixtures—Excess Property Description

For a liquid mixture formed by mixing pure liquids at constant temperature and pressure the following description is used:

$$\Theta_{\text{mix}}(T, P, x_i) = \Theta_{\text{mix}}^{\text{IM}}(T, P, x_i) + \Theta^{\text{EX}}(T, P, x_i) \quad (45.5)$$

Here,  $\Theta^{\text{EX}}(T, P, x_i)$  is the additional change in the thermodynamic property on forming a mixture from the pure components above that on forming an ideal mixture. This excess thermodynamic property change on mixing is a function of temperature, pressure, and mixture composition. Expressions for several thermodynamic properties of real mixtures are given below:

$$\begin{aligned} v_{\text{mix}}(T, P, x_i) &= \sum_i x_i v_i(T, P) + v^{\text{EX}}(T, P, x_i) \\ u_{\text{mix}}(T, P, x_i) &= \sum_i x_i u_i(T, P) + u^{\text{EX}}(T, P, x_i) \\ s_{\text{mix}}(T, P, x_i) &= \sum_i x_i s_i(T, P) - R \sum_i x_i \ln x_i + s^{\text{EX}}(T, P, x_i) \\ g_{\text{mix}}(T, P, x_i) &= \sum_i x_i g_i(T, P) + RT \sum_i x_i \ln x_i + g^{\text{EX}}(T, P, x_i) \\ a_{\text{mix}}(T, P, x_i) &= \sum_i x_i a_i(T, P) + RT \sum_i x_i \ln x_i + a^{\text{EX}}(T, P, x_i) \end{aligned} \quad (45.6)$$

An important concept in mixture solution thermodynamics is the partial molar property  $\bar{\Theta}_i(T, P, x_i)$ , defined as follows:

$$\bar{\Theta}_i(T, P, x_i) = \left( \frac{\partial n_i \Theta_{\text{mix}}}{\partial n_i} \right)_{T, P, n_{j \neq i}} \quad (45.7)$$

It follows [Van Ness, 1964] from the fact that any solution property that is linear and homogeneous in the number of moles is related to its partial molar properties as follows:

$$\Theta_{\text{mix}}(T, P, x_i) = \sum_i x_i \bar{\Theta}_i(T, P, x_i) \quad (45.8)$$

Also, from Eq. (45.6),

$$\bar{\Theta}_i(T, P, x_i) = \bar{\Theta}_i^{IM}(T, P, x_i) + \bar{\Theta}_i^{EX}(T, P, x_i) \quad (45.9)$$

From this we have, as examples,

$$\begin{aligned} \bar{v}_i(T, P, x_i) &= v_i(T, P) + \bar{v}_i^{EX}(T, P, x_i) \\ \bar{u}_i(T, P, x_i) &= u_i(T, P) + \bar{u}_i^{EX}(T, P, x_i) \\ \bar{s}_i(T, P, x_i) &= s_i(T, P) - R \ln x_i + \bar{s}_i^{EX}(T, P, x_i) \\ \bar{g}_i(T, P, x_i) &= g_i(T, P) + RT \ln x_i + \bar{g}_i^{EX}(T, P, x_i) \end{aligned} \quad (45.10)$$

By definition, all excess properties must vanish in the limit of a pure component. Several other thermodynamic variables of special interest are the chemical potential,  $\mu_i$ ,

$$\begin{aligned} \mu_i(T, P, x_i) &= \mu_i^o(T, P) + RT \ln x_i(T, P, x_i) + \bar{g}_i^{EX}(T, P, x_i) \\ &= \mu_i^o(T, P) + RT \ln x_i \gamma_i(T, P, x_i) \end{aligned} \quad (45.11)$$

the activity,  $a_i$ ,

$$a_i(T, P, x_i) = \exp \left( \frac{\mu_i - \mu_i^o}{RT} \right) \quad (45.12)$$

the activity coefficient,  $\gamma_i$ ,

$$\gamma_i(T, P, x_i) = \exp \left( \frac{\bar{g}_i^{EX}(T, P, x_i)}{RT} \right) \quad (45.13)$$

the fugacity,  $f_i$ ,

$$RT \ln \left( \frac{\hat{f}_i(T, P, x_i)}{\hat{f}_i^o(T, P)} \right) = \mu_i(T, P, x_i) - \mu_i^o(T, P) \quad (45.14)$$

and the fugacity coefficient,  $\hat{\phi}_i$ ,

$$\hat{\phi}_i(T, P, x_i) = \frac{\hat{f}_i}{x_i P} \quad (45.15)$$

In the definitions above,  $\mu_i^o$  is the chemical potential of pure species  $i$  at the temperature, pressure, and state of aggregation of the mixture, and  $\hat{f}_i^o$  is its fugacity in that state; this pure component state is referred to as the standard state for the component. Also,  $\gamma_i(x_i \rightarrow 1) = 1$ . If the pure species does not exist in the same state of aggregation as the mixture, as with a gas or solid dissolved in a liquid solvent, then other standard states are chosen. For example, if a hypothetical pure component state is chosen as the standard state with properties extrapolated based on the behavior of the component at infinite dilution, then

$$\mu_i(T, P, x_i) = \mu_i^*(T, P) + RT \ln(x_i \gamma_i^*) \quad (45.16)$$

where  $\mu_i^*$  is the chemical potential in this so-called Henry's law standard state and  $\gamma_i^*$  is an activity coefficient defined so that  $\gamma_i^*(T, P, x_i \rightarrow 0) = 1$ . Another choice for the standard state of such components is the hypothetical ideal 1 molal solution, which is commonly used for electrolytes and other compounds.

The excess properties have generally been determined by experiment and then fitted to algebraic models. These models must satisfy the boundary condition of vanishing in the limit of a mixture going to a pure component. That is, in a binary mixture,  $\Theta^{EX}(T, P, x_i)$  must go to zero as either  $x_1 \rightarrow 1$  or  $x_2 \rightarrow 1$ . One solution model of historical interest is the strictly regular solution model, for which  $u^{EX} = g^{EX} = x_1 x_2 \omega$  and  $s^{EX} = 0$ , where, from simple molecular theory,  $\omega$  is the energy change on interchanging molecules between the pure species and the mixture. Another model is the athermal solution model, for which there is no excess internal energy change of mixing—that is,  $u^{EX} = 0$ —but there is an excess entropy change so that  $g^{EX} = -T s^{EX}$ . Most excess property models currently in use are a combination of an athermal part to account for the entropy change on mixing and a modified regular solution part to account for energy change on mixing. Several excess property models are listed in Table 45.1. There are also excess property models that are predictive (instead of merely correlative), such as the UNIFAC and ASOG models. These are described in detail elsewhere [Fredenslund, 1977; Sandler, 1989].

**Table 45.1** Excess Gibbs Free and Activity Coefficient Models

---

*The two-constant Margules equation*

$$g^{EX} = x_1 x_2 \{ A + B(x_1 - x_2) \}$$

which leads to

$$RT \ln \gamma_1 = \alpha_1 x_2^2 + \beta_1 x_2^3 \quad \text{and} \quad RT \ln \gamma_2 = \alpha_2 x_1^2 + \beta_2 x_1^3$$

with  $\alpha_i = A + 3(-1)^{i+1} B$  and  $\beta_i = 4(-1)^i B$ .

*The van Laar model*

$$g^{EX} = x_1 x_2 \frac{2a q_1 q_2}{x_1 q_1 + x_2 q_2}$$

which leads to

$$\ln \gamma_1 = \frac{\alpha}{\left[1 + \frac{\alpha x_1}{\beta x_2}\right]^2} \quad \text{and} \quad \ln \gamma_2 = \frac{\beta}{\left[1 + \frac{\beta x_2}{\alpha x_1}\right]^2}$$

with  $\alpha = 2q_1 a$  and  $\beta = 2q_2 a$ .

*The nonrandom two-liquid (NRTL) model*

$$\frac{g^{EX}}{RT} = x_1 x_2 \left( \frac{\tau_{21} G_{21}}{x_1 + x_2 G_{21}} + \frac{\tau_{12} G_{12}}{x_2 + x_1 G_{12}} \right)$$

which leads to

$$\ln \gamma_1 = x_2^2 \left[ \tau_{21} \left( \frac{G_{21}}{x_1 + x_2 G_{21}} \right)^2 + \frac{\tau_{12} G_{12}}{(x_2 + x_1 G_{12})^2} \right]$$

with  $\ln G_{12} = -\alpha \tau_{12}$  and  $\ln G_{21} = -\alpha \tau_{21}$ .

*The Flory-Huggins model (for polymer solutions)*

$$\frac{g^{EX}}{RT} = x_1 \ln \frac{\phi_1}{x_1} + x_2 \ln \frac{\phi_2}{x_2} + \chi (x_1 + m x_2) \phi_1 \phi_2$$

which leads to

$$\ln \gamma_1 = \ln \frac{\phi_1}{x_1} + \left(1 - \frac{1}{m}\right) \phi_2 + \chi \phi_2^2 \quad \text{and} \quad \ln \gamma_2 = \ln \frac{\phi_2}{x_2} + (m - 1) \phi_1 + \chi \phi_1^2$$

with  $\phi_1 = x_1 / (x_1 + m x_2)$  and  $\phi_2 = m x_2 / (x_1 + m x_2)$ .

---

## Real Mixtures—Equation of State Description

Whereas the excess property description is used for the thermodynamic description of solids and liquids, a volumetric equation of state is typically used for vapor mixtures and may be used for liquid mixtures of hydrocarbons (including with light gases). The simplest volumetric equation of state is the ideal gas (which may be applicable only at low pressures):

$$P = \frac{RT}{v} \quad (45.17)$$

where  $R$  is the gas constant. The first equation of state that qualitatively described both vapors and liquids was that of van der Waals [1890]:

$$P = \frac{RT}{v - b} - \frac{a}{v^2} \quad (45.18)$$

Here  $b$  is interpreted as the hard-core volume of the molecules, and  $a$  is a parameter that represents the strength of the attractive interaction energy. Consequently, the first term results from molecular repulsions, and the second term from attractions. Many equations of state now used have this same structure, such as the Peng-Robinson equation:

$$P = \frac{RT}{v - b} - \frac{a(T)}{v(v + b) + b(v - b)} \quad (45.19)$$

where the  $a$  parameter has been made a function of temperature to better describe the pure component vapor pressure.

In order to apply equations of state developed for pure components to mixtures, expressions are needed that relate the mixture parameters, such as  $a$  and  $b$  above, to those of the pure fluids. This is done by using mixing and combining rules. The simplest mixing rules are those also attributed to van der Waals:

$$a_{\text{mix}} = \sum_i \sum_j x_i x_j a_{ij} \quad \text{and} \quad b_{\text{mix}} = \sum_i \sum_j x_i x_j b_{ij} \quad (45.20)$$

The most commonly used combining rules are

$$b_{ij} = \frac{1}{2}(b_{ii} + b_{jj}) \quad \text{and} \quad a_{ij} = \sqrt{a_{ii} a_{jj}}(1 - k_{ij}) \quad (45.21)$$

where  $a_{ii}$  and  $b_{ii}$  are pure component equation of state parameters and  $k_{ij}$  is an adjustable parameter. A more complete review of equations of state and their mixing and combining rules can be found in Sandler [1994].

Once the volumetric equation of state for a mixture has been specified, the fugacity of each species in the mixture can be computed using the rigorous thermodynamic relation

$$\begin{aligned} RT \ln \hat{\phi}_i(T, P, x_i) &= \int_0^P \left( \bar{v}_i - \frac{RT}{P} \right) dP \\ &= \int_V^\infty \left[ \left( \frac{\partial P}{\partial n_i} \right)_{T, V, n_{j \neq i}} - \frac{RT}{V} \right] dV - RT \ln Z \end{aligned} \quad (45.22)$$

Here,  $Z = Pv/RT$  is the compressibility factor,  $\bar{v}_i$  is partial molar volume of component as calculated from an equation of state,  $V$  is total volume, and  $n$  is the number of moles.

## Solutions in the Solid Phase

Various solid solutions are encountered in engineering applications. In some cases, pure solids can exist in more than one form (for example, sulfur in rhombic and monoclinic allotropic forms, or carbon in diamond and in graphite forms), with each phase exhibiting different physical characteristics. In such cases, equilibrium may exist between the pure solid phases. Solids may also form compounds, and solutions of these compounds are common in metallurgical applications. These phenomena are beyond the scope of this chapter and the reader is referred to other sources [Gaskell, 1981; Kyle, 1992; Sandler, 1989].

## 45.2 Applications

---

The fundamental principle of equilibrium between two phases in contact is that the temperature, pressure, and the chemical potentials of all species must be identical in both phases. This principle can be reduced to the following relation for vapor-liquid equilibrium [Sandler, 1989] of mixtures of condensible liquids when one of the activity coefficient models discussed earlier is used for the liquid phase:

$$x_i \gamma_i(T, P, x_i) \hat{f}_i^o(T, P, x_i) = y_i \hat{\phi}_i(T, P, y_i) P \quad (45.23)$$

To use this equation, an appropriate reference state for the liquid phase must be selected, which dictates the term  $\hat{f}_i^o$ , and the fugacity coefficient in the gas phase,  $\hat{\phi}_i$ , is evaluated from an equation of state. If the liquid phase is ideal ( $\gamma_i = 1$  and  $\hat{f}_i^o = P_i^{\text{vap}}$ , which is the pure component vapor pressure) and the gas phase is an ideal gas mixture, the well-known Raoult's law is obtained:

$$x_i P_i^{\text{vap}} = y_i P \quad \text{and} \quad \sum_i x_i P_i^{\text{vap}} = P \quad (45.24)$$

For liquid-liquid phase separation the equilibrium relation is

$$x_i^{\text{I}} \gamma_i(T, x_i^{\text{I}}) = x_i^{\text{II}} \gamma_i(T, x_i^{\text{II}}) \quad (45.25)$$

where I and II refer to the coexisting equilibrium liquid phases. This equation has to be solved for each species, with activity coefficients obtained from the models above.

In the case where both phases can be described by an equation of state, the fundamental equation of vapor-liquid equilibrium is

$$x_i \hat{\phi}_i^L(T, P, x_i) = y_i \hat{\phi}_i^V(T, P, y_i) \quad (45.26)$$

where the fugacity coefficient of each species in each phase is calculated from Eq. (45.22) and an equation of state. Note that a cubic equation of state such as Eq. (45.19) may have three roots at the selected pressure and at temperatures below the critical temperature. In this case, the smallest root is used for the evaluation of the liquid phase fugacity coefficients, and the largest root is used for the vapor phase fugacities.

Detailed discussions of the application of equations of state and activity coefficient models to phase equilibrium problems can be found in **Chapter 47** of this text and in thermodynamics textbooks [e.g., Kyle, 1992; Chapter 8 of Sandler, 1989]. Calculations of equilibria of solids with solids, liquids, and vapors are important for heterogeneous chemical reactions, precipitation from solutions, and in metallurgical applications. These cases are beyond the scope of this chapter and can be found elsewhere [Kyle, 1992; Gaskell, 1981; Sandler, 1989].

### Defining Terms

**Activity coefficient,  $\gamma_i$ :** A term that accounts for the nonideality of a liquid mixture; related to the

excess partial molar Gibbs free energy.

**Athermal solution:** A mixture in which the excess internal energy of mixing is zero.

**Chemical potential,  $\mu_i$ :** Equal to the partial molar Gibbs free energy.

**Equation of state:** An equation relating the volume, temperature, and pressure of a pure fluid or mixture.

**Excess property,  $\Theta^{EX}(T, P, x_i)$ :** The additional change in a thermodynamic property on forming a mixture from the pure components above that on forming an ideal mixture.

**Fugacity,  $f_i$ , and fugacity coefficient,  $\hat{\phi}_i$ :** Terms related to the partial molar Gibbs free energy or chemical potential that are computed from an equation of state.

**Mixing rule:** Equations to compute the parameters in a mixture equation of state from those for the pure components.

**Partial molar property,  $\bar{\Theta}_i(T, P, x_i)$ :** The contribution to a thermodynamic property of a mixture made by adding a small amount on component  $i$ , reported on a molar basis.

**Regular solution:** A mixture in which the excess entropy of mixing is zero.

## References

- Fredenslund, A. 1977. *Vapor-Liquid Equilibria Using UNIFAC: A Group Contribution Method*. Elsevier, New York.
- Gaskell, D. R. 1981. *Introduction to Metallurgical Thermodynamics*, 2nd ed. Hemisphere, New York.
- Kyle, B. G. 1992. *Chemical and Process Thermodynamics*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.
- Sandler, S. I. 1989. *Chemical and Engineering Thermodynamics*, 2nd ed. John Wiley & Sons, New York.
- Sandler, S. I., ed. 1994. *Models for Thermodynamic and Phase Equilibria Calculations*. Marcel Dekker, New York.
- van der Waals, J. H. 1890. Physical memoirs. (English translation by Therelfall & Adair.) *Physical Society*. 1, iii, 333.
- Van Ness, H. 1964. *Classical Thermodynamics of Nonelectrolyte Solutions*. Pergamon Press, London.

## Further Information

- Prausnitz, J. M., Lichtenthaler, R. N., and Azevedo, E. G. 1986. *Molecular Thermodynamics of Fluid-Phase Equilibria*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.
- Reid, R. C., Prausnitz, J. M., and Poling, B. E. 1987. *The Properties of Gases and Liquids*, 4th ed. McGraw-Hill, New York.
- Walas, S. M. 1985. *Phase Equilibria in Chemical Engineering*. Butterworths, Boston, MA.



Krantz, W. B. "Thermodynamics of Surfaces"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Thermodynamics of Surfaces

---

## 46.1 Basic Concepts

The Interface • Curvature • The Gibbs Phase Rule

## 46.2 First Law of Thermodynamics

## 46.3 Effects of Curved Interfaces

Young and Laplace Equation • The Kelvin Equation • The Gibbs-Thompson Equation

## 46.4 Adsorption at Interfaces

The Gibbs Adsorption Equation

## 46.5 Wettability and Adhesion

Wetting and Spreading • Adhesion and Cohesion • Contact Angles

### **William B. Krantz**

*University of Colorado*

In writing this brief overview it was necessary to focus on those aspects of the thermodynamics of surfaces that are of considerable importance in engineering applications and that could be summarized well in this compact format. We begin by considering the basic concepts of the interface as a surface, curvature, and the generalized Gibbs phase rule. The generalized first law of thermodynamics is then stated, which forms the basis for subsequent considerations of work done on or by interfaces. The implications of surface curvature on capillary pressure, vapor pressure and solubility, and phase transition temperature are reviewed. Adsorption at interfaces is then considered with a brief introduction to surface equations of state. The first law is then used to introduce the subjects of wettability, adhesion, and contact angles. Unfortunately, space does not permit reviewing adequately other important areas involving the thermodynamics of surfaces, such as electrical aspects of surface chemistry, nucleation and growth phenomena, colloids, emulsions, foams, aerosols, and thin films. An excellent overview of these and other topics in surface science is provided in Adamson [1990].

## 46.1 Basic Concepts

---

### **The Interface**

Two bulk phases in contact will be separated by a thin region within which the intensive thermodynamic properties change continuously from those of the one phase to those of the other. Since this region is typically only a few nanometers thick, we replace it with a surface whose thermodynamic properties are assigned to conserve the extensive thermodynamic properties of the

overall system composed of the interface and the two bulk phases, which now are assumed to maintain their bulk properties up to the dividing surface. There are several conventions used to locate the dividing plane, henceforth to be referred to as the *interface*; however, the interfacial intensive thermodynamic properties are independent of its location.

An interface can exist between a liquid and a gas (L/G interface), two liquid phases (L/L), a liquid and a solid (L/S), a gas and a solid (G/S), or two solid phases (S/S). Interfaces between fluid phases frequently can be assumed to be at equilibrium. However, interfaces formed with a solid phase often cannot be described by equilibrium thermodynamics, owing to slow diffusion in solids.

## Curvature

The interface between two fluids takes on solidlike properties in that it can sustain a state of stress characterized by the *surface tension*. Consequently, a pressure drop can be sustained across curved interfaces, which increases with their curvature. The latter is uniquely defined at a point on the interface by the curvatures in two mutually perpendicular directions. The curvature  $C$  of an interface in a given plane is defined by

$$C \equiv \vec{n} \cdot \frac{d\vec{t}}{ds} \quad (46.1)$$

where  $\vec{n}$  and  $\vec{t}$  are the normal and tangential unit vectors at a point and  $s$  is the arc length along the curved interface. For a curved surface described in rectangular Cartesian coordinates by  $y = y(x, z)$ , Eq. (46.1) implies the following for the curvatures in the  $yx$  and  $yz$  planes:

$$C_{xy} = \frac{y_{xx}}{(1 + y_x^2)^{3/2}} \quad (46.2)$$

$$C_{zy} = \frac{y_{zz}}{(1 + y_z^2)^{3/2}} \quad (46.3)$$

where  $y_c$  and  $y_{cc}$  denote the first and second partial derivatives of  $y$  with respect to  $c$  ( $c = x$  or  $z$ ). For axisymmetric bodies the two curvatures  $C_1$  and  $C_2$  can be represented solely in terms of  $y_x$  and  $y_{xx}$ :

$$C_1 = \frac{y_{xx}}{(1 + y_x^2)^{3/2}} \quad (46.4)$$

$$C_2 = \frac{y_x}{x(1 + y_x^2)^{1/2}} \quad (46.5)$$

Let us determine the curvature of a sphere of radius  $r$  defined by  $x^2 + y^2 + z^2 = r^2$ . Equations (46.4) and (46.5) then give  $C_1 = C_2 = -r^{-1}$ ; that is, the curvature is equal to the reciprocal of the radius of the sphere. The negative sign follows the convention that the curvature is viewed as if one is looking at the surface down the  $y$  axis. The curvature is positive when viewed from the

concave side (e.g., interior of the sphere) and negative when viewed from the convex side.

## The Gibbs Phase Rule

The Gibbs phase rule specifies the number of thermodynamic intensive variables  $f$  required to uniquely determine the thermodynamic state of the system. The phase rule is derived by adding up the intensive variables and then subtracting the number of independent relationships between them. For a system containing  $c$  components distributed between  $p$  phases separated by curved interfaces, the phase rule is given by

$$f = c + 1 \quad (46.6)$$

Equation (46.6) differs from the phase rule for bulk systems for which  $f = c - p + 2$  because the  $p$  phases can sustain different pressures, owing to the interfacial curvature. For example, this generalized phase rule implies that a droplet of pure liquid water in equilibrium with its own vapor requires two intensive thermodynamic variables to be specified in order to determine its boiling point. Specifying the liquid and gas phase pressures, which is equivalent to specifying the curvature of the drop, is then sufficient to determine a unique boiling point.

## 46.2 First Law of Thermodynamics

---

Consider two contacting multicomponent phases  $\alpha$  and  $\beta$  separated by a single interface having area  $A$ . The generalized first law of thermodynamics is then given by

$$\begin{aligned} dU &= T dS - P^\alpha dV^\alpha - P^\beta dV^\beta + \gamma dA + \sum_i \mu_i dN_i \\ &= dQ - dW + \sum_i \mu_i dN_i \end{aligned} \quad (46.7)$$

where  $U$ ,  $S$ , and  $N_i$  are the total system internal energy, entropy, and moles of species  $i$ , respectively;  $T$  is the absolute temperature;  $\mu_i$  is the chemical potential of  $i$ ;  $P^p$  and  $V^p$  are the pressure and total volume of phase  $p$  ( $p = \alpha$  or  $\beta$ );  $\gamma$  is the surface tension, an intensive thermodynamic property having units of energy per unit area ( $\text{J/m}^2$ ) or force per unit length ( $\text{N/m}$ );  $dQ = TdS$  is the heat transferred to the system; and  $dW = P^\alpha dV^\alpha + P^\beta dV^\beta - \gamma dA$  is the work done by the system. Hence, changing the interfacial area can result in work being done on or by the system.

## 46.3 Effects of Curved Interfaces

---

### Young and Laplace Equation

Consider the simplification of Eq. (46.7) for an isolated system at equilibrium (that is,

$dU = dS = dN_i = 0$  ). For an arbitrary displacement of the interface between  $\alpha$  and  $\beta$ ,  $dA = (C_1 + C_2)dV^\alpha$  and  $dV^\alpha = -dV^\beta$  ; hence,

$$\Delta P \equiv P^\alpha - P^\beta = \gamma(C_1 + C_2) \quad (46.8)$$

Equation (46.8), the Young and Laplace equation, predicts a pressure drop across a curved interface, which is referred to as a *capillary pressure effect*. For planar interfaces,  $\Delta P = 0$  , whereas for spherical liquid drops having a radius  $r$ ,  $\Delta P = 2\gamma/r$  . Note the pressure is higher on the concave (positive curvature) side of the interface.

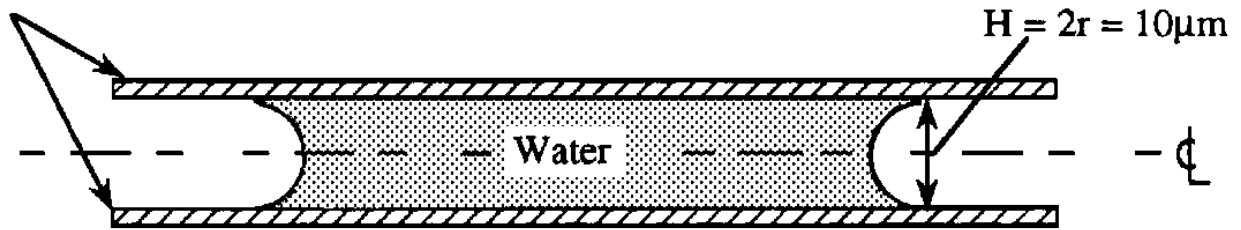
As an example of the use of Eq. (46.8) consider the force required to separate two glass plates between which  $V = 1$  ml of water is spread to a thickness of  $H = 10 \mu\text{m}$  , as shown in Fig. 46.1. Since  $C_2 \ll C_1$  , we have

$$\begin{aligned} F &= \Delta P A = \gamma C_1 \frac{V}{H} \\ &= \frac{2(72.94 \cdot 10^{-3} \text{ N/m})(1 \cdot 10^{-6} \text{ m}^3)}{(10 \cdot 10^{-6} \text{ m})^2} \quad (46.9) \\ &= 1459 \text{ N (328 lbf)} \end{aligned}$$

Capillary pressure effects can be quite large and account for phenomena such as the cohesiveness of wet soils.

**Figure 46.1** A wetting liquid between two parallel glass plates:  $C = r^{-1} = H/2$ .

Glass Plates



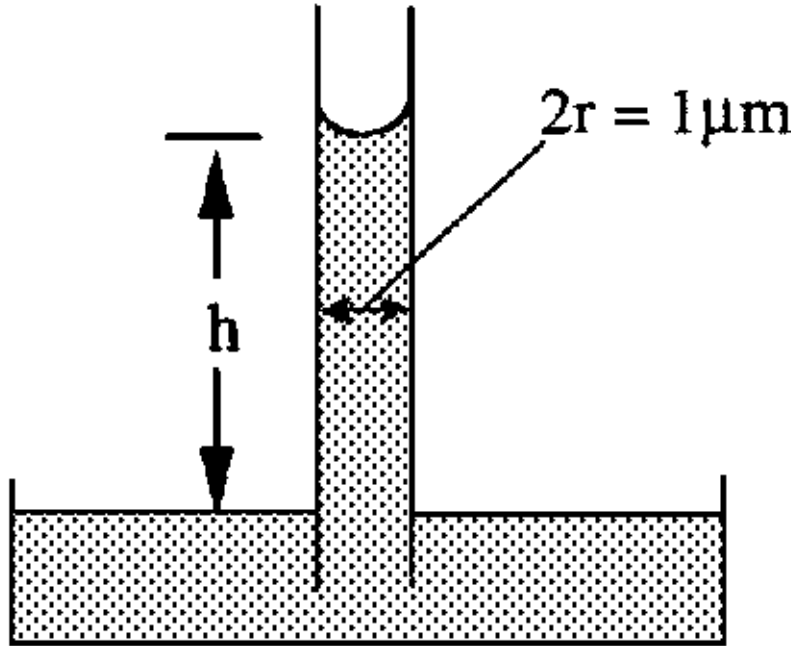
As a second example consider the placing of a capillary tube vertically into a wetting liquid (one having a convex interface when viewed through the liquid) as shown in Fig. 46.2; the liquid will rise until the pressure drop across the curved interface is equal to the hydrostatic pressure exerted by the liquid column of height  $h$ ; that is,

$$\Delta P = \frac{2\gamma}{r} = \Delta\rho gh \quad (46.10)$$

where  $\Delta\rho$  is the density difference between the liquid (phase  $\beta$ ) and the ambient fluid (phase  $\alpha$ ), and  $g$  is the gravitational acceleration. For example, a  $1 \mu\text{m}$  capillary can draw water to a height of 30 m (98 ft) in air. Capillary rise is involved in absorbent wicking and also is used as a method for

measuring L/G and L/L surface tensions.

**Figure 46.2** Capillary rise of a wetting liquid  $\beta$  in an ambient gas phase  $\alpha$ :  $C_1 = C_2 = r^{-1}$ .



As a third example consider the insertion of a capillary tube containing an insoluble gas vertically into a wetting liquid. If the volume of the gas is slowly increased, the curvature of the L/G interface will adjust to satisfy Eq. (46.8). The gas-phase pressure will pass through a maximum when the bubble radius is equal to the tube radius. This is the basis of the maximum bubble pressure method for determining surface tensions.

## The Kelvin Equation

The Young and Laplace equation also implies that liquids bounded by an interface having positive curvature will exert a higher vapor pressure  $P^v$  than that exerted by liquids bounded by a planar interface  $P_o^v$ ; these two vapor pressures are related by the Kelvin equation:

$$P^v = P_o^v \exp \left( \frac{2\gamma \bar{V}_l}{RT r} \right) \quad (46.11)$$

where  $\bar{V}_l$  is the molar volume of the liquid and  $R$  is the gas constant. A water droplet having a radius of  $0.01\ \mu\text{m}$  will have a  $P^v = 1.11P_o^v$ . The Kelvin equation can also be applied to S/L systems to predict the increased solubility of small crystals in solution. Equation (46.11) explains the formation of supersaturated vapors and Ostwald ripening whereby large crystals grow at the expense of small crystals. It also explains the falling rate period in drying during which water must be removed from progressively smaller pores, causing reduced vapor pressures.

## The Gibbs-Thompson Equation

The Young and Laplace equation also implies that wetting liquids in small capillaries will have a lower freezing temperature  $T$  than the normal freezing temperature  $T_o$ ; these two freezing temperatures are related by the Gibbs-Thompson equation:

$$T = T_o - \frac{\gamma T_o \bar{V}_s}{\Delta H^f} (C_1 + C_2) \quad (46.12)$$

where  $\bar{V}_s$  is the molar volume of the solid phase and  $\Delta H^f$  is the latent heat of fusion. Water in a capillary having a radius  $0.01 \mu\text{m}$  will freeze at  $-5.4^\circ\text{C}$ . Equation (46.12) explains why freezing in wet soils occurs over a zone rather than at a discrete plane.

## 46.4 Adsorption at Interfaces

---

### The Gibbs Adsorption Equation

The interfacial concentration  $\Gamma_i \equiv N_i^s/A$ —where  $N_i^s$  denotes the moles of  $i$  in the interface—depends on the location of the interface. It can be determined from the concentration dependence of the surface tension using the Gibbs adsorption equation, which is the interfacial analogue of the Gibbs-Duhem equation derived from the generalized first law and given by

$$d\gamma = - \sum_i \Gamma_i d\mu_i \quad (46.13)$$

Consider determining the interfacial concentration for the special case of a binary ideal solution for which the surface tension is a linear function of the bulk concentration of solute,  $c_2$ , given by  $\gamma = \gamma_o - bc_2$ . We will locate the interface to ensure no net adsorption of solvent 1; that is,  $\gamma_1 = 0$ . This is the commonly used Gibbs convention for locating the dividing surface. Hence, Eq. (46.13) implies that the interfacial concentration of solute in this convention,  $\Gamma_2^1$  (where the superscript 1 denotes the Gibbs convention), is given by

$$\Gamma_2^1 = -\frac{\partial\gamma}{\partial\mu_2} = -\frac{N_2}{RT} \frac{\partial\gamma}{\partial N_2} = \frac{bc_2}{RT} = \frac{\gamma_o - \gamma}{RT} = \frac{\pi}{RT} \quad (46.14)$$

where  $\pi \equiv \gamma_o - \gamma$  is referred to as the *surface pressure*, having units of force per unit length. Equation (46.14) implies that  $\pi\sigma_i = RT$ , where  $\sigma_i \equiv 1/\Gamma_i^1$  is the area per mole of solute in the interface. This is a two-dimensional analogue to the familiar ideal gas law. Hence, by analogy one would expect a linear dependence of surface tension on bulk concentration at low surface pressures where intermolecular forces and the size of the molecules can be ignored. The two-dimensional equations of state are used to describe the  $\pi - \sigma_i - T$  behavior of interfaces just as the three-dimensional analogues are used to describe the  $P - V - T$  behavior of bulk phases.

Note that  $b > 0 \Rightarrow \pi > 0 \Rightarrow \Gamma_2^1 > 0$  ; that is, surface tension decreasing with increasing solute concentration implies positive adsorption of the solute at the interface. Solutes displaying positive adsorption are said to be *surface-active* and are referred to as *surfactants*. Surfactants are used to alter the wettability of surfaces in applications such as the manufacture of cleaning detergents and tertiary oil recovery. If  $b < 0 \Rightarrow \pi < 0 \Rightarrow \Gamma_2^1 < 0$  , negative adsorption occurs; this is characteristic of ionizing salts in aqueous solution, which avoid the air/water interface, which has a lower dielectric constant.

If the surface pressure is eliminated from the equation of state using the prescribed equation for the surface tension dependence on bulk concentration, one obtains a relationship between the surface and bulk concentrations referred to as the *adsorption isotherm*. The Gibbs adsorption equation provides the link between the interfacial equation of state, the adsorption isotherm, and the surface tension dependence on bulk concentration.

## 46.5 Wettability and Adhesion

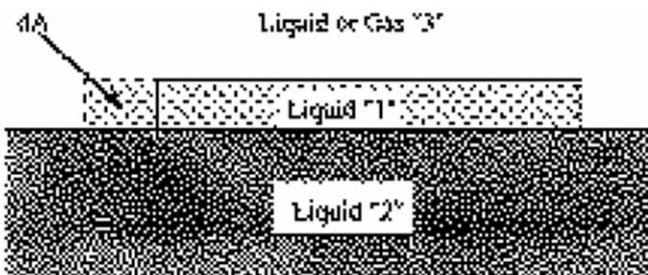
### Wetting and Spreading

Wetting relates to the propensity of a liquid to spread over another liquid or a solid. Consider the spreading of liquid 1 on liquid 2 in the presence of gas or liquid 3, as shown in Fig. 46.3, to create  $dA$  of interface between 1 and 2 and between 1 and 3 while destroying  $dA$  of interface between 2 and 3. Equation (46.7) implies that the reversible work of spreading  $dW_s/dA$  at constant  $T$  and  $P$  is given by

$$\frac{dW}{dA} = -\gamma_{12} - \gamma_{13} + \gamma_{23} \equiv S_{12} \quad (46.15)$$

where  $\gamma_{ij}$  denotes the surface tension between phases  $i$  and  $j$  and  $S_{12}$  is defined as the *spreading coefficient* of phase 1 on phase 2. Note that if  $dW/dA = S_{12} > 0$  , spreading will occur spontaneously since work can be done by the system to increase the contact area between phases 1 and 2. Consider the spreading of heptane (1) on water (2) in air (3) at 20°C , for which  $\gamma_{12} = 50.2 \text{ mN/m}$  ,  $\gamma_{13} = 20.14 \text{ mN/m}$  , and  $\gamma_{23} = 72.94 \text{ mN/m}$  . Hence,  $dW/dA = S_{12} = 2.6 \text{ mN/m}$  and we conclude that spreading will occur spontaneously. Wetting is an important consideration in developing coatings of various types, solders and welding fluxes, water-repellent fabrics, and so on.

**Figure 46.3** Spreading of liquid 1 on liquid 2 in the presence of gas or liquid 3.





## Adhesion and Cohesion

Adhesion relates to the cohesiveness of the "bond" between two contacting phases. A measure of this property is the work of adhesion  $W_{ad}$ , the work that must be done on the system to separate the two phases in contact. Consider the separation of phase 1 from phase 2 in the presence of phase 3, as shown in Fig. 46.4, to create  $dA$  of interface between 1 and 3 and between 2 and 3 while destroying  $dA$  of interface between 1 and 2. Equation (46.7) implies that the reversible work of adhesion  $W_{ad}$  at constant  $T$  and  $P$  is given by

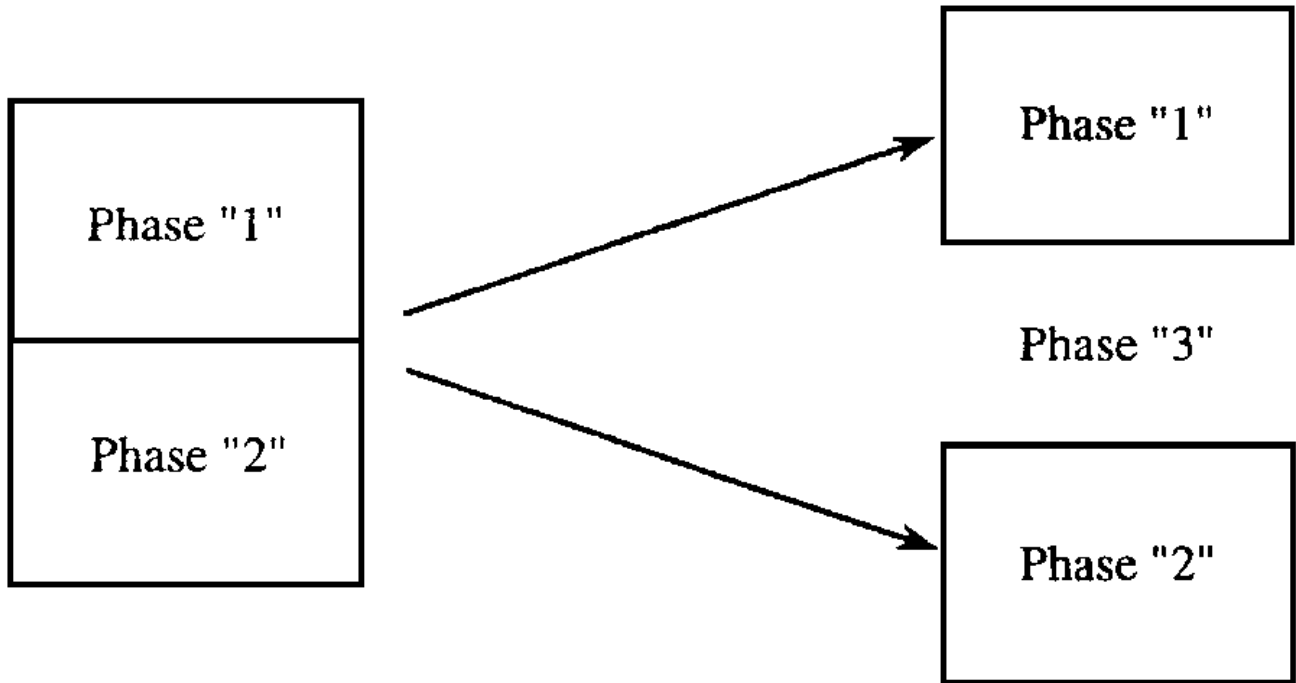
$$W_{ad} = -\frac{dW}{dA} = \gamma_{13} + \gamma_{23} - \gamma_{12} \quad (46.16)$$

A special case of the above is the work of cohesion  $W_{co}$ , which is the work required to separate a single phase 1 in the presence of phase 3, thereby creating two interfaces between phases 1 and 3. Equation (46.16) implies that

$$W_{co} = -\frac{dW}{dA} = \gamma_{13} + \gamma_{13} = 2\gamma_{13} \quad (46.17)$$

Comparison between Eqs. (46.15), (46.16), and (46.17) indicates that  $S_{12} = W_{ad} - W_{co}$ ; that is, the spreading coefficient is equal to the difference between the work of adhesion and cohesion. Adhesion is involved in developing bonding materials, laminates, paints, printing inks, and so on.

**Figure 46.4** Work of adhesion necessary to separate phase 1 from phase 2 in the presence of phase 3.



## Contact Angles

If phase 1 completely "wets" phase 2 in the presence of phase 3, it will spread over phase 2. If phase 1 does not wet phase 2, it will form a drop having a contact angle  $\theta$ , where  $\theta$  is measured from the interface between phases 1 and 2 through phase 1 to the tangent line at the contact line between phases 1, 2, and 3, as shown in Fig. 46.5. The contact angle is an intensive thermodynamic property of the system uniquely determined by the surface tensions via Young's equation, given by

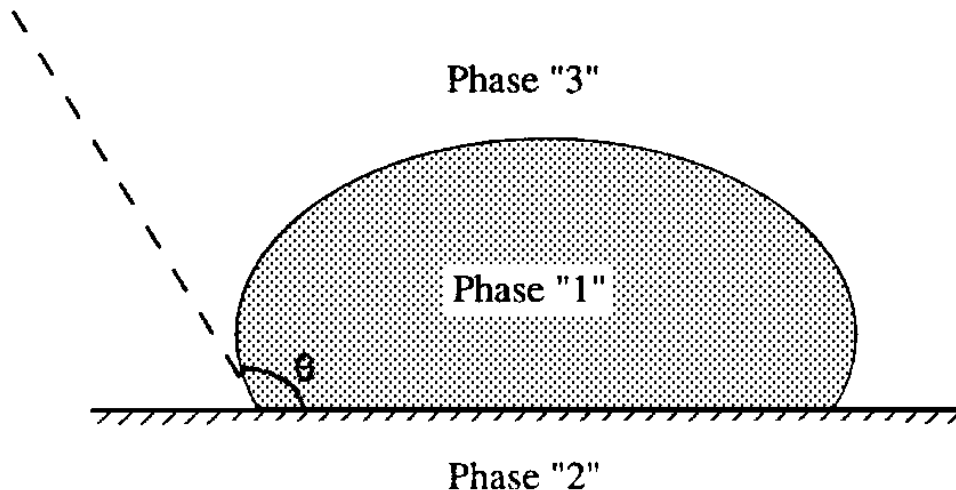
$$\gamma_{12} - \gamma_{23} + \gamma_{13} \cos \theta = 0 \quad (46.18)$$

If  $\theta = 0$ , phase 1 is said to *wet* phase 2; if  $\theta > 90^\circ$ , phase 1 is said to be *nonwetting* to phase 2 in the presence of phase 3. Equation (46.16) can be combined with Eq. (46.18) to obtain

$$W_{ad} = \gamma_{13}(1 + \cos \theta) \quad (46.19)$$

Equation (46.19) implies that the work of adhesion of phase 1 to phase 2 can be obtained by merely measuring the interfacial tension  $\gamma_{13}$  and the contact angle of phase 1 on phase 2 in the presence of phase 3. This provides a simple means for determining the strength of adhesion between two materials.

**Figure 46.5** A drop of phase 1 resting on phase 2 in the presence of phase 3, showing the contact angle  $\theta$ .



## Defining Terms

**Capillary pressure effects:** A pressure difference sustained across curved interfaces owing to surface tension.

**Contact angle:** The angle measured through a liquid phase to the tangent at the contact line where three phases meet.

**Interface:** A dividing surface between two phases to which intensive thermodynamic properties are assigned that satisfy the conservation of extensive thermodynamic properties of the overall system consisting of the interface and the two bulk phases whose intensive properties are assumed to persist up to the dividing plane.

**Surface tension:** An intensive property of an interface that characterizes the distinct energy state whereby it sustains a tensile stress at equilibrium.

**Surfactants:** Solutes that have a propensity to positively adsorb at the interface.

## References

- Adamson, A. W. 1990. *Physical Chemistry of Surfaces*, 5th ed. John Wiley & Sons, New York.
- Aveyard, R. and Haydon, D. A. 1973. *An Introduction to the Principles of Surface Chemistry*. Cambridge University Press, Cambridge.
- Buff, F. P. 1960. The theory of capillarity. In *Handbuch der Physik*, vol. X, pp. 281–304. Springer-Verlag, Berlin.
- Davies, J. T. and Rideal, E. K. 1963. *Interfacial Phenomena*, 2nd ed. Academic, New York.
- Edwards, D. A., Brenner, H., and Wasan, D. T. 1991. *Interfacial Transport Processes and Rheology*. Butterworth-Heinemann, Boston.
- Gibbs, J. W. 1961. *Scientific Papers*, vol. 1. Dover, New York.
- Hiemenz, P. C. 1986. *Principles of Colloid and Surface Chemistry*, 2nd ed. Marcel Dekker, New York.
- Miller, C. A. and Neogi, P. 1985. *Interfacial Phenomena*. Surfactant Science Series, vol. 17. Marcel Dekker, New York.
- Morrison, S. R. 1977. *The Chemical Physics of Surfaces*. Plenum, London.
- Rosen, M. J. 1989. *Surfactants and Interfacial Phenomena*, 2nd ed. John Wiley & Sons, New York.

## Further Information

Interested readers are referred to the following journals that publish research on interfacial phenomena:

*Journal of Adhesion*  
*Journal of Colloid and Interface Science*  
*Colloids and Surfaces*  
*Langmuir*

Kyle, B. G. "Phase Equilibrium"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Phase Equilibrium

---

## 47.1 Pure-Component Phase Equilibrium

The Clapeyron Equation • The Clausius-Clapeyron Equation

## 47.2 Phase Equilibrium in Mixtures

Vapor-Liquid Equilibrium (VLE) • Liquid-Liquid Equilibrium (LLE) • Solid Liquid Equilibrium (SLE)

## 47.3 Perspective

### Benjamin G. Kyle

*Kansas State University*

The best way to understand the rationale underlying the application of thermodynamics to phase equilibrium is to regard thermodynamics as a method or framework for processing experimentally gained information. From known properties for the system under study (information), other properties may be determined through the network of thermodynamic relationships. Thus, information about a system can be extended. The thermodynamic framework is also useful for conceptualizing and analyzing phase equilibrium. This allows extrapolation and interpolation of data and the establishment of correlations as well as evaluation and application of theories based on molecular considerations.

For the most part, the variables of interest in phase equilibrium are intensive: temperature, pressure, composition, and specific or molar properties. The phase rule determines the number of intensive variables,  $F$ , required to define a system containing  $\pi$  phases and  $C$  components.

$$F = C + 2 - \pi \quad (47.1)$$

The thermodynamic network also contains other intensive variables such as **fugacity** and the activity coefficient, which are necessary for computational purposes but possess no intrinsic value; they are simply useful thermodynamic artifacts. Specifically, the basis for the thermodynamic treatment of phase equilibrium is the equality of the fugacity of each component in each phase in which it exists.

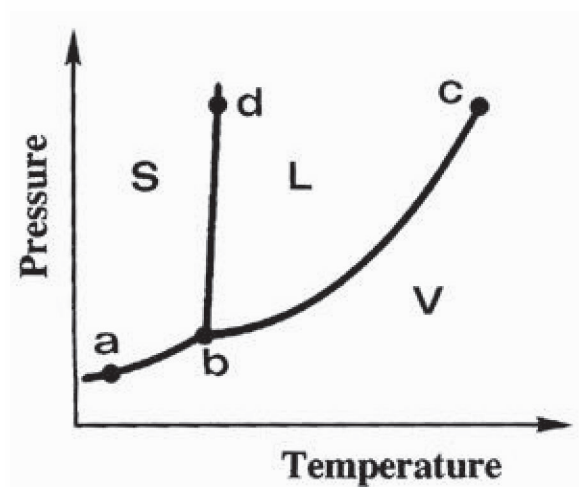
## 47.1 Pure-Component Phase Equilibrium

---

Equilibrium between phases of a pure component can be visualized with the aid of [Fig. 47.1](#), which shows areas in the pressure-temperature ( $PT$ ) plane labeled S, L, and V and representing conditions where respectively solid, liquid, and vapor exist. In accordance with the phase rule, specification of two intensive variables (here,  $T$  and  $P$ ) is required to define the state of the system.

Conditions where two phases are in equilibrium are represented on this diagram by lines (such as  $ab$ ,  $bc$ , and  $bd$ ) and therefore only one variable need be specified. The intersection of three lines results in a triple point, where three phases are in equilibrium and no variables may be specified. Some systems exhibit more than one triple point.

**Figure 47.1** Pressure-temperature graph of pure-component phase equilibrium.



## The Clapeyron Equation

For two phases in equilibrium, say the  $\alpha$  and  $\beta$  phases, the equating of fugacities leads to the Clapeyron equation:

$$\frac{dP}{dT} = \frac{h^\alpha - h^\beta}{(v^\alpha - v^\beta)T} \quad (47.2)$$

This equation relates the slope of a coexistence line on [Fig. 47.1](#) (e.g.,  $ab$ ) to differences in molar enthalpy and molar volume and can be used to determine the value of one variable when all the others are known.

## The Clausius-Clapeyron Equation

When one of the phases is a vapor, the number of variables in the Clapeyron equation can be reduced; the result is the Clausius-Clapeyron equation:

$$\frac{d \ln P^0}{d(1/T)} = -\frac{\Delta h}{R} \quad (47.3)$$

If Eq. (47.3) is integrated under the assumption of constant  $\Delta h$ , the result is

$$\ln P^0 = c - \frac{\Delta h}{RT} \quad (47.4)$$

where  $c$  is a constant of integration. Equations (47.3) and (47.4) relate the vapor pressure,  $P^0$ , to temperature and the heat of vaporization or heat of sublimation,  $\Delta h$ , and predict that a plot of  $\ln P^0$  versus reciprocal of absolute temperature should be linear with a slope of  $-\Delta h/R$ . Equations (47.3) and (47.4) find use in correlating, interpolating, and extrapolating vapor pressure data and in determining the heat of vaporization or sublimation from vapor pressure data.

Despite some questionable assumptions made in arriving at Eqs. (47.3) and (47.4), the linear relationship has been found to hold over a wide range of temperature. In fact, the Antoine equation,

$$\log P^0 = A - \frac{B}{C + t} \quad (47.5)$$

which has been found to give excellent representation of vapor pressure data, can be seen to be an empiricized version of Eq. (47.4). Extensive compilations of Antoine parameters ( $A$ ,  $B$ , and  $C$ ) are available [Boublik *et al.*, 1973; Reid *et al.*, 1987].

## 47.2 Phase Equilibrium in Mixtures

---

There are two approaches to the treatment of phase equilibrium in mixtures—the use of an equation of state and the use of activity coefficients. The latter is developed here; the former is delineated by Kyle [1992] and also in **Chapter 45** of this text.

### Vapor-Liquid Equilibrium (VLE)

Although the activity coefficient approach can be used for systems at any pressure, it is predominantly applied at low to moderate pressure. Here, consideration will be restricted to low pressure. This method is based on the ideal solution model for representing the fugacity of a component in a solution. Unlike the pure-component fugacity, which can be easily calculated or estimated, it is usually not possible to directly calculate the fugacity of a component in a solution, and a model—the ideal solution model—is therefore employed. This model is applied to both the vapor and liquid phases. It has been found to represent the vapor phase reasonably well and will be used here without correction. On the other hand, the model fits very few liquid-phase systems and needs to be corrected. The activity coefficient,  $\gamma_i$ , is the correction factor and is expected to depend upon temperature, pressure, and the liquid-phase composition. The effect of pressure is usually neglected and activity-coefficient equations of demonstrated efficacy are used to represent the temperature and composition dependence [Reid *et al.*, 1987].

Employment of the equal fugacity criterion along with these models yields the basic equation for treating vapor-liquid equilibrium,

$$Py_i = P_i^0 x_i \gamma_i \quad (47.6)$$

where  $P$  is the system pressure,  $P_i^0$  is the vapor pressure of component  $i$ ,  $x_i$  is its liquid mole fraction, and  $y_i$  is its vapor mole fraction.

### Determination of Activity Coefficients

Although methods for estimating activity coefficients are available (see **Chapter 45**), it is preferable to use experimental phase equilibrium data for their evaluation. The exception to this rule is when it is known that a system forms an ideal liquid solution and  $\gamma$  values are therefore unity. These systems, however, are rare and occur only when all the constituents belong to the same chemical family (e.g., paraffin hydrocarbons). Equation (47.6) is used to calculate  $\gamma_i$  from experimental values of  $x_i$ ,  $y_i$ , and  $P$  at a given temperature where  $P_i^0$  is known. Thus, for a binary system,  $\gamma$  for each component can be evaluated from a single VLE data point.

Once a set of  $\gamma_1$ ,  $\gamma_2$ , and  $x_1$  data has been determined from binary VLE data, it is customary to fit this data to an activity-coefficient equation. The Wilson equation, shown below for a binary system, is outstanding among equations of proven efficacy:

$$\ln \gamma_1 = -\ln(x_1 + x_2 G_{12}) + x_2 \left( \frac{G_{12}}{x_1 + x_2 G_{12}} - \frac{G_{21}}{x_2 + x_1 G_{21}} \right) \quad (47.7)$$

$$\ln \gamma_2 = -\ln(x_2 + x_1 G_{21}) - x_1 \left( \frac{G_{12}}{x_1 + x_2 G_{12}} - \frac{G_{21}}{x_2 + x_1 G_{21}} \right) \quad (47.8)$$

The  $G$  values are empirically determined parameters that are expected to show the following exponential dependence on temperature:

$$G_{12} = \frac{v_2}{v_1} \exp \left( -\frac{a_{12}}{RT} \right); \quad G_{21} = \frac{v_1}{v_2} \exp \left( -\frac{a_{21}}{RT} \right) \quad (47.9)$$

where  $v_1$  and  $v_2$  are the liquid molar volumes of components 1 and 2, and  $a_{12}$  and  $a_{21}$  are empirically determined parameters. Some type of parameter estimation technique is used to determine the best values of  $G$ . This can easily be done on a spreadsheet using the optimization feature to find the parameters that minimize the following objective function:

$$\sum_{i=1}^n (Q^{\text{exp}} - Q^{\text{cal}})_i^2 \quad (47.10)$$

where  $Q$  is defined as

$$Q = x_1 \ln \gamma_1 + x_2 \ln \gamma_2 \quad (47.11)$$

In Eq. (47.10),  $Q^{\text{exp}}$  is determined from Eq. (47.11) and values of  $\gamma_1$  and  $\gamma_2$  from experimental VLE data, while  $Q^{\text{cal}}$  is determined at the same composition from Eq. (47.11) using Eqs. (47.7) and (47.8). The difference in these  $Q$  values and, consequently, the objective function [Eq. (47.10)]



is a function only of the  $G$  values.

Sensing a certain amount of circularity, one may well question the value of this procedure that uses experimental VLE data to determine parameters in the Wilson equation that will then be used to calculate VLE by means of Eq. (47.6). However, the procedure becomes more attractive when one recognizes that VLE data at one condition can be used to evaluate Wilson parameters that can then be used to calculate VLE at another condition. Or one might use the minimum amount of VLE data (one data point) to evaluate Wilson parameters and then proceed to calculate, via Eq. (47.6), sufficient data to represent the system. When no experimental information is available, activity coefficients may be estimated by the UNIFAC method, which is based on the behavior of functional groups within the molecules composing the solution rather than on the molecules themselves [Reid *et al.*, 1987].

### Azeotropes

An azeotrope is the condition in which the vapor and liquid compositions in a system are identical. As the knowledge of the existence of azeotropes is crucial to the design of distillation columns, much effort has been expended in determining and compiling this information [Horsely, 1952, 1962]. These compilations represent a source of VLE data: values of  $T$ ,  $P$ , and  $x_1$ , where  $x_i = y_i$ . At a specified  $T$ (or  $P$ ) a value of  $\gamma$  can be calculated for each component via Eq. (47.6), the minimum data required to determine Wilson parameters.

For an azeotrope in the 1-2 system, Eq. (47.6) can be employed to obtain

$$\frac{P_1^0}{P_2^0} = \frac{\gamma_2}{\gamma_1} \quad (47.12)$$

This equation is useful in estimating the effect of temperature (which is manifested as a corresponding change in pressure) on the composition of the azeotrope. The vapor pressure ratio depends only on temperature, whereas the  $\gamma$  values depend on temperature and composition. If one assumes that the ratio  $\gamma_2/\gamma_1$  is independent of temperature and uses the azeotropic data point to evaluate Wilson parameters, the right-hand side becomes a known function of  $x_1$ , and, thus, a value of  $x_1$  can be determined from the vapor pressure ratio evaluated at any temperature.

### Computational Procedures

The computations for VLE involving Eq. (47.6) cannot be directly executed because all of the necessary variables cannot be specified prior to the calculation. According to the phase rule, a system of two phases possesses  $C$  degrees of freedom. If the composition of a phase is specified at the expense of  $C - 1$  variables, there remains only one additional variable to be specified— $T$  or  $P$ . Because the quantities in Eq. (47.6) depend on  $T$  and  $P$ , it is necessary to adopt a trial-and-error computational procedure. If, for example,  $T$  and the  $x_i$  values are specified, a value of  $P$  would be assumed and Eq. (47.6) used to calculate the  $y_i$  values. If these calculated  $y_i$  values all sum to unity, the assumed value of  $P$  would be correct: if not, other values of  $P$  would be assumed until this condition is met. A similar procedure would be used for a system described in different terms.

## Multicomponent Vapor-Liquid Equilibrium

In order to effectively treat multicomponent systems, it is necessary to be able to represent activity coefficients as a function of temperature and composition for use in Eq. (47.6). The Wilson equation can be extended to any number of components and requires only two parameters ( $G_{ij}$  and  $G_{ji}$ ) for each binary pair that can be formed from the components of the system [Kyle, 1992]. A comprehensive study [Holmes and Van Winkle, 1970] has established the ability of the Wilson equation to calculate ternary VLE with acceptable accuracy from parameters obtained from the three constituent binary systems.

## Liquid-Liquid Equilibrium (LLE)

When two liquid phases are in equilibrium, the equating of fugacities results in the following expression for each component:

$$\gamma'_i x'_i = \gamma''_i x''_i \quad (47.13)$$

where ' and '' distinguish the liquid phases. Because activity coefficients are required for each phase, calculations involving LLE are computationally intensive. These calculations entail either the computation of LLE from available information concerning activity coefficients or the use of LLE to obtain information concerning activity coefficients [Kyle, 1992].

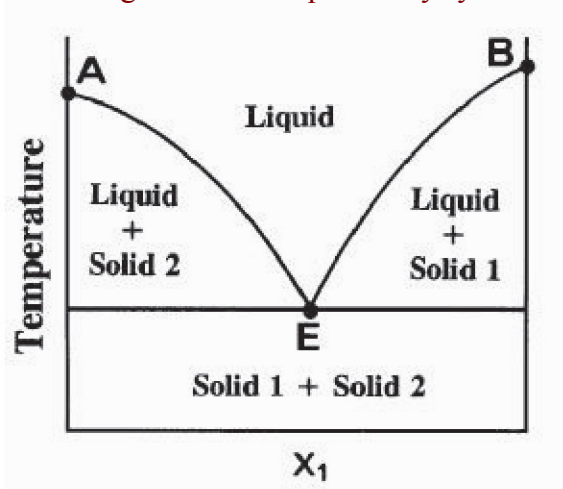
## Solid-Liquid Equilibrium (SLE)

For pure solid component  $i$  in equilibrium with a liquid solution containing component  $i$ , the equating of fugacities results in

$$RT \ln x_i \gamma_i = \frac{L_{mi}(T - T_{mi})}{T_{mi}} + \Delta c_{Pi}(T_{mi} - T) + T \Delta c_{Pi} \ln \frac{T}{T_{mi}} \quad (47.14)$$

where  $T_{mi}$  is the freezing point and  $L_{mi}$  is the heat of fusion of component  $i$ . Usually, the data for the evaluation of  $\Delta c_{Pi}$  are unavailable and only the first right-hand term is used. Figure 47.2 shows a phase diagram for a simple binary system where the curves  $AE$  and  $BE$  are represented by Eq. (47.14) with  $i$  equal to 2 and 1, respectively. The **eutectic point**  $E$  is merely the intersection of these two curves.

**Figure 47.2** Phase diagram for a simple binary system.



To apply Eq. (47.14) where liquid solutions are not ideal, it will be necessary to have access to enough experimental phase equilibrium data to evaluate parameters in the Wilson, or equivalent, activity-coefficient equation. When reliable parameters are available, it has been demonstrated [Gmehling *et al.*, 1978] that good estimates of SLE can be obtained for systems of nonelectrolytes. Alternatively, SLE data can be used as a source of activity coefficient information for use in other types of phase equilibrium calculations involving a liquid phase.

The solid phase can be either pure or a solid solution. For a given system, solid solutions will exhibit larger deviations from the ideal solution model than will liquid solutions, and, therefore, an activity coefficient will also be required for the solid phase. Thus, calculations for this type of system will require some experimental data for the evaluation of liquid- and solid-phase activity coefficients [Kyle, 1992].

## 47.3 Perspective

In this work the emphasis has been placed on delineating the rationale for the application of thermodynamics to phase equilibrium rather than on amassing a set of "working equations." Such a set would be useful in dealing with routine or standard problems; however, with the exception of pure components, there are few such problems in phase equilibrium. Each problem is defined by the nature of the system, what information is desired, and what information is available; judgments concerning the suitability of data and the fitting of data may also be involved. This endeavor can be regarded as information processing in which the activity coefficient provides the means by which information (experimental phase equilibrium data) at one condition is processed in order to generate information at other conditions.

### Defining Terms

**Eutectic point:** The existence of two solid phases and a liquid phase in equilibrium at a temperature below the melting points of the pure solid components.

**Fugacity:** A thermodynamic function that requires for its evaluation PVT data for a pure

substance and PVT-composition data for a component in a mixture. It is often regarded as a thermodynamic pressure because it has the units of pressure and is equal to pressure for an ideal gas and equal to partial pressure for a component in an ideal gas mixture.

## References

- Boublik, T., Fried, V., and Hala, E. 1973. *The Vapor Pressures of Pure Substances*. Elsevier, Amsterdam.
- Gmehling, J. G., Anderson, T. F., and Prausnitz, J. M. 1978. Solid-liquid equilibria using UNIFAC. *Ind. Eng. Chem. Fundam.* 17(4):269–273.
- Holmes, M. J. and Van Winkle, M. 1970. Prediction of ternary vapor-liquid equilibria from binary data. *Ind. Eng. Chem.* 62(1):21–31.
- Horsely, L. H. 1952. *Azeotropic Data*. Advances in Chemistry Series, No. 6. American Chemical Society, Washington, D.C.
- Horsely, L. H. 1962. *Azeotropic Data—II*. Advances in Chemistry Series, No. 35. American Chemical Society, Washington, D.C.
- Kyle, B. G. 1992. *Chemical and Process Thermodynamics*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.
- Reid, R. C., Prausnitz, J. M., and Poling, B. E. 1987. *The Properties of Gases and Liquids*, 4th ed. McGraw-Hill, New York.

## Further Information

A comprehensive treatment of phase equilibrium consistent with this work and a list of sources of phase equilibrium data is available in Kyle [1992].

A vast quantity of various types of phase equilibrium data is available in the Dortmund Data Bank via the on-line service DETHERM provided by STN International.

See Reid *et al.* [1987] for a detailed explication of the UNIFAC method for estimating activity coefficients and for comprehensive lists of sources of experimental data and books about phase equilibrium.

Cook, W. J. "Thermodynamic Cycles"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Thermodynamic Cycles

## 48.1 Power Cycles

## 48.2 Refrigeration Cycles

**William J. Cook**

*Iowa State University*

A thermodynamic cycle is a continuous series of thermodynamic processes that periodically returns the **working fluid** of the cycle to a given state. Although cycles can be executed in closed systems, the focus here is on the cycles most frequently encountered in practice: **steady-flow** cycles, cycles in which each process occurs in a steady-flow manner. Practical cycles can be classified into two groups: power-producing cycles (power cycles) and power-consuming cycles (refrigeration cycles). The working fluid typically undergoes phase changes during either a power cycle or a refrigeration cycle. Devices that operate on thermodynamic cycles are widely used in energy conversion and utilization processes since such devices operate continuously as the working fluid undergoes repeated thermodynamic cycles.

The fundamentals of cycle analysis begin with the **first law of thermodynamics**. Since each process is a steady-flow process, only the first law as it applies to steady-flow processes will be considered. For a steady-flow process occurring in a **control volume** with multiple inflows and outflows, the first law is written on a time rate basis as

$$\dot{Q} + \sum [\dot{m}(h + V^2/2 + gz)]_{\text{in}} = \dot{W} + \sum [\dot{m}(h + V^2/2 + gz)]_{\text{out}} \quad (48.1)$$

For a single-stream process between states  $i$  and  $j$ , Eq. (48.1) on a unit mass basis becomes

$${}_iq_j + h_i + V_i^2/2 + gz_i = {}_iw_j + h_j + V_j^2/2 + gz_j \quad (48.2)$$

where  ${}_iq_j = \dot{Q}_{ij}/\dot{m}$ ,  ${}_iw_j = \dot{W}_{ij}/\dot{m}$ , and  $\dot{m}$  is the mass rate of flow. See Van Wylen *et al.* [1994]. In processes involved with the cycles considered here, changes in kinetic and potential energies ( $V^2/2$  and  $gz$  terms, respectively) are small and are neglected. Power  $\dot{W}$  is considered positive when it is transferred out of the control volume, and heat transfer rate  $\dot{Q}$  is considered positive when heat transfer is to the control volume. In the figures herein that describe the transfer of power and heat energy to and from cycles, arrows indicate direction and take the place of signs. The accompanying  $\dot{W}$  or  $\dot{Q}$  is then an absolute quantity. Where confusion might arise, absolute value signs are used. Only power transfer and heat transfer occur across a closed boundary that encloses the complete cycle. For such a boundary,

$$\left[ \sum \dot{Q} = \sum \dot{W} \right]_{\text{cycle}} \quad (48.3)$$

## 48.1 Power Cycles

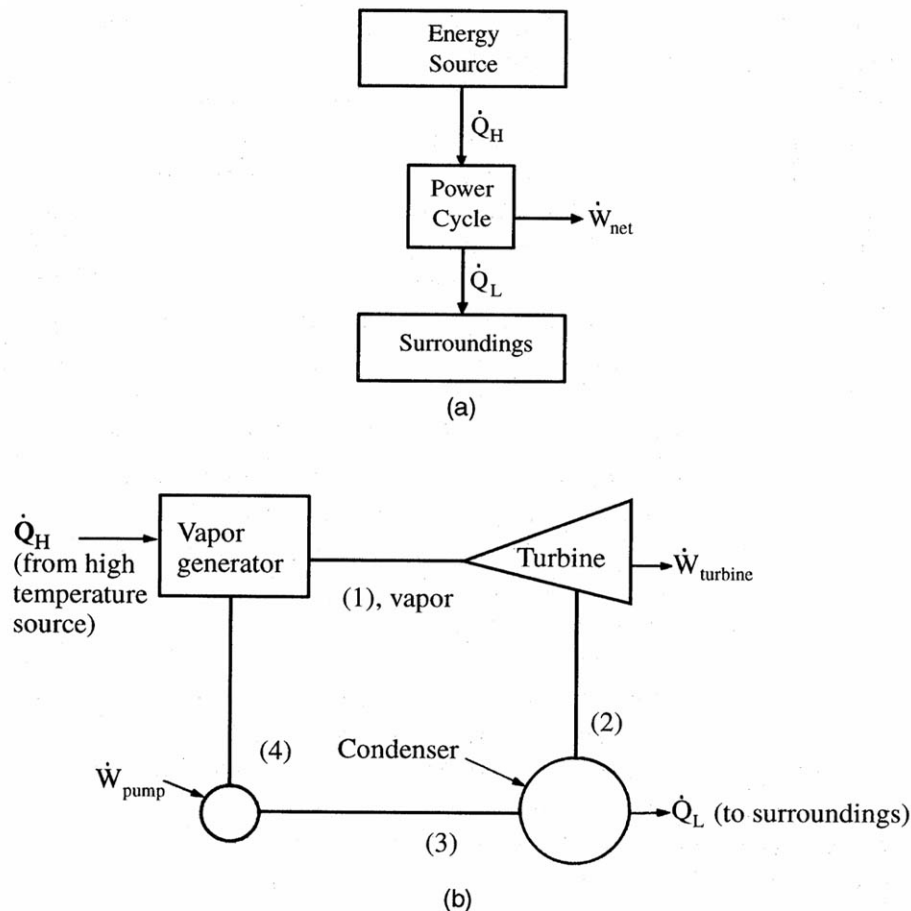
The purpose of a power cycle is to produce a net power output on a continuous basis from **heat energy** supplied to it from a high-temperature energy source. The device in which the power cycle is executed is sometimes referred to as a *heat engine*. Gas-turbine engines and reciprocating internal combustion engines are used widely to produce power. Strictly speaking, these engines are not classified as power cycles because their working fluids do not undergo thermodynamic cycles. [Figure 48.1\(a\)](#) shows a heat engine that receives heat energy at the rate  $\dot{Q}_H$  from a high-temperature energy source and produces net power  $\dot{W}_{\text{net}}$ . As a consequence of its operation it rejects heat energy to the lower-temperature surroundings at the rate  $\dot{Q}_L$ . A widely used performance parameter for a power cycle is  $\eta$ , the cycle thermal efficiency, defined as

$$\eta = \dot{W}_{\text{net}} / \dot{Q}_H \quad (48.4)$$

Second law considerations for thermodynamic power cycles restrict  $\eta$  to a value less than unity. Thus,  $\dot{W}_{\text{net}}$  in [Fig. 48.1\(a\)](#) is less than  $\dot{Q}_H$ . By Eq. (48.3), the rate at which heat energy is rejected to the surroundings is

$$|\dot{Q}_L| = \dot{Q}_H - \dot{W}_{\text{net}} \quad (48.5)$$

**Figure 48.1** Descriptions of power cycles: (a) Power cycle operation. (b) The simple vapor power cycle.



It is useful to consider cycles for which the energy source and the surrounding temperatures—denoted respectively as  $T_H$  and  $T_L$ —are uniform. The maximum thermal efficiency any power cycle can have while operating between a source and its surroundings, each at a uniform temperature, is that for a totally reversible thermodynamic cycle (a Carnot cycle, for example) and is given by the expression

$$\eta_{\max} = (T_H - T_L)/T_H \quad (48.6)$$

where the temperatures are on an absolute scale [Wark, 1983].

Figure 48.1(b) illustrates a simple vapor power cycle. Each component operates in a steady-flow manner. The vapor generator delivers high-pressure high-temperature vapor at state 1 to the turbine. The vapor flows through the turbine to the turbine exit state, state 2, and produces power  $\dot{W}_{\text{turbine}}$  at the turbine output shaft. The vapor is condensed to liquid, state 3, as it passes through the condenser, which is typically cooled by a water supply at a temperature near that of the surroundings. The pump, which consumes power  $\dot{W}_{\text{pump}}$ , compresses the liquid from state 3 to state 4, the state at which it enters the vapor generator. Heat energy at the rate  $\dot{Q}_H$  is supplied to the vapor generator from the energy source to produce vapor at state 1. Thus, the working fluid executes a cycle, in that an element of the working fluid initially at state 1 is periodically returned to that state through the series of thermodynamic processes as it flows through the various hardware components. The net power produced,  $\dot{W}_{\text{net}}$ , is the algebraic sum of the positive turbine power  $\dot{W}_{\text{turbine}}$  and the negative pump power  $\dot{W}_{\text{pump}}$ .

**Example 48.1.** Consider the power cycle shown in Fig. 48.1(b) and let water be the working fluid. The mass flow rate  $\dot{m}$  through each component is 100 kg/h, the turbine inlet pressure  $P_1$  is 1000 kPa, and turbine inlet temperature  $T_1$  is 480°C. The condenser pressure is 7 kPa and saturated liquid leaves the condenser. The processes through the turbine and the pump are isentropic (adiabatic and reversible, constant entropy). The pressure drop in the flow direction is assumed to be negligible in both the steam generator and the condenser as well as in the connecting lines. Compute  $\dot{W}_{\text{net}}$ ,  $\dot{Q}_H$ ,  $\eta$ , and  $\dot{Q}_L$ .

**Solution.** Table 48.1 lists the properties at each state and Fig. 48.2 shows the temperature ( $T$ ) versus entropy ( $s$ ) diagram for the cycle. Property values were obtained from *Steam Tables* by Keenan *et al.* [1978]. Evaluation of properties using such tables is covered in most basic textbooks on engineering thermodynamics, for example, Wark [1983] and Van Wylen *et al.* [1994]. Properties at the various states were established as follows. State 1 is in the superheat region and the values of entropy and enthalpy were obtained from the superheat table of *Steam Tables* at the noted values of  $P_1$  and  $T_1$ . Also, since  $s_2$  is equal to  $s_1$ ,

$$s_2 = s_f + x_2(s_g - s_f) = 7.7055 = 0.5592 + x_2(8.2758 - 0.5592)$$

This yields the value for the quality  $x_2$  as 0.9224 and allows  $h_2$  to be calculated as

$$h_2 = h_f + x_2(h_g - h_f) = 163.4 + 0.9261(2572.5 - 163.4) = 2394.4 \text{ kJ/kg}$$

In these equations, quantities with  $f$  and  $g$  subscripts were obtained from the saturation table of *Steam Tables* at  $P_2$ . The value of enthalpy at state 4 was determined by first computing  ${}_3w_4$ , the work per unit mass for the process through the pump, using the expression for the reversible steady-flow work with negligible kinetic and potential energy changes [Wark, 1983]:



$${}_3w_4 = - \int_1^2 v \, dP = -v_3(P_4 - P_3)$$

where the specific volume  $v$  is assumed constant at  $v_3$  since a liquid is pumped. With  $v_3$  obtained from *Steam Tables* as  $v_f$  at  $P_3$ ,

$${}_3w_4 = -0.001\,007(1000 - 7) = -1.00 \text{ kJ/kg}$$

Writing Eq. (48.2) for the adiabatic process from state 3 to state 4,

$$h_4 = h_3 - {}_3w_4 = 163.39 - (-1.00) = 164.4 \text{ kJ/kg}$$

Proceeding with the solution for  $\dot{W}_{\text{net}}$ ,

$$\begin{aligned}\dot{W}_{\text{net}} &= \dot{W}_{\text{turbine}} + \dot{W}_{\text{pump}} = \dot{m}_1 w_2 + \dot{m}_3 w_4 = \dot{m}(h_1 - h_2) + \dot{m}_3 w_4 \\ &= 100.0(3435.2 - 2394.4) + 100.0(-1.00) = 103\,980 \text{ kW}\end{aligned}$$

where  ${}_1w_2$  was obtained by writing Eq. (48.2) between states 1 and 2. Next,  $\dot{Q}_H$  is determined by writing Eq. (48.1) for a control volume enclosing the steam generator and noting that there is no power transmitted across its surface. Equation (48.1) reduces to

$$\dot{Q}_H = \dot{m}h_1 - \dot{m}h_4 = 100.0(3435.2 - 164.4) = 327\,080 \text{ kW}$$

To find  $\eta$ , substitution into Eq. (48.4) yields

$$\eta = 103\,980/327\,080 = 0.318 \text{ or } 31.8\%$$

The solution for  $\dot{Q}_L$  can be obtained by either of two approaches. First, by Eq. (48.5),

$$|\dot{Q}_L| = \dot{Q}_H - \dot{W}_{\text{net}} = 327\,080 - 103\,980 = 223\,100 \text{ kW}$$

The solution is also obtained by writing the first law for the process between state 2 and state 3. The result is

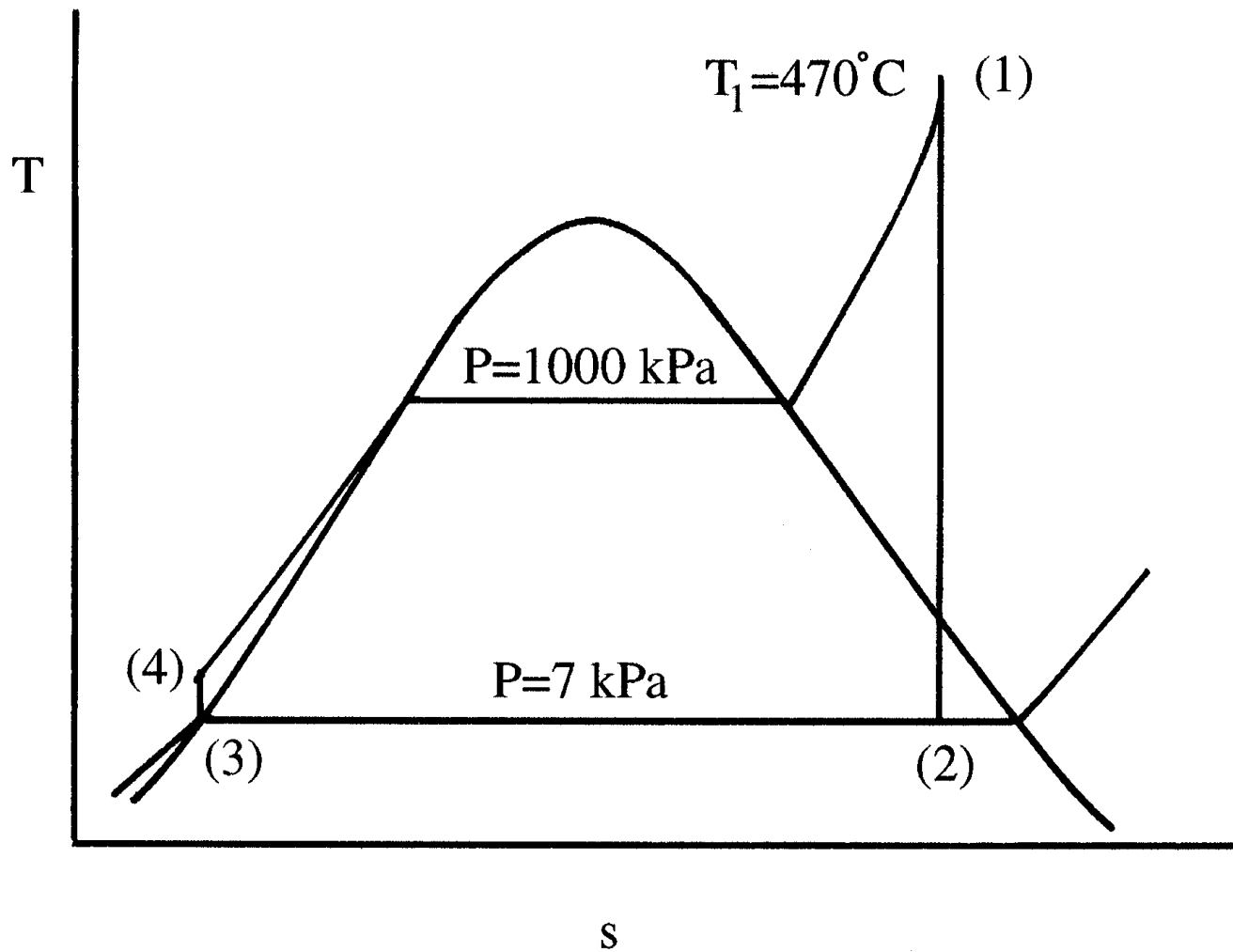
$$|\dot{Q}_L| = |\dot{m}(h_2 - h_3)| = |100(163.4 - 2394.4)| = 223\,100 \text{ kW}$$

The cycle in this example is known as the *Rankine cycle with superheat*. Modified forms of this cycle are widely used to provide shaft power to drive electric generators in steam-electric power plants and other power applications.

**Table 48.1** Properties at Cycle States for Example 48.1

State	Pressure, kPa	Temperature, °C	Quality, kg/kg	Entropy, kJ/kg K	Enthalpy, kJ/kg	Condition
1	1000	480	*	7.7055	3435.2	Superheated vapor
2	7	39	0.9261	7.7055	2394.4	Liquid-vapor mixture
3	7	39	0	—	163.4	Saturated liquid
4	1000	—	*	—	164.4	Subcooled liquid

\*Not applicable

**Figure 48.2** Temperature-entropy diagram for the steam power cycle in Example 48.1.

**Example 48.2.** Let  $\dot{Q}_H$  in Example 48.1 be supplied from a high-temperature source at a fixed temperature of  $500^\circ\text{C}$  and let the surrounding temperature be  $20^\circ\text{C}$ . Find the maximum thermal efficiency a cycle could have while operating between these regions and compare this value with  $\eta$  calculated in Example 48.1.

**Solution.** Equation (48.6) gives the expression for maximum thermal efficiency:

$$\eta_{\max} = (T_H - T_L)/T_H = [(500 + 273) - (20 + 273)]/[500 + 273] = 0.621 \text{ or } 62.1\%$$

compared to 31.8% for Example 48.1. The maximum value of cycle thermal efficiency was not realized because of the inherent **irreversibilities** associated with heat transfer across finite temperature differences in the heat reception and heat rejection processes for the cycle.

## 48.2 Refrigeration Cycles

The function of a refrigeration cycle is to cause heat energy to continuously flow from a low-temperature region to a region at a higher temperature. The operation of a refrigeration cycle is illustrated in Fig. 48.3(a), in which heat energy flows at the rate  $\dot{Q}_L$  from the low-temperature refrigerated region, heat is rejected at the rate  $\dot{Q}_H$  to the higher-temperature surroundings, and power  $\dot{W}_{\text{net}}$  is required. From Eq. (48.3) these are related as

$$|\dot{Q}_H| = \dot{Q}_L + |\dot{W}_{\text{net}}| \quad (48.7)$$

The performance parameter for conventional refrigeration cycles is termed *coefficient of performance* and is defined as

$$\beta = \dot{Q}_L / |\dot{W}_{\text{net}}| \quad (48.8)$$

The maximum value  $\beta$  can have when regions at uniform temperature  $T_H$  and  $T_L$  are considered is again derived from consideration of totally reversible cycles [Wark, 1983]. The expression, in terms of absolute temperatures, is

$$\beta_{\text{max}} = T_L / (T_H - T_L) \quad (48.9)$$

**Figure 48.3** Descriptions of refrigeration cycles: (a) Refrigeration cycle operation. (b) The simple vapor compression refrigeration cycle.

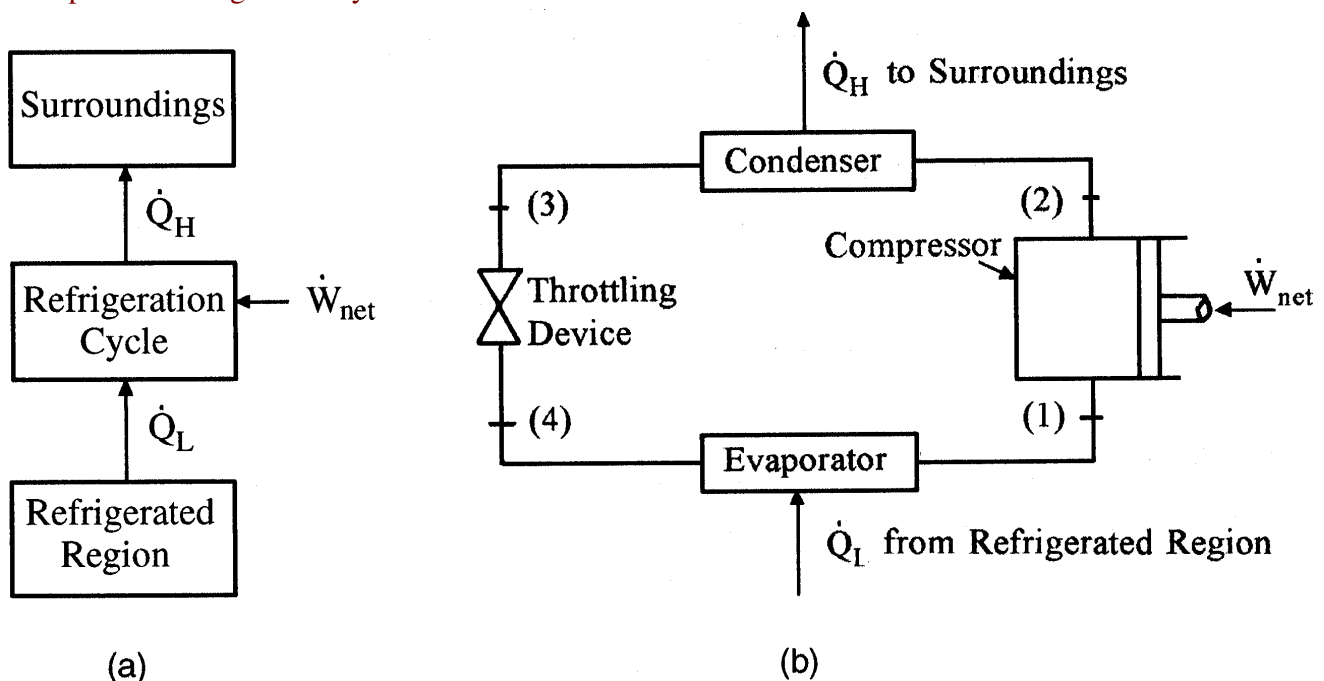


Figure 48.3(b) illustrates a simple vapor-compression refrigeration cycle. The compressor receives the refrigerant (working fluid) in the vapor phase at low pressure, state 1, and compresses it to state 2, where  $P_2 > P_1$ . Cooling at the condenser by means of a liquid or air coolant causes the vapor to condense to a liquid, state 3, after which it passes through a throttling device to the evaporator pressure. The refrigerant is a mixture of saturated liquid and saturated vapor at state 4. The liquid in the evaporator undergoes a phase change to vapor that is caused by the transfer of heat energy from the refrigerated region. The refrigerant leaves the evaporator as vapor at state 1, completing its thermodynamic cycle. The cycle illustrated in Fig. 48.3(b) is the basis for practical refrigeration cycles.

**Example 48.3.** A simple vapor compression refrigeration cycle, Fig. 48.3(b), has a refrigerating capacity of three tons (36 000 Btu/h) and operates with R134a as the refrigerant. The temperature of the refrigerated region is 15°F and the surroundings are at 90°F. Saturated vapor leaves the evaporator at 5°F and is compressed isentropically by the compressor to 150 psia. The refrigerant leaves the condenser as saturated liquid at 150 psia and flows through the throttling device to the condenser, in which the temperature is uniform at 5°F. Determine  $\dot{W}_{\text{net}}$ ,  $\beta$ , and the maximum coefficient of performance a refrigerator could have while operating between the refrigerated region and the surroundings.

**Solution.** Figure 48.4 shows the  $T$ - $s$  diagram for the cycle and the temperatures of the two regions. Table 48.2 lists values for the various properties obtained for R134a from the *ASHRAE Handbook [1993]* at the four states. The mass rate of flow is determined by applying Eq. (48.1) to the evaporator. The result is

$$\begin{aligned}\dot{Q} + \dot{m}h_1 &= \dot{m}h_2 = 36\,000 + \dot{m}(46.78) = \dot{m}(103.745) \\ \dot{m} &= 632.0 \text{ lbm/h}\end{aligned}$$

To find  $\dot{W}_{\text{net}}$ , application of Eq. (48.1) to the compressor yields

$$\dot{W}_{\text{net}} = {}_1\dot{W}_2 = \dot{m}(h_1 - h_2) = 632.0(103.745 - 120.3) = -10\,460 \text{ Btu/h}$$

To find  $\beta$ , Eq. (48.8) yields

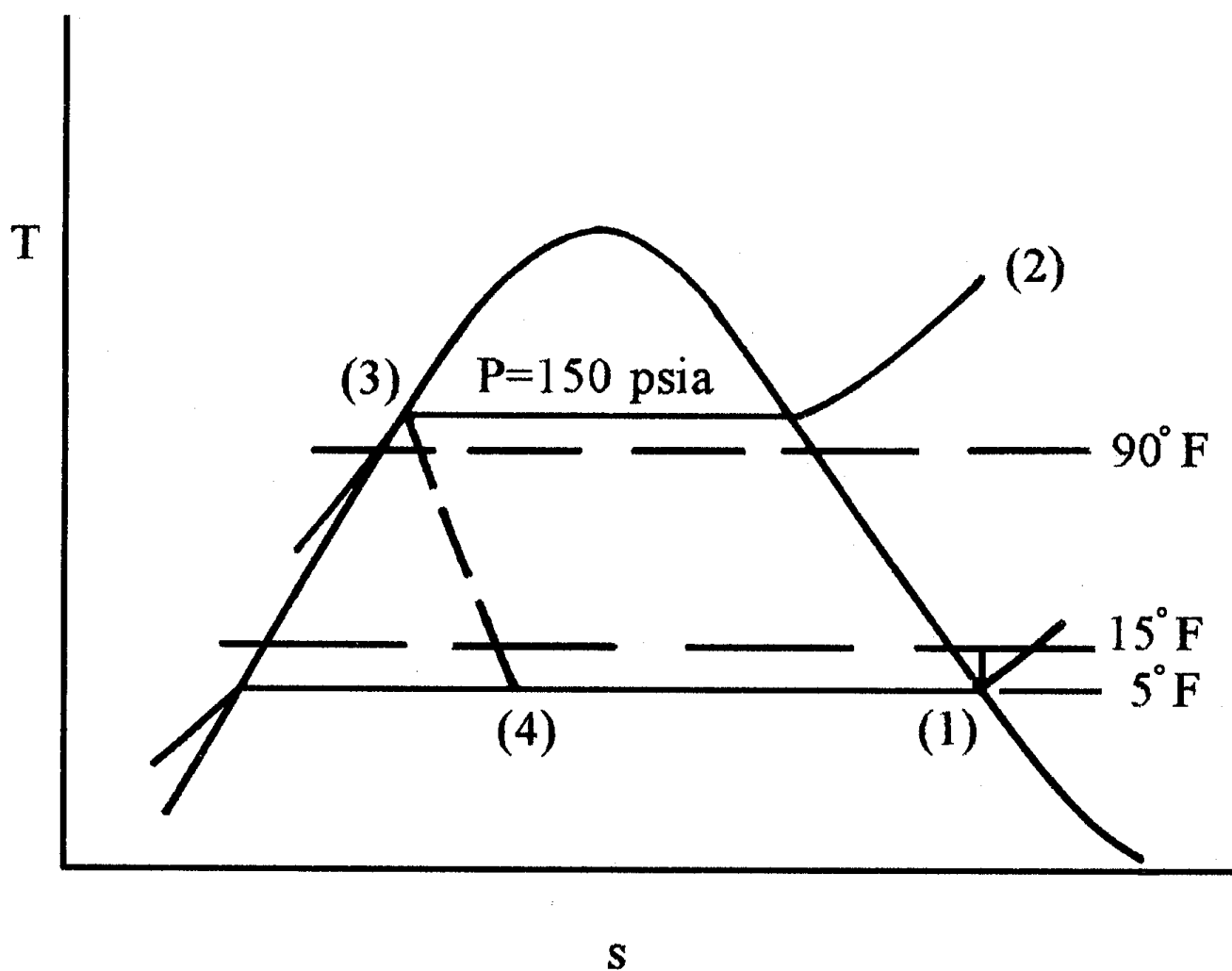
$$\beta = 36\,000/10\,460 = 3.44$$

The solution for maximum coefficient of performance is obtained by applying Eq. (48.9) as follows:

$$\beta_{\text{max}} = [15 + 460]/[(90 + 460) - (15 + 460)] = 6.33$$

Irreversibilities present due to finite temperature differences associated with the heat transfer processes and the irreversibility related to the throttling process cause  $\beta$  to be less than  $\beta_{\text{max}}$ .

**Figure 48.4** Temperature-entropy diagram for the refrigeration cycle in Example 48.3.



**Table 48.2** Properties at Cycle States for Example 48.3

State	Pressure, psia	Temperature, °F	Quality, lbm/lbm	Entropy, Btu/lbm R	Enthalpy, Btu/lbm	Condition
1	23.767	5	1.00	0.22470	103.745	Saturated vapor
2	150	118.7	*	0.22470	120.3	Superheated vapor
3	150	105.17	0.0	0.09464	46.78	Saturated liquid
4	23.767	5	0.368	0.10210	46.78	Liquid-vapor mixture

\*Not applicable

## Defining Terms

**Control volume:** A region specified by a control surface through which mass flows.

**First law of thermodynamics:** An empirical law that in its simplest form states that energy in its

various forms must be conserved.

**Heat energy:** Energy that is transferred across a control surface solely because of a temperature difference between the control volume and its surroundings. This form of energy transfer is frequently referred to simply as heat transfer.

**Irreversibilities:** Undesirable phenomena that reduce the work potential of heat energy. Such phenomena include friction, unrestrained expansions, and heat transfer across a finite temperature difference.

**Steady flow:** A condition that prevails in a flow process after all time transients related to the process have died out.

**Working fluid:** The substance that is contained within the apparatus in which the cycle is executed. The substance undergoes the series of processes that constitute the cycle.

## References

- ASHRAE. 1993. Refrigeration systems and applications. *ASHRAE Handbook*, I-P Edition. American Society of Heating, Refrigeration and Air-Conditioning Engineers, Atlanta, GA.
- Keenan, J. H., Keys, F. G., Hill, P. G., and Moore, J. G. 1978. *Steam Tables, SI Units*. John Wiley & Sons, New York.
- Van Wylen, G., Sonntag, R., and Borgnakke, C. 1994. *Fundamentals of Classical Thermodynamics*, 4th ed. John Wiley & Sons, New York.
- Wark, K. 1983. *Thermodynamics*, 4th ed. McGraw-Hill, New York.

## Further Information

- Proceedings of the American Power Conference*, Illinois Institute of Technology, Chicago, IL. Published annually.
- Stoecker, W. F. and Jones, J. W. 1982. *Refrigeration and Air Conditioning*, 2nd ed. McGraw-Hill, New York.
- Threlkeld, J. L. 1970. Mechanical vapor compression refrigeration cycles (chapter 3). *Thermal Environmental Engineering*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.

Bayazitoglu, Y., Sathuvalli, U. B. "Heat Transfer"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

### 49.1 Conduction

Fourier's Law of Heat Conduction • Thermal Conductivity of Materials • The Energy Equation • Limits of Fourier's Law • Dimensionless Variables in Heat Conduction

### 49.2 Convection

Thermal Boundary Layer • Heat Transfer Coefficient • Similarity Parameters of Convection • Forced Convection over Bodies (External Flows) • Forced Convection in Ducts (Internal Flows) • Free Convection

### 49.3 Radiation

Basic Quantities of Radiation • Radiation from a Blackbody • Intensity of Radiation • Radiative Properties of Real (Nonblack) Surfaces • Kirchhoff's Law • Shape Factors • Radiative Transfer Equation

### 49.4 Phase Change

Melting and Freezing • Condensation • Pool Boiling

**Yildiz Bayazitoglu**

*Rice University*

**Udaya B. Sathuvalli**

*Rice University*

There are two modes of **heat transfer**—diffusion and radiation. A *diffusion* process occurs due to the presence of a gradient (say, of temperature, density, pressure, concentration, electric potential, etc.) and *requires* a material medium. Both conduction and convection are diffusion processes. A radiative process, on the other hand, does *not* require a material medium.

This chapter will cover the basic ideas of conduction, convection, radiation, and phase change. The mode of heat transfer in solids and fluids at rest due to the exchange of the kinetic energy of the molecules or the drift of free electrons due to the application of a temperature gradient is known as *conduction*. The heat transfer that takes place due to mixing of one portion of a fluid with another via bulk movement is known as *convection*. All bodies that are at a temperature above absolute zero are thermally excited, and emit electromagnetic waves. The heat transfer that occurs between two bodies by the propagation of these waves is known as *radiation*. *Phase change* takes place when a body changes its state (i.e., solid, liquid, or gas) either by absorption or release of thermal energy.



## 49.1 Conduction

---

Heat conduction in solids can be attributed to two basic phenomena—lattice vibrations and transport of free electrons due to a thermal gradient.

The mechanism of heat conduction in a solid can be qualitatively described as follows. The geometrical structure in which the atoms in a solid are arranged is known as a *crystal lattice*. The positions of the atoms in the lattice are determined by the interatomic forces between them. When one part of a solid body is at a higher temperature than the rest of it, there is an increased amplitude of the lattice vibrations of the atoms in that region. These vibrations are eventually transmitted to the neighboring cooler parts of the solid, thus ensuring that conduction is in the direction of the temperature gradient. The mechanical oscillations of these atoms are known as *vibration modes*. In classical physics, each of these vibrations can have an *arbitrary* amount of energy. However, according to quantum mechanics, this energy is *quantized*. These quanta (in analogy with the theory of light) are called **phonons** and are the primary carriers of heat in dielectrics. The mathematical theory that studies heat conduction due to such lattice vibrations is known as the *phonon theory of conduction* [see [Ziman \(1960\)](#)].

On the other hand, in a metal there is an abundance of free electrons and they are the primary carriers of heat. There is a certain amount of conduction due to phonons, but it is usually negligible. A metal can be modeled as a lattice of positively charged nuclei surrounded by closed shells of electrons and immersed in a sea of free valence electrons. The free valence electrons are the carriers of energy and their motion in the absence of a thermal gradient is determined by the combined electric fields of the ions in the lattice. At thermal equilibrium (no net temperature gradient) there are as many electrons moving in one direction as in the opposite, and there is no net energy transport. However, when a temperature gradient is imposed, electrons that cross a given cross section of the solid have different temperatures (different velocities) at different points because the electrons collide with the ions in the lattice and reach local thermal equilibrium. This directed migration of the electrons eventually ceases when the thermal gradient disappears and the temperature becomes uniform throughout. This model can be used to determine the heat flux in terms of the applied temperature gradient and a material property known as the *thermal conductivity*.

### Fourier's Law of Heat Conduction

The important quantities in the study of heat conduction are the temperature, the heat flux, and the thermal conductivity of the material. The total amount of energy per unit time that crosses a given surface is known as the *heat flow* across the surface,  $Q$ . The energy per unit time per unit area across the surface is called the *heat flux*,  $\mathbf{q}$ . Heat flux (unlike heat flow) is a vector quantity and is directed along the normal to the surface at which it is measured.

The fundamental relation between the heat flux and the temperature gradient is **Fourier's law**, which states that heat flow due to conduction in a given direction is proportional (see [Fig. 49.1](#)) to the temperature gradient in that direction and the area normal to the direction of heat flow. Mathematically,

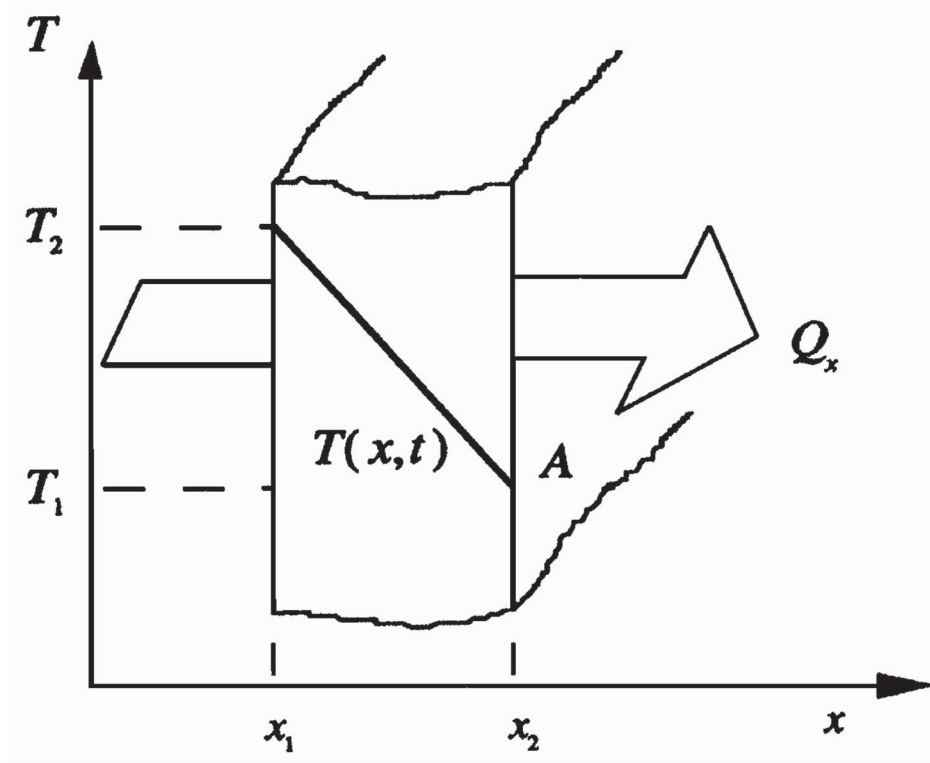
$$Q_x(x, t) = -\kappa A \frac{dT(x, t)}{dx} \quad (49.1)$$

where  $Q_x(x, t)$  is the heat flow in the positive  $x$  direction through the area  $A$ ,  $dT(x, t)/dx$  is the temperature gradient at point  $x$  at time  $t$ , and  $\kappa$  is the **thermal conductivity** of the material. The minus sign on the right-hand side of Eq. (49.1) is in accordance with the second law of thermodynamics and implies that heat flow due to conduction must occur from a point at higher temperature to a point at lower temperature. Equation (49.1) can be generalized for a solid to yield

$$\mathbf{q}(\mathbf{r}, t) = -\kappa \nabla T(\mathbf{r}, t) \quad (49.2)$$

where  $\mathbf{q}(\mathbf{r}, t)$  is the heat flux vector (i.e., heat flow per unit area,  $Q_x/A$ ) and  $T(\mathbf{r}, t)$  is the temperature at a point in the body whose position vector is  $\mathbf{r}$  at time  $t$ . In SI units, heat flux is measured in  $\text{W}/\text{m}^2$ , temperature gradient in  $^\circ\text{C}/\text{m}$ , and thermal conductivity in  $\text{W}/\text{m}^\circ\text{C}$ .

**Figure 49.1** Flow of heat in a slab.



## Thermal Conductivity of Materials

Materials display a wide range of thermal conductivities. Between gases (e.g., air) and highly conducting metals (e.g., copper, silver),  $\kappa$  varies by a factor of  $10^4$ . Metals have the highest thermal conductivity ( $\sim 10^4 \text{ W}/\text{m}^\circ\text{C}$ ) and gases such as hydrogen and helium are at the lower end of the range ( $\sim 10^{-2} \text{ W}/\text{m}^\circ\text{C}$ ). Materials such as oils and nonmetallic oxides have conductivities that range between  $10^{-1} \text{ W}/\text{m}^\circ\text{C}$  and  $10^2 \text{ W}/\text{m}^\circ\text{C}$  and are known as *thermal insulators*. Thermal conductivity is also known to vary with temperature. For some materials the variation over certain temperatures is small enough to be neglected. Generally, the thermal conductivity of metals decreases with increase in temperature. For gases (air,  $\text{CO}_2$ ,  $\text{H}_2$ , etc.) and most insulators (asbestos, amorphous carbon, etc.) it increases with increase in temperature.

Generally speaking, good thermal conductors are also good electrical conductors and vice versa. The free electron theory of metals accounts for their high conductivities and shows that the thermal and electrical conductivities of metals are related by the *Weidemann-Franz law*, which states that at absolute temperature  $T$ ,

$$\frac{\kappa}{\sigma_e T} = 2.23 \cdot 10^{-8} \text{ W}^- / \text{K}^2$$

where  $\sigma_e$  is the electrical conductivity of the metal.

## The Energy Equation

The study of heat conduction is primarily concerned with knowing the temperature distribution in a given body as a function of position and time. Consider an arbitrary body of volume  $V$  and area  $A$ , bounded by a surface  $S$ . Let the body have an internal heat source that generates heat at the rate of  $g(\mathbf{r}, t)$  per unit volume (in SI units, the units for this quantity are  $\text{W}/\text{m}^3$ ). Then, *conservation of energy* requires that the sum of heat fluxes that enter the body through its surface and the internal rate of heat generation in the body should equal the net rate of accumulation of energy in it. Mathematically,

$$\begin{aligned} - \oint_S \mathbf{q}(\mathbf{r}, t) \cdot \hat{\mathbf{n}} \, dA \Big|_{\text{conduction}} &= \oint_S \mathbf{q}(\mathbf{r}, t) \cdot \hat{\mathbf{n}} \, dA \Big|_{\text{other modes}} - \oint_V g(\mathbf{r}, t) \, dV \\ &+ \oint_V \rho C_p \frac{\partial T(\mathbf{r}, t)}{\partial t} \, dV \end{aligned} \quad (49.3)$$

where the left-hand side represents the net heat flux that enters the body by conduction, the first term on the right-hand side represents the corresponding term for other modes of heat transfer (i.e., convection or radiation), the second term on the right-hand side represents the total heat generated in the body, and the last term represents the accumulation of thermal energy in it. Here  $\hat{\mathbf{n}}$  is the outward drawn unit normal to the surface of the body,  $\rho$  is the density of the body, and  $C_p$  its specific heat. Equation (49.3) is known as the *integral form* of the **energy equation**. In Eq. (49.3) the internal heat generation can be due to electromagnetic heating, a chemical or nuclear reaction, etc.

By using Gauss's divergence theorem, the surface integrals of the heat flux can be converted into volume integrals, and this leads to

$$-\nabla \cdot \mathbf{q}(\mathbf{r}, t) \Big|_{\text{conduction}} = \nabla \cdot \mathbf{q}(\mathbf{r}, t) \Big|_{\text{other modes}} - g(\mathbf{r}, t) + \rho C_p \frac{\partial T(\mathbf{r}, t)}{\partial t} \quad (49.4)$$

The above equation is known as the *distributed form* of the energy equation. When conduction is the only mode of heat transfer, Fourier's law [Eq. (49.2)] can be used for a *homogeneous isotropic* medium (i.e., a medium whose thermal conductivity does not vary with position or

direction) in Eq. (49.4) to give

$$\nabla^2 T(\mathbf{r}, t) + \frac{1}{\kappa} g(\mathbf{r}, t) = \frac{1}{\alpha} \frac{\partial T(\mathbf{r}, t)}{\partial t} \quad (49.5)$$

where  $\alpha$  is the **thermal diffusivity** of the medium and is defined as

$$\alpha = \kappa / (\rho C_p) \quad (49.6)$$

The thermal diffusivity of a material governs the rate of propagation of heat during transient processes. In SI units it is measured in m<sup>2</sup>/s. When there is no internal heat generation in the body, Eq. (49.5) is known as the *diffusion equation*.

In many instances the temperature in the body may be assumed to be constant over its volume and is treated only as a function of time. Then Eq. (49.3) may be integrated to give

$$-Q|_{\text{conduction}} = Q|_{\text{other modes}} - G + \rho C_p V \frac{dT}{dt} \quad (49.7)$$

where  $T$  is the temperature of the body and  $G$  is the net heat generated in it. This is known as the *lumped form* of the energy equation.

One of the main problems of heat conduction is to obtain the solution of Eq. (49.5) subject to appropriate boundary conditions and an initial condition. The boundary conditions for the solution of Eq. (49.5) are of three kinds. The *constant temperature boundary condition* is obtained by prescribing the temperature at all points on the surface  $S$ . The *constant heat flux boundary condition* is invoked when the body is losing a known amount of heat to (or receiving heat from) the external ambient, for example, by conduction, convection, or radiation. The *mixed boundary condition* is used when a linear combination of the surface temperature and the heat flux leaving the surface is known. This situation typically occurs when a body is losing heat by convection from its surface to an ambient at a lower temperature. These boundary conditions are also known as *Dirichlet*, *Neumann*, and *Cauchy* boundary conditions, respectively. Once the boundary conditions and the initial condition for a given heat conduction situation are formulated, the problem reduces to a boundary value problem. The solutions to these problems are well documented in standard texts on heat conduction such as Carslaw and Jaeger [1959] and Ozisik [1993].

## Limits of Fourier's Law

In certain applications that deal with transient heat flow in very small periods of time, temperatures approaching absolute zero, heat flow due to large thermal gradients, and heat flow on a nano- or microscale (such as in thin films), Fourier's law of heat conduction [Eq. (49.1)] is known to be unreliable. This is because the diffusion equation predicts that the effect of a temperature gradient at a point  $\mathbf{r}$  that is established at time  $t$  should be instantly felt everywhere in the medium, implying that temperature disturbances travel at an *infinite* speed. In order to account for the *finite* speed of

propagation of heat, Eq. (49.1) has been modified as

$$\tau \frac{\partial \mathbf{q}(\mathbf{r}, t)}{\partial t} + \mathbf{q}(\mathbf{r}, t) = -\kappa \nabla T(\mathbf{r}, t) \quad (49.8)$$

where  $\tau$  is a *relaxation time*. When this equation is used in the differential form of the energy equation, Eq. (49.5), it results in a hyperbolic differential equation known as the *telegrapher's equation*:

$$\nabla^2 T(\mathbf{r}, t) + \frac{1}{\kappa} \left[ g(\mathbf{r}, t) + \frac{\alpha}{c^2} \frac{\partial g(\mathbf{r}, t)}{\partial t} \right] = \frac{1}{c^2} \frac{\partial^2 T(\mathbf{r}, t)}{\partial t^2} \quad (49.9)$$

where  $c$  is the wave propagation speed and is given by  $c = (\alpha/\tau)^{-1/2}$ . The solution for Eq. (49.9) indicates that heat propagates as a wave at a finite speed and is the basis of the theory of *heat waves*. For  $\tau \rightarrow 0$ , Eq. (49.8) reduces to Fourier's law [Eq. (49.2)], and when  $c \rightarrow \infty$ , Eq. (49.9) becomes the diffusion equation, Eq. (49.5).

## Dimensionless Variables in Heat Conduction

If a body is subject to a mixed boundary condition given by

$$-\kappa \hat{\mathbf{n}} \cdot \nabla T(\mathbf{r}, t) = h(T(\mathbf{r}, t) - T_\infty)$$

on the surface  $S$ , where  $h$  is the heat transfer coefficient from the surface of the body, then Eq. (49.5) can be nondimensionalized with the following variables: the *Fourier number*,  $\text{Fo} = \alpha t/L^2$  (i.e., the dimensionless time variable), the dimensionless heat generation variable  $[g(\mathbf{r}, t)L^2]/[\kappa(T_o - T_\infty)]$ , and the *Biot number*,  $\text{Bi} = hL/\kappa$ , where  $L$  and  $T_o$  are suitable length and temperature scales, respectively. The Fourier number denotes the ratio of the heat transferred by conduction to the heat stored in the body and is useful in solving transient problems. When the Fourier number is very large, transient terms in the solution of the diffusion equation may be neglected. The Biot number is a measure of the relative magnitudes of the heat transfer due to conduction and convection (or radiation) in the body. When the Biot number is very small ( $< 0.1$ ), the temperature in the body may be assumed to be a constant, and the lumped form of the energy equation may be used.

## 49.2 Convection

Convection takes place when a fluid moves over a solid and their temperatures are unequal. Heat transfer occurs due to actual material transport, unlike in the case of conduction. In many engineering applications the heat transfer due to convection may be calculated by using **Newton's law of cooling**. It states that if the heat transfer coefficient is  $h$ , the heat flux  $q$  due to convection from a surface at temperature  $T_w$  into a fluid at temperature  $T_f$  is given by

$$q = h(T_w - T_f) \quad (49.10)$$

There are three kinds of convection: forced, natural, and mixed. *Forced convection* takes place when the motion of the fluid that causes convection is sustained by an externally imposed pressure gradient. Forced convection typically occurs in systems such as blowers and air conditioners. Sometimes, even in the absence of external forces, pressure gradients are created due to differences in density that are caused by local heating in the fluid. This heat transfer is known as *free* or *natural* convection. *Mixed* convection, as the name implies, is the situation in which both forced and free convection are present.

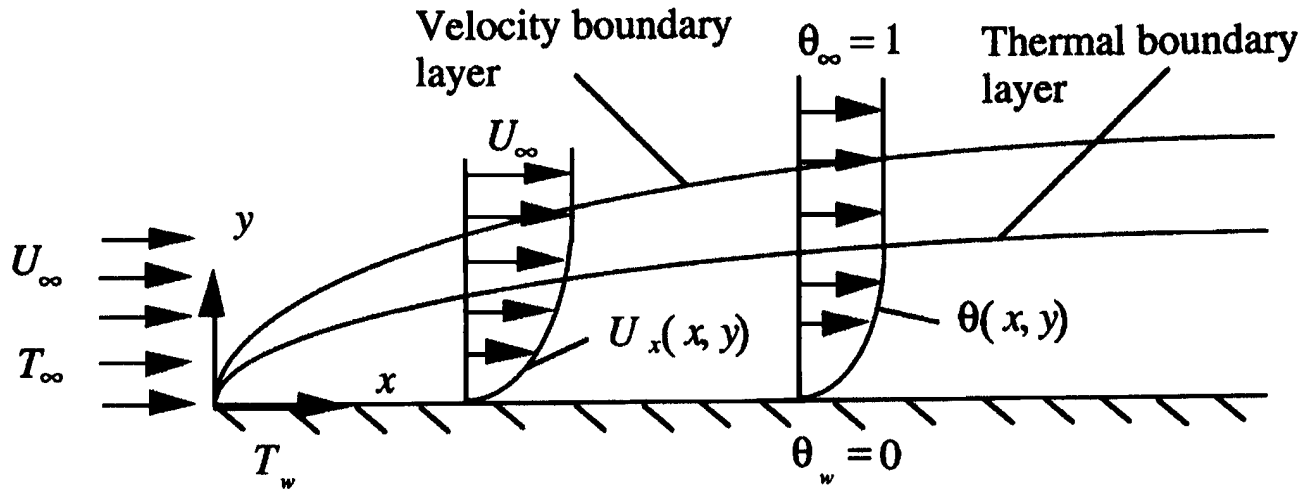
## Thermal Boundary Layer

In analogy with the momentum (or the velocity) boundary layer, a thermal boundary layer can be defined. Imagine a fluid of uniform temperature  $T_\infty$  flowing with a uniform velocity  $U_\infty$  (the free stream values) along a flat plate maintained at temperature  $T_w$  ( $< T_\infty$ ) (see Fig. 49.2). When the fluid comes into contact with the flat plate, the layers of fluid immediately adjacent to the plate attain the temperature  $T_w$ . Deeper into the fluid (away from the plate) the temperature of the fluid increases slowly. Finally, the temperature of the fluid becomes uniform at  $T_\infty$ . This suggests that there is a "temperature profile" and hence a "temperature gradient" in the layers of the fluid close to the flat plate. It is now convenient to define a dimensionless temperature  $\theta(x, y)$  at each point  $(x, y)$  of the fluid as

$$\theta(x, y) = \frac{T(x, y) - T_w}{T_\infty - T_w}$$

For a point very close to the wall, for example,  $y = 0$ , the temperature of the fluid is the same as that of the wall and, hence,  $\theta(x, y) = 0$ . At points sufficiently far away from the wall, the temperature of the fluid is  $T_\infty$  and, hence,  $\theta(x, y) \rightarrow 1$ . For each location  $x$  along the flat plate, there is a distance  $y$  from the plate at which  $\theta(x, y) = 0.99$ . The locus of all such points is called the *thermal boundary layer*. For practical purposes, locations *outside* the boundary layer are thermally not affected by the plate.

**Figure 49.2** The thermal boundary layer.



## Heat Transfer Coefficient

The **heat transfer coefficient**  $h$  [Eq. (49.10)] is a very useful concept in the determination of the heat transfer due to convection. Consider the case shown in Fig. 49.2. Very close to the wall, the fluid particles are stationary (due to the no-slip boundary condition) and the heat transfer is due to conduction in the fluid. Then the heat flux due to conduction at the wall is

$$q|_{\text{wall}} = -\kappa_f \left. \frac{\partial T(x, y)}{\partial y} \right|_{y=0} \quad (49.11)$$

where  $\kappa_f$  is the thermal conductivity of the fluid. In engineering applications the heat transfer between the fluid and the wall is related to the heat transfer coefficient using Newton's law of cooling by

$$h_x = -\kappa_f \frac{[\partial T(x, y)/\partial y]_{y=0}}{(T_\infty - T_w)} \quad (49.12)$$

The heat transfer coefficient  $h_x$  in Eq. (49.12) may vary along the length of the surface shown in Fig. 49.2, that is, it may be a function of  $x$ . It is therefore useful to define a *mean* or *bulk* value of the heat transfer coefficient over a finite length  $L$  as

$$h_m = \frac{1}{L} \int_0^L h_x \, dx \quad (49.13)$$

In SI units  $h_x$  is measured in  $\text{W}/\text{m}^2\text{ }^\circ\text{C}$ . The heat transfer coefficient is often expressed via an important dimensionless quantity known as the *Nusselt number*. The Nusselt number is defined as

$$\text{Nu}_x = \frac{h_x x}{\kappa_f} \quad (49.14)$$

and denotes the ratio of the actual convection heat flux to the conduction heat flux that would occur through a fluid slab of thickness  $x$ . Note the similarity of the Nusselt number with the Biot number (defined earlier). As in the case of  $h$ , it is useful to define a mean or bulk Nusselt number as follows:

$$\text{Nu}_m = \frac{h_m L}{\kappa_f} \int_0^L \frac{1}{x} \text{Nu}_x \, dx$$

## Similarity Parameters of Convection

Heat transfer by convection depends on the characteristics of the flow pattern of the fluid, its thermophysical properties, the geometry of the flow passage, and surface conditions. The flow patterns may be characterized as *laminar*, *transitional*, or *turbulent*. The thermophysical properties that determine heat transfer are the fluid density  $\rho$ , thermal conductivity  $\kappa_f$ , kinematic viscosity  $\nu$ , and specific heat  $C_p$ . The flow may be *external* or *internal* and the flow geometry may take such forms as flow over a flat plate, a cylinder, or a sphere, flow in a channel or an enclosed space, etc. Consider the laminar flow along a flat plate in which the local heat transfer coefficient at a location  $x$  is  $h_x$ . The local heat transfer coefficient  $h_x$  is a function of  $x$ ,  $\kappa_f$ , the local velocity  $U_x$ ,  $\rho$ ,  $\nu$ , the volume expansion coefficient of the fluid  $\beta$ , a characteristic temperature difference  $\Delta T$  ( $\sim [T_\infty - T_w]$ ), and  $C_p$ . It has been found that these variables can be grouped into a set of five nondimensional numbers in a functional relationship of the form

$$\text{Nu}_x = f(\text{Re}_x, \text{Pr}, \text{Gr}_x, \text{Ec}) \quad (49.16)$$

where the arguments of  $f$  are the nondimensional numbers discussed as follows. The importance of Eq. (49.16) is that it suggests that instead of determining  $h_x$  as a function of nine variables, it may be sought as a function of four nondimensional numbers. Further, Eq. (49.16) is obtained by a method known as *similarity analysis*. Similarity analyses usually indicate the actual functional form of Eq. (49.16) in advance.

### Reynolds Number

The local *Reynolds number* is defined by

$$\text{Re}_x = \frac{U_\infty x}{\nu} = \frac{U_\infty x \rho}{\mu} \quad (49.17)$$

where  $\mu$  is the dynamic viscosity of the fluid and is related to  $\nu$  by  $\nu = \mu/\rho$ . It denotes the relative magnitudes of the inertial to viscous forces that govern the flow. For large values of the Reynolds number, the inertial forces dominate (low-viscosity flows), whereas the viscous forces dominate in



small Reynolds number flows (creeping flows, viscous fluids).

### Prandtl Number

The *Prandtl number* is defined as

$$\text{Pr} = \frac{\nu}{\alpha_f} \quad (49.18)$$

and is the ratio of the kinematic viscosity and the thermal diffusivity  $\alpha_f (= \kappa_f / \rho C_p)$  of the fluid. It represents the relative rates of the diffusion of the momentum and thermal boundary layers. In liquid metals,  $\text{Pr} \ll 1$  and therefore the rate diffusion of thermal energy greatly exceeds that of momentum. In oils,  $\text{Pr} \gg 1$  and the converse holds.

### Grashof Number

The *Grashof number* is significant in the analysis of free convection, which is discussed later. It is defined as

$$\text{Gr}_x = \frac{g\beta x^3 (T_w - T_\infty)}{\nu^2} \quad (49.19)$$

where  $g$  is the acceleration due to gravity,  $\beta$  is the coefficient of volume expansion of the fluid, and  $T_w - T_\infty$  is a characteristic temperature difference that generates free convection. This number signifies the relative importance of the buoyant and the viscous forces in the flow. When there is no natural convection (for example, in microgravity situations  $g \approx 0$ , or when there is not enough thermal gradient to cause appreciable convection, i.e.,  $\Delta T$  is very small) and there is only forced convection,  $\text{Gr} = 0$ .

### Eckert Number

The *Eckert number* appears due to the inclusion of the *viscous dissipation* term in the equation for the conservation of the energy in the boundary layer. The viscous dissipation term denotes the rate at which mechanical energy is being dissipated into thermal energy due to the presence of viscous forces in the fluid. The viscous dissipation is proportional to the square of the velocity gradient and the dynamic viscosity of the fluid. The Eckert number is defined as

$$\text{Ec} = \frac{U_\infty^2}{C_p \Delta T} \quad (49.20)$$

and measures the kinetic energy of the flow with respect to the enthalpy difference across the thermal boundary layer.

The principal aim of convection can be said to be the determination of the Nusselt number as a function of the problem parameters that are expressed in terms of the above nondimensional numbers.

## Forced Convection over Bodies (External Flows)

When there is only forced convection and the viscous dissipation is small enough ( $Ec \approx 0$ ), Eq. (49.16) becomes

$$Nu_x = f(Re_x, Pr) \quad (49.21a)$$

A more specific functional form for the above relation may be chosen as

$$Nu_x = K(Re_x)^a(Pr)^b \quad (49.21b)$$

where  $K$ ,  $a$ , and  $b$  must be chosen by fitting curves to the experimental data or by analysis of the governing equations of convection. It is convenient to combine the Reynolds and Prandtl numbers to define a new nondimensional number, the *Peclet number*, as

$$Pe_x = Re_x Pr = \frac{U_\infty x}{\alpha_f}$$

Actual correlations in terms of these nondimensional numbers for different flow situations are given in standard textbooks of heat transfer [Bayazitoglu and Ozisik, 1988; Kreith and Bohn, 1986].

For laminar flow, correlations of the form given by Eq. (49.21) may be obtained by analytical methods. However, when the flow is turbulent, it is not easy to determine such correlations analytically and experimental methods must be used. In experiments it is often simpler to measure the drag coefficient  $c_x$  rather than the heat transfer coefficient. Furthermore, there are correlations between the heat transfer and drag coefficients. One such relation is the *Reynolds-Colburn* analogy given by

$$St_x Pr^{2/3} = \frac{1}{2} c_x \quad (49.22)$$

where

$$St_x = \frac{Nu_x}{Re_x Pr} = \frac{h_x}{\rho U_\infty C_p} \quad (49.23)$$

is the local *Stanton number* and is the nondimensional local heat transfer coefficient. The Reynolds-Colburn analogy may be used to develop expressions for local heat transfer coefficients. The relevant physical fluid properties that appear in these correlations are functions of the temperature and are usually evaluated at the **film temperature**  $T_f$ , which is defined as the mean boundary layer temperature:

$$T_f = \frac{T_w + T_\infty}{2} \quad (49.24)$$

## Forced Convection in Ducts (Internal Flows)

The following concepts (in addition to the ones discussed in the previous section) are needed to understand convection due to fluid flow in ducts. When the fluid (see Fig. 49.3) enters a tube whose wall is maintained at a temperature different from that of the fluid, there is heat transfer and the temperature distribution is such that, starting at the tube inlet, a thermal boundary layer develops and grows along the length of the tube until the thermal boundary layer occupies the entire width of the duct. The region from the entrance of the duct to the point where the thermal boundary layer reaches the axis of the duct is known as the *thermal entrance region*. The length of this region is called the *thermal entry length*,  $L_t$ . In this region the temperature distribution varies along the radius and the length of the duct. Beyond this region the temperature is only a function of the radial coordinate. This region is known as the *thermally developed region*. In the thermal entrance region the average temperature of the fluid at any cross section (averaged along the radial direction) is known as the *mean fluid temperature*. If the axial velocity of the fluid inside the duct is  $U(r)$ , the mean temperature is given by

$$T_m(z) = \frac{\int_0^R \rho C_p U(r) T(r, z) (2\pi r) dr}{\int_0^R \rho C_p U(r) (2\pi r) dr} \quad (49.25)$$

Based on this equation, the local heat transfer coefficient  $h_z$  is defined as

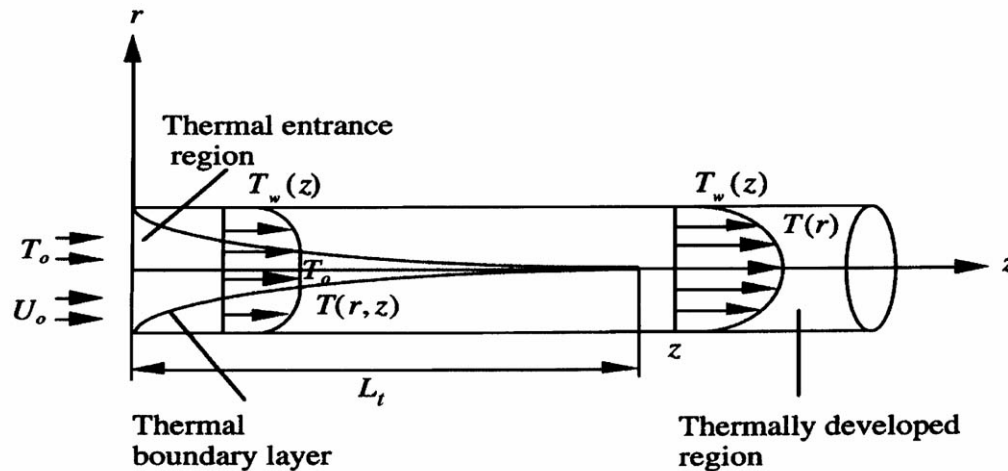
$$h_z = - \frac{\kappa_f}{T_m(z) - T_w(z)} \left. \frac{\partial T(r, z)}{\partial r} \right|_{r=R, \text{ wall}} \quad (49.26)$$

where  $T_w(z)$  is the tube wall temperature at  $z$ . A quantity known as the *logarithmic mean temperature difference (LMTD)*,  $\Delta T_{\ln}$ , which is useful in the determination of the total heat transfer between the fluid and wall of the duct (for example, in heat exchangers), is defined as

$$\Delta T_{\ln} = \frac{\Delta T_1 - \Delta T_2}{\ln(\Delta T_1 / \Delta T_2)} \quad (49.27)$$

where  $\Delta T_1$  and  $\Delta T_2$  are temperature differences between the fluid and the wall at the inlet and the outlet, respectively. Finally, in internal flows, the characteristic length parameter that is used to define the Nusselt number is known as the *hydraulic diameter* and is defined as  $D_h = 4A_c/P$ , where  $A_c$  is the cross-sectional area of the duct and  $P$  is its perimeter.

**Figure 49.3** Thermal entrance and thermally developed regions for flow in a duct.

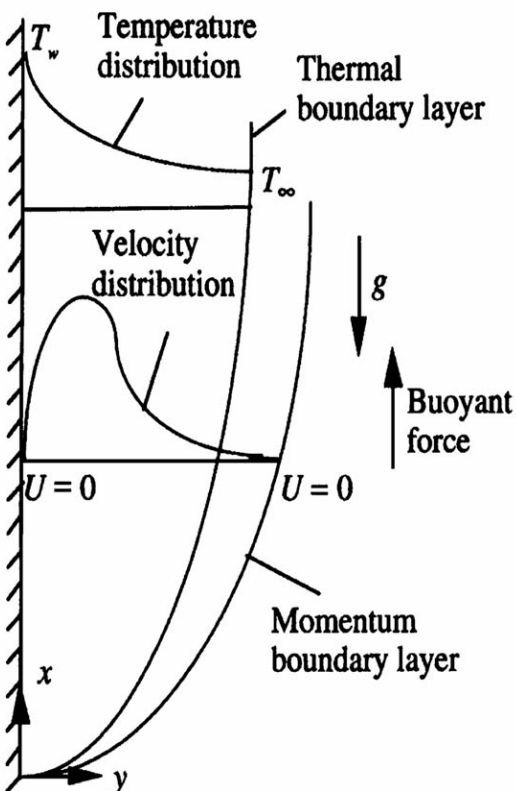


When the flow is turbulent, the heat transfer analysis is more complicated. There are several empirical correlations to calculate the heat transfer in ducts that carry turbulent flows. For flow in a smooth pipe of length  $L$  and diameter  $D$ , the *Dittus-Boelter* and the *Petukhov* equations are most commonly used [see [Bayazitoglu and Ozisik \(1988\)](#) for the actual relations].

## Free Convection

Consider a vertical flat plate at a temperature  $T_w$  immersed in a fluid at temperature  $T_\infty$ . If  $T_w > T_\infty$ , there is heat transfer from the plate to the fluid. The velocity and temperature profiles for this case of free convection are shown in [Fig. 49.4](#).

**Figure 49.4** Free convection over a heated vertical plate.



The Grashof number [defined in Eq. (49.19)] plays an important role in free convection. In the absence of forced convection and viscous dissipation, Eq. (49.16) becomes

$$\text{Nu}_x = f(\text{Gr}_x, \text{Pr}) \quad (49.28)$$

For gases,  $\text{Pr} \cong 1$  and Eq. (49.28) suggests that  $\text{Nu}_x = f(\text{Gr}_x)$ .

The role of the Grashof number in free convection is similar to that of the Reynolds number in forced convection. However, free convection may take place even in situations where forced convection dominates. In such cases the relative importance of these two modes of convection is determined by the parameter  $\text{Gr}/\text{Re}^2$ . If  $\text{Gr}/\text{Re}^2 \ll 1$ , forced convection dominates and free convection may be neglected. When  $\text{Gr}/\text{Re}^2 \gg 1$ , free convection dominates and the heat transfer correlations are of the form given by Eq. (49.28). And when  $\text{Gr}/\text{Re}^2 \sim 1$ , both free and forced convection must be considered. The heat transfer correlations are sometimes defined in terms of the *Rayleigh number*,  $\text{Ra}$ , given by

$$\text{Ra}_x = \text{Gr}_x \text{Pr} = \frac{g\beta(T_w - T_\infty)x^3}{\nu\alpha}$$

In some of the free convection correlations, the physical properties are evaluated at a mean temperature given by

$$T_m = T_w - 0.25(T_w - T_\infty)$$

Free convection inside closed volumes involves an interesting but complicated flow phenomenon. Consider a fluid between two large horizontal plates maintained at temperatures  $T_h$  and  $T_c$  and separated by a distance  $d$ . If the bottom plate is at temperature  $T_h$  ( $> T_c$ ), heat flows upward through the fluid, thus establishing a temperature profile that decreases upward. The hotter layers of fluid are at the bottom, whereas the colder (and less dense) layers are at the top. This arrangement of the fluid layers remains stationary and heat transfer is by conduction alone. This unstable state of affairs lasts as long as the buoyancy force does not exceed the viscous force. If the temperature difference (gradient) between the top and bottom plates is strong enough to overcome the viscous forces, fluid convection takes place. Theoretical and experimental investigations have shown that the fluid convection in such an enclosure occurs when the Rayleigh number reaches a critical value  $\text{Ra}_c$ , which is given in terms of the (characteristic) spacing  $d$  between the plates. The convective flow patterns that arise in such situations are known as *Bernard cells*.

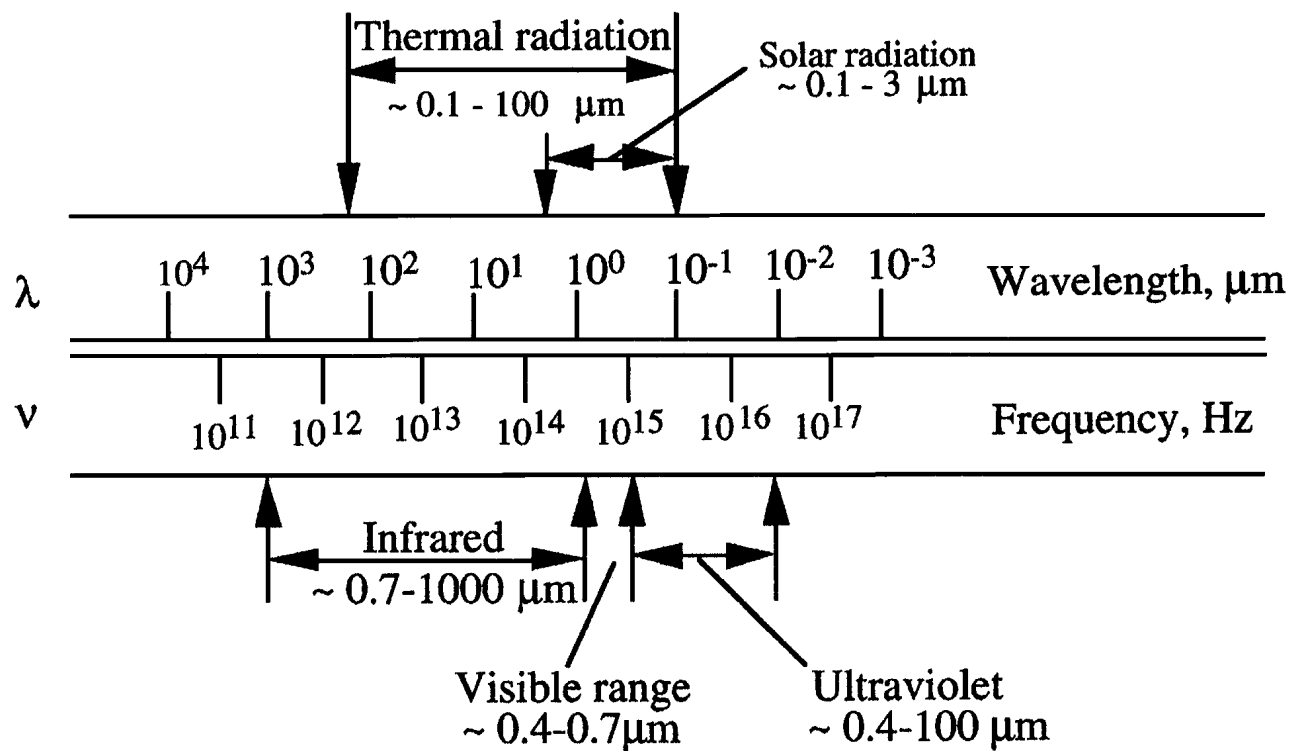
## 49.3 Radiation

All bodies that are above absolute zero temperature are said to be thermally excited. When bodies are thermally excited, they emit *electromagnetic waves*, which travel at the speed of light ( $c = 3 \cdot 10^8$  m/s in free space or vacuum). This radiation takes place at different wavelengths.

[Figure 49.5](#) shows the typical electromagnetic spectrum. The radiation that occurs at wavelengths

between  $\lambda = 0.1$  and  $100 \mu\text{m}$  is called *thermal radiation*, and the ensuing heat transfer is known as *radiative heat transfer*. The wave nature of radiation implies that the wavelength  $\lambda$  should be associated with a frequency  $\nu$ , which is given by  $\nu = c/\lambda$ . Bodies emit radiation at different wavelengths, and the intensity of the radiation varies with the wavelength. This is referred to as the *spectral distribution* of the radiation. Since electromagnetic waves can travel in vacuum, radiative transfer as opposed to conductive and convective transfer does *not* require a material medium.

**Figure 49.5** The range of the electromagnetic spectrum.



## Basic Quantities of Radiation

The radiation that is incident on the surface of a body may be absorbed, reflected, transmitted, or scattered. When all the incident radiation is absorbed by the body, the body is said to be *opaque*. If the material thickness required to absorb the radiation is very large compared to the characteristic thickness of the body, the radiation may be transmitted entirely. For example, glass is a very poor absorber of electromagnetic waves in the visible spectrum and is *transparent* to radiation in the visible spectrum. If the surface of the body is shiny and smooth, a good part of the radiation may be reflected (as with metallic surfaces). A material property can be associated with each of these phenomena: *absorptivity*,  $\alpha$ ; *reflectivity*,  $\rho$ ; and *transmissivity*,  $\tau$ . When these properties depend on the frequency of the incident radiation, they are said to exhibit *spectral dependence*, and when they depend on the direction of the incident radiation, they are said to exhibit *directional*

*dependence*. When radiation travels through a relatively transparent medium that contains inhomogeneities such as very small particles of dust, it gets *scattered*. Scattering is defined as the process in which a photon collides with one or more material particles and does not lose all its energy.

When radiative properties (or quantities) depend upon the wavelength and direction of the incident radiation, they are known as *monochromatic directional* properties. When these properties are summed over all directions *above* the surface in question, they are called *hemispherical properties*. When they are summed over the entire spectrum of radiation, they are referred to as *total properties*. When summed over all the directions above the surface in question as well as the entire spectrum of radiation, they are called *total hemispherical properties*.

The *hemispherical monochromatic emissive power*,  $E_\lambda(\lambda, T)$ , of a body is defined as the emitted energy leaving its surface per unit time per unit area at a given wavelength  $\lambda$  and temperature  $T$ . It is worth emphasizing that "emitted" refers to the *original* emission from the body—that is, the radiation due to thermal excitation of the atoms in the body. The *hemispherical monochromatic radiosity*,  $J_\lambda(\lambda, T)$ , refers to all the radiant energy (emitted and reflected) per unit area per unit time that leaves the surface. Radiant energy that is incident on a surface from all directions is known as *hemispherical monochromatic irradiation*,  $G_\lambda(\lambda)$ . In each of these cases the total quantity is obtained by integrating the monochromatic quantity over the entire spectrum, that is, from  $\lambda = 0$  to  $\infty$ .

## Radiation from a Blackbody

In the study of radiation it is useful to define an ideal surface in order to compare the properties of real surfaces. A body that absorbs *all* incident radiation regardless of its spectral distribution and directional character is known as a **blackbody**. Therefore, a blackbody is a perfect absorber. The radiation emitted by a blackbody at a given temperature is the maximum radiation that can be emitted by any body.

The emissive power  $E_{b,\lambda}(\lambda, T)$  of a blackbody at a given wavelength  $\lambda$  and absolute temperature  $T$  is given by **Planck's law** as

$$E_{b,\lambda}(\lambda, T) = \frac{c_1}{\lambda^5 \{e^{c_2/\lambda T} - 1\}} \quad (49.29)$$

where  $c_1 = 3.743 \cdot 10^8 \text{ W} \cdot \mu\text{m}^4/\text{m}^2$  and  $c_2 = 1.4387 \cdot 10^4 \mu\text{m K}$ . The units for  $c_1$  and  $c_2$  indicate that  $\lambda$  is measured in  $\mu\text{m}$  and  $E_{b,\lambda}(\lambda, T)$  in  $\text{W}/\text{m}^2$ . The *total emissive power*  $E_b(T)$  of the blackbody is the radiation emitted by it at all wavelengths. Mathematically,

$$E_b(T) = \int_{\lambda=0}^{\infty} E_{b,\lambda}(\lambda, T) d\lambda = \sigma T^4 \quad (49.30)$$

where  $\sigma$  is the *Stefan-Boltzmann constant*, given by

$$\sigma = \left( \frac{\pi}{c_2} \right)^4 \frac{c_1}{15} = 5.67 \cdot 10^{-8} \text{ W/m}^2 \text{K}^4$$

Equation (49.30) is known as the **Stefan-Boltzmann law** of radiation. Experiments have shown that the peak values of the emission increase with  $T$  and occur at shorter wavelengths. This fact is formally given by *Wien's displacement law*, which states that

$$(\lambda T)_{\max} = 2897.6 \text{ } \mu\text{m K} \quad (49.31)$$

[This relation can also be obtained by differentiating Planck's law, Eq. (49.29).] Often, in practice, it is required to know what fraction of the total emission occurs between two given wavelengths. The *blackbody radiation function*,  $f_{0,\lambda}(T)$ , is defined by

$$f_{0,\lambda}(T) = (\sigma T^4)^{-1} \int_{\lambda'=0}^{\lambda} E_{b,\lambda}(\lambda', T) d\lambda' \quad (49.32)$$

and tabulated in standard textbooks. It may be used to find the emission of a blackbody between any two given wavelengths  $\lambda_1$  and  $\lambda_2$  by

$$f_{\lambda_1,\lambda_2}(T) = f_{0,\lambda_2}(T) - f_{0,\lambda_1}(T) \quad (49.33)$$

## Intensity of Radiation

The fundamental quantity that is used to represent the amount of radiant energy transmitted in a given direction is the *intensity of radiation*. The *spectral intensity of radiation* is defined as the radiative energy per unit time per unit area normal to the direction of propagation per unit solid angle per unit wavelength. The total intensity refers to the intensity summed over all wavelengths. With reference to Fig. 49.6, let  $\Delta Q$  represent the energy radiated per unit time and confined to a solid angle  $d\omega$  around the solid angle  $\omega$ . The *total intensity of radiation* due to the radiating surface at  $Q$  is then defined mathematically as

$$I(\mathbf{r}, \mathbf{s}) = \frac{dq}{d\omega} \frac{1}{\cos \theta} = \frac{dq}{da} \frac{r^2}{\cos \theta} \quad (49.34)$$

where  $q = \lim_{\Delta A \rightarrow 0} (\Delta Q / \Delta A)$  is the heat flux that is measured at the radiation source  $Q$ , and  $d\omega$  is the solid angle subtended by the area  $da$  at  $Q$  (which is equal to  $da/r^2$ ). If the intensity at a point  $P$  is constant, Eq. (49.34) shows that the flux intercepted by an area varies inversely with the square of its distance from the source and directly as the cosine of the angle between the normals of the radiating and intercepting surfaces. This is known as *Lambert's cosine law*. Since energy conservation implies that all the flux (emitted as well as reflected) that leaves  $Q$  must cross the hemisphere above  $Q$ ,



$$q = \int_{\text{hemisphere}} I(\mathbf{r}, \hat{\mathbf{s}}) \hat{\mathbf{n}} \cdot \hat{\mathbf{s}} d\Omega \quad (49.35)$$

where  $q$  is the total flux radiated by the surface (i.e., at all wavelengths). If only the intensity due to the emission from the surface is considered, the flux that leaves the hemisphere above the surface is known as the *total hemispherical emissive power* and is given by

$$E = \int_{\text{hemisphere}} I(\mathbf{r}, \hat{\mathbf{s}}) \hat{\mathbf{n}} \cdot \hat{\mathbf{s}} d\Omega \quad (49.36)$$

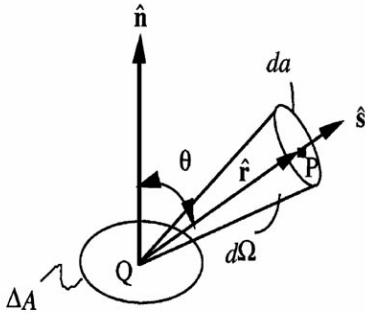
If  $I(\mathbf{r}, \hat{\mathbf{s}})$  is uniform, then the radiating surface is known as a **diffuse** emitter, and Eq. (49.36) becomes

$$q = \pi I \quad (49.37)$$

The radiation emitted by a blackbody is diffuse. Therefore, the total intensity of radiation from a blackbody [due to Eqs. (49.30) and (49.37)] is

$$I_b = \frac{\sigma T^4}{\pi} \quad (49.38)$$

**Figure 49.6** The definition of intensity of radiation.



## Radiative Properties of Real (Nonblack) Surfaces

The important radiative properties of a real (nonblack) surface are its emissivity, absorptivity, transmissivity, and reflectivity. In the following sections we consider only the hemispherical monochromatic and total properties. For directional monochromatic quantities see Siegel and Howell [1981] or Modest [1993].

### Emissivity

The *hemispherical monochromatic emissivity* of a nonblack surface at a temperature  $T$  is defined as the ratio of the hemispherical monochromatic emissive power of the surface  $E_\lambda(\lambda, T)$  and the corresponding hemispherical monochromatic emissive power of a blackbody. That is,

$$\varepsilon_\lambda(\lambda, T) = \frac{E_\lambda(\lambda, T)}{E_{b,\lambda}(\lambda, T)} \quad (49.39)$$

The hemispherical total emissivity is the ratio of the corresponding hemispherical total

quantities:

$$\varepsilon(T) = \frac{E(T)}{E_b(T)} = (\sigma T^4)^{-1} \int_{\lambda=0}^{\infty} \varepsilon_{\lambda}(\lambda, T) E_{b,\lambda}(\lambda, T) d\lambda \quad (49.40)$$

A surface whose monochromatic emissivity does not depend on the wavelength is called an ideal **gray body**. For a gray body,

$$\varepsilon_{\lambda}(\lambda, T) = \varepsilon(T) \quad (49.41)$$

### Absorptivity

The *hemispherical monochromatic absorptivity* is defined as the fraction of the incident hemispherical monochromatic irradiation that is absorbed by the surface, that is,

$$\alpha_{\lambda}(\lambda, T) = \frac{G_{\lambda}(\lambda)|_{\text{absorbed}}}{G_{\lambda}(\lambda)} \quad (49.42)$$

From the definitions of intensity and radiosity, the *hemispherical total absorptivity* of a surface may be defined as

$$\alpha(T, \text{source}) = \frac{\int_{\lambda=0}^{\infty} \alpha_{\lambda}(\lambda, T) \left[ \int_{\text{hemisphere}} I_{\lambda,i}(\mathbf{r}, \lambda, \mathbf{S}) \cos \theta_i d\Omega \right] d\lambda}{\int_{\lambda=0}^{\infty} \int_{\text{hemisphere}} I_{\lambda,i}(\mathbf{r}, \lambda, \mathbf{S}) \cos \theta_i d\Omega d\lambda} \quad (49.43)$$

where  $I_{\lambda,i}(\mathbf{r}, \lambda, \mathbf{S})$  refers to the incident monochromatic intensity of radiation and  $\theta_i$  is the angle between the incident radiation and the outward drawn normal to the surface. Equations (49.42) and (49.43) indicate that the monochromatic and the total absorptivities depend on  $I_{\lambda,i}$  and, hence, on the source of radiation. Consequently, unlike the values of  $\varepsilon$ , the values of  $\alpha$  cannot be easily tabulated.

### Reflectivity

Radiation that is incident on a surface gets reflected. If the surface is smooth, the incident and reflected rays are symmetric with respect to the normal at the point of incidence, and this is known as *specular reflection*. If the surface is rough, the incident radiation is scattered in all directions. An idealized reflection law assumes that, in this situation, the intensity of the reflected radiation is constant for all angles of reflection and independent of the direction of radiation. The *hemispherical monochromatic reflectivity*,  $\rho_{\lambda}(\lambda)$ , is defined as the fraction of the reflected radiant energy that is incident on a surface. The hemispherical total reflectivity is merely

$$\rho = \int_{\lambda=0}^{\infty} \rho_{\lambda}(\lambda) d\lambda$$

### Transmissivity

If a body is semitransparent to radiation (as glass is to solar radiation), part of the incident radiation is reflected by the surface and part of it is absorbed. In this case the sum of the absorptivity and reflectivity is less than unity, and the difference is known as *transmissivity*,  $\tau$ . Therefore

$$\alpha_{\lambda} + \rho_{\lambda} + \tau_{\lambda} = 1 \quad \text{and} \quad \alpha + \rho + \tau = 1$$

For an opaque body  $\tau = 0$ , implying that  $\rho_{\lambda} = 1 - \alpha_{\lambda}$  and  $\rho = 1 - \alpha$ .

### Kirchhoff's Law

*Kirchhoff's law* states that the hemispherical monochromatic emissivity  $\varepsilon_{\lambda}(\lambda, T)$  is equal to the hemispherical monochromatic absorptivity  $\alpha_{\lambda}(\lambda, T)$  at a given temperature  $T$ . That is,

$$\varepsilon_{\lambda}(\lambda, T) = \alpha_{\lambda}(\lambda, T) \quad (49.44)$$

However, the hemispherical total values of the emissivity and absorptivity are equal only when one of the two following conditions is met: either (1) the incident radiation of the receiving surface must have a spectral distribution that is the same as that of the emission from a blackbody at the same temperature, or (2) the receiving surface must be an ideal gray surface—that is,  $\varepsilon_{\lambda}(\lambda, T)$  is not a function of  $\lambda$ .

### Shape Factors

In the exchange of radiant energy between two surfaces, it is important to find the fraction of energy that leaves one surface and strikes the other. The flux of the energy between any two given surfaces can be expressed in terms of the *shape factor* (also known as *view factor*). The shape factor is purely a function of the geometry of the two surfaces and their relative orientation. The shape factors for the two surfaces shown in [Fig. 49.7](#) are given by

$$F_{1,2} = \frac{1}{A_1} \int_{A_1} \int_{A_2} \frac{\cos \theta_1 \cos \theta_2}{\pi r^2} dA_2 dA_1$$

and

$$F_{2,1} = \frac{1}{A_2} \int_{A_2} \int_{A_1} \frac{\cos \theta_2 \cos \theta_1}{\pi r^2} dA_1 dA_2 \quad (49.45)$$

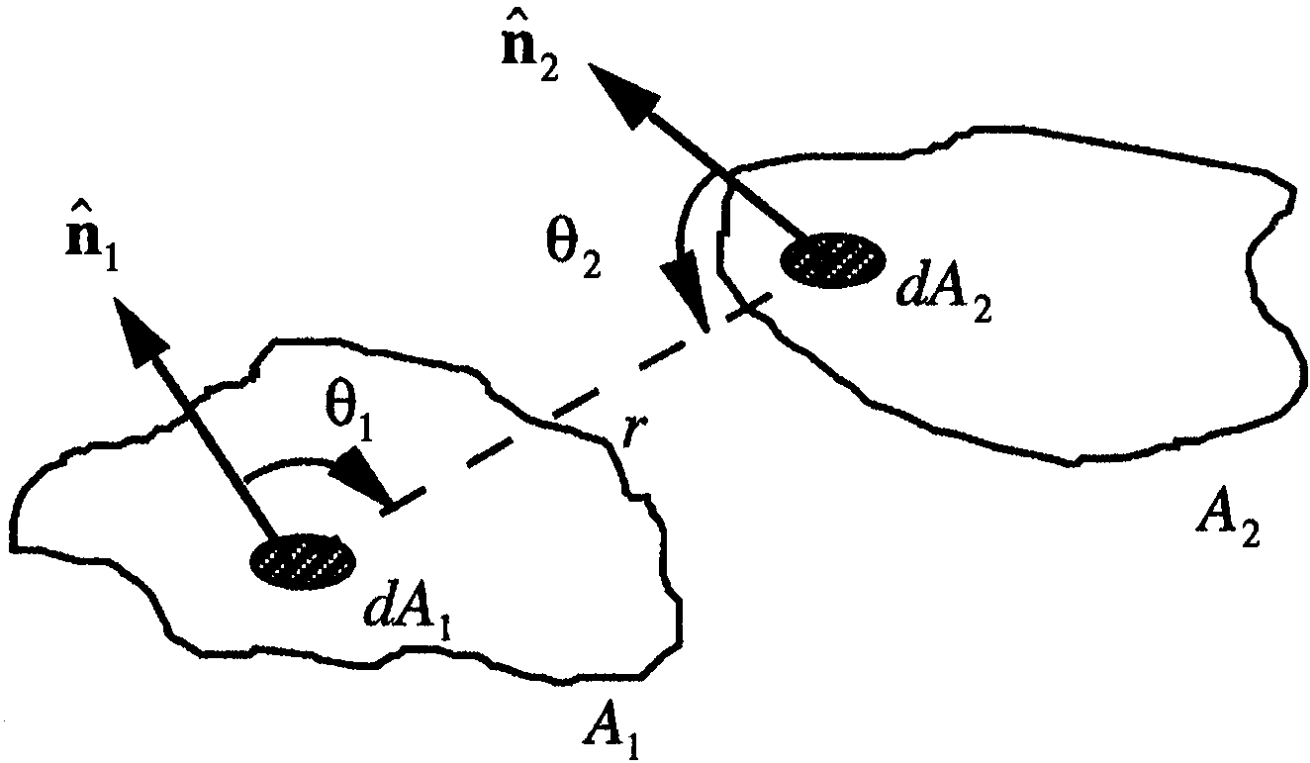
If the radiation between the two surfaces is diffuse, the fraction  $Q_{1,2}$  of the radiant energy that

leaves surface  $A_1$  and is intercepted by  $A_2$  is given by

$$Q_{1,2} = A_1 J_1 F_{1,2} \quad (49.46)$$

where  $J_1$  is the radiosity of surface  $A_1$ . In the above relations the order of the subscripts is important since  $F_{1,2} \neq F_{2,1}$ .

**Figure 49.7** Determination of the shape factor.



Further, shape factors possess the following properties. Any two given surfaces  $A_1$  and  $A_2$  satisfy the *reciprocal* property:

$$A_1 F_{1,2} = A_2 F_{2,1} \quad (49.47)$$

If a surface  $A_1$  is divided into  $n$  parts ( $A_{1_1}, A_{1_2}, \dots, A_{1_n}$ ) and the surface  $A_2$  into  $m$  parts, the following *additive property* holds:

$$A_1 F_{1,2} = \sum_n \sum_m A_{1_n} F_{1_n,2_m} \quad (49.48)$$

If the interior of an enclosure is divided into  $n$  parts, each with a finite area  $A_i, i = 1, 2, \dots, n$ , then the *enclosure property* states that

$$\sum_{j=1}^n F_{i-j} = 1, \quad i = 1, 2, \dots, n \quad (49.49)$$

## Radiative Transfer Equation

When a medium absorbs, emits, and scatters the radiant energy flowing through it, it is called a **participating medium**. An example of a participating medium is a gas with particles. Experiment shows that the intensity of radiation decays exponentially with the distance traveled in the medium. Further, if it encounters the particles in the gas, it gets scattered. Scattering takes place when one or more of diffraction, reflection, or refraction occur. Taking absorption and scattering into account, the transmissivity of a medium [also see "Radiative Properties of Real (Nonblack) Surfaces"] may be written as

$$\tau_\lambda = e^{-(\delta_\lambda + \varsigma_\lambda)s} = e^{\beta_\lambda s}$$

where  $\delta_\lambda$  is the *monochromatic scattering coefficient*,  $\varsigma_\lambda$  is the *monochromatic absorption coefficient*, and  $s$  is the thickness of the medium.  $\beta_\lambda$  is known as the *monochromatic extinction coefficient* of the medium.

In the solution of heat transfer problems that involve multiple modes of heat transfer, the energy equation must be solved. In such cases the term  $\nabla \cdot q|_{\text{radiation}}$  [also see Eq. (49.4)] must be calculated. Consider an absorbing, emitting, scattering medium with a monochromatic scattering coefficient  $\delta_\lambda$  and a monochromatic absorption coefficient  $\varsigma_\lambda$ . Let a beam of monochromatic radiation of intensity  $I_\lambda(s, \hat{\Omega}, t)$  travel in this medium along the direction  $\hat{\Omega}$ . The *radiative transfer equation* (RTE) is obtained by considering the energy balance for radiation due to emission, absorption, and in and out scattering in a small volume of this medium. The steady divergence of the heat flux for this case is then given by

$$\nabla \cdot q|_{\text{radiation}} = 4\pi \int_0^\infty \varsigma_\lambda I_{b,\lambda}(T) d\lambda - \int_0^\infty \varsigma_\lambda \left[ \int_{4\pi} I_\lambda(s, \hat{\Omega}) d\hat{\Omega} \right] d\lambda$$

where  $I_{b,\lambda} = E_{b,\lambda}/\pi$  and  $I_\lambda(s, \hat{\Omega})$  is obtained as a solution to the RTE.

## 49.4 Phase Change

Phase change occurs when a substance (solid, liquid, or gas) changes its state due to the absorption or release of energy. Below we briefly discuss melting/freezing (solid  $\leftrightarrow$  liquid) and condensation/boiling (liquid  $\leftrightarrow$  vapor).

### Melting and Freezing

The analysis of situations in which a solid melts or a liquid freezes involves the *phase change* or

*moving boundary problem*. The solutions to such problems are important in the making of ice, the freezing of food, the solidification of metals in casting, the cooling of large masses of igneous rock, the casting and welding of metals and alloys, etc. The solution of such problems is difficult because the *interface* between the solid and the liquid phases is in motion, due to the absorption or release of the latent heat of fusion or solidification. Thus, the location of the moving interface is not known beforehand and is a part of the solution to the problem. The heat transfer problem in these situations can be treated as a problem in heat conduction with moving boundaries, which requires a solution to Eq. (49.6) under appropriate boundary conditions. The solution to these problems is beyond the scope of this article, and the interested reader is referred to works such as Ozisik [1993].

## Condensation

Consider a vapor that comes into contact with a surface that is maintained at a temperature that is lower than the saturation temperature of the vapor. The resulting heat transfer from the vapor to the surface causes immediate *condensation* to occur on the surface. If the surface is cooled continuously so as to maintain it at a constant temperature, and the condensate is removed by motion due to gravity, it is covered with a thin film of condensed liquid. This process is known as *filmwise condensation*. Filmwise condensation usually occurs when the vapor is relatively free of impurities. Sometimes, when there are oily impurities on surfaces or when the surfaces are highly polished, the film of condensate breaks into droplets. This situation is known as *dropwise condensation*. In filmwise condensation the presence of the liquid film acts as a barrier for the heat transfer between the vapor and the surface. In dropwise condensation there is less of a barrier between the vapor and the surface. As a result, heat transfer coefficients are five to ten times the corresponding values for film condensation. Since fluid motion plays a significant role in condensation and boiling, they are studied along with convective heat transfer processes. However, there are significant differences between convective heat transfer during phase change and single-phase processes. The study of condensation and boiling is useful in the design of condensers and boilers, which are two widely used types of heat exchangers.

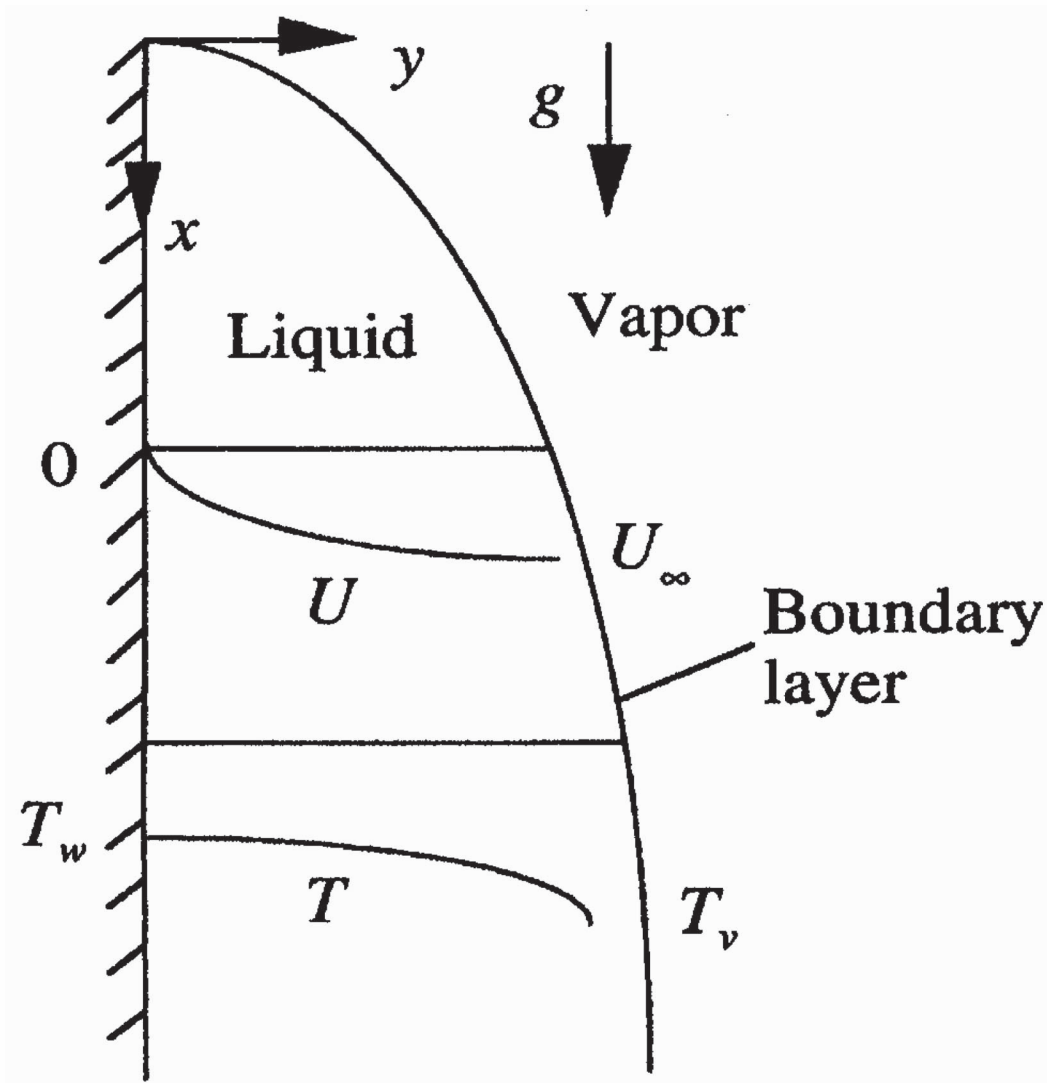
### Filmwise Condensation

Consider a cold vertical plate of length  $L$  maintained at a constant temperature  $T_w$  and exposed to saturated vapor at temperature  $T_v$ . Figure 49.8 shows the formation of the film of condensate on the plate. By the analysis of the boundary layer equations for the film of condensate, Nusselt obtained the heat transfer coefficient  $h_x$  at location  $x$  (with respect to Fig. 49.8) for this case. Nusselt's analysis is also valid for condensation outside a long cylinder or tube if its radius is large compared to the thickness of the condensing film. The actual correlation for this case and for filmwise condensation on other surfaces can be found in standard textbooks [see, for instance, Bejan (1984) or Chapman (1987)]. In condensation problems the Reynolds number of the condensate flow is given by

$$\text{Re} = \frac{4\dot{m}}{\mu_f P} = \frac{4A_t h_m (T_v - T_w)}{h_{fg} \mu_f P}$$

where  $\dot{m}$  is the mass flow rate of the condensate at the lowest part of the condensing surface. The wetted perimeter  $P$  is defined as  $w$  for a vertical plate of width  $w$  and  $D$  for a vertical cylinder/tube with outside diameter  $D$ .  $A_t$  is the total condensing surface area,  $h_m$  is the mean heat transfer coefficient, and  $h_{fg}$  is the latent heat of condensation.

**Figure 49.8** Filmwise condensation on a vertical surface.



Condensers are usually designed with horizontal tubes arranged in vertical tiers. If the drainage from one tube is assumed to flow smoothly onto the tube below, then, for a vertical tier of  $N$  tubes each of diameter  $D$ ,

$$[\text{Nu}_m]_{\text{for } N \text{ tubes}} = \frac{1}{N^{1/4}} [\text{Nu}_m]_{\text{for one tube}}$$

## Dropwise Condensation

Experiments have shown that if traces of oil or other selected substances (known as **promoters**) are present either in steam or on the condensing surface, the condensate film breaks into droplets. The droplets grow, coalesce, and run off the surface, leaving a greater portion of the surface exposed to the incoming steam. Therefore, dropwise condensation is a more efficient method of heat transfer. Typical heat transfer coefficients of  $5.7 \cdot 10^4$  to  $50 \cdot 10^4 \text{ W/m}^2\text{°C}$  may be obtained as opposed to heat transfer coefficients in the order of  $5 \cdot 10^3 \text{ W/m}^2\text{°C}$  for filmwise condensation. If sustained dropwise condensation can be maintained, it will result in considerable reduction in the size of condensers and their cost. Research in this area has thus been aimed at producing long-lasting dropwise condensation via the use of promoters. A satisfactory promoter must repel the condensate, stick tenaciously to the substrate, and prevent oxidation of the surface. Some of the most popular promoters are fatty acids such as oleic, stearic, and linoleic acids. In order to prevent failure of dropwise condensation as a result of oxidation, coatings of noble metals such as gold, silver, and palladium have been used in the laboratory. Such coatings have sustained over 10 000 hours of continuous dropwise condensation. However, this method is too expensive for use on an industrial scale.

Another problem in dropwise condensation is the presence of a noncondensable gas. If a noncondensable gas such as air is present in the vapor even in small amounts, the heat transfer coefficient is significantly reduced. When the vapor condenses, the noncondensable gas is left at the surface of the condensate and the incoming vapor must diffuse through this body of vapor-gas mixture before reaching the condensing surface. This diffusion process reduces the partial pressure of the condensing vapor and in turn its saturation temperature; that is, the temperature of the surface of the layer of condensate is lower than the bulk saturation temperature of the vapor. Therefore, in practical applications, there is a provision to vent the noncondensable gas that accumulates inside the condenser.

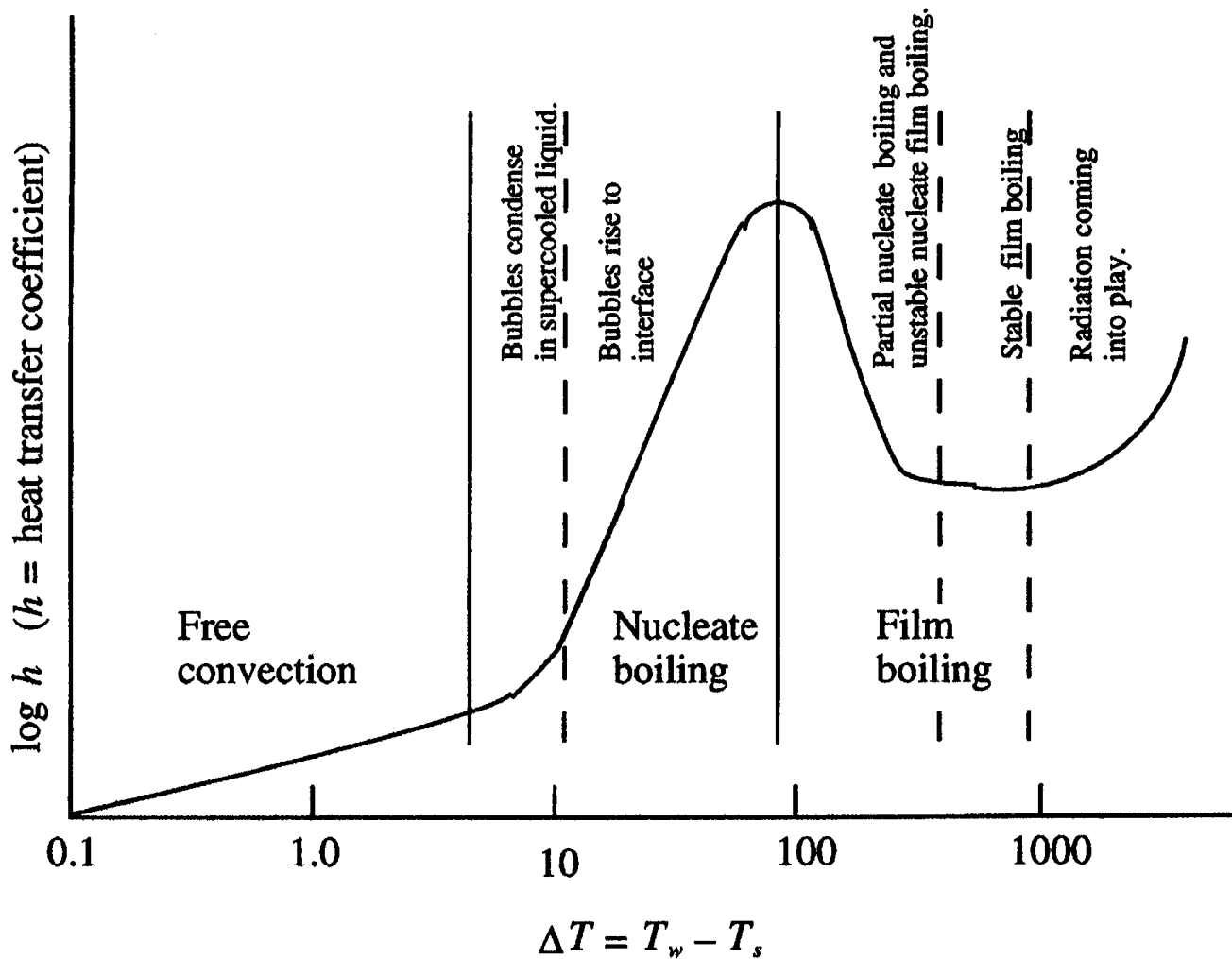
## Pool Boiling

Boiling occurs when a liquid changes phase and becomes vapor due to the absorption of heat. When the heating surface is submerged in a quiescent pool of liquid, the boiling phenomenon is known as *pool boiling*. The boiling that might occur in a fluid that moves due to an externally imposed pressure gradient is called *forced convection boiling*.

In 1934 Nukiyama performed a systematic study of pool boiling by immersing an electric wire into a body of saturated water and heating it. He determined the heat flux and temperature from current and voltage measurements. [Figure 49.9](#) illustrates the characteristics of pool boiling for water at atmospheric pressure. It shows the heat transfer coefficient (or the heat flux) as a function of the wire and water saturation temperatures. This curve shows that there are three distinct regimes of pool boiling: the *free-convection* regime, nucleate boiling regime, and *film boiling* regime.



**Figure 49.9** Principal regimes in pool boiling of water at atmospheric pressure and saturation temperature.



Initially, heat transfer is by free convection. The temperature of the heater surface is a few degrees above the saturation temperature, and free convection is sufficient to remove heat from it. The heat transfer correlations, as expected, are of the form given by Eq. (49.28).

In the **nucleate boiling** regime, bubbles are formed on the surface of the heater. There are two distinct regions in this regime. In the first, bubbles are formed at certain favored sites but are dissipated in the liquid as soon as they are detached from the surface. In the second region, the bubble generation is high enough to sustain a continuous column. Thus, large heat fluxes may be obtained in this region. In the nucleate boiling regime the heat flux increases rapidly with increasing temperature difference until the peak heat flux is reached. The location of this peak heat flux is known as the *burnout point*, *departure from nucleate boiling* (DNB), or *critical heat flux*. After the peak heat flux is exceeded, an extremely large temperature difference is needed to realize the resulting heat flux, and such high temperature differences may cause "burnout" of the

heating element.

As indicated in [Fig. 49.9](#), the peak heat flux occurs in the nucleation boiling regime. This maximum value must be known beforehand because of burnout considerations. That is, if the applied heat flux is greater than the peak heat flux, the transition takes place from the nucleate to the stable film boiling regime, in which (depending on the kind of fluid) boiling may occur at temperature differences well above the melting point of the heating surface.

After the peak heat flux is reached, the *unstable film boiling region* begins. No correlations are available for the heat flux in this region until the minimum point in the boiling curve is reached and the stable film boiling region starts. In the stable film boiling region the heating surface is separated from the liquid by a vapor layer across which heat must be transferred. Since vapors have low thermal conductivities, large temperature differences are needed for heat transfer in this region. Therefore, heat transfer in this regime is generally avoided when high temperatures are involved.

## Defining Terms

**Blackbody:** A body that absorbs all incident radiation from all directions at all wavelengths without reflecting, transmitting, or scattering it.

**Diffuse:** The radiation from a body or a radiative property that is independent of direction.

**Energy equation:** The equation that makes use of the principle of conservation of energy in a given process [see Eqs. (49.3)–(49.5)].

**Film temperature:** The mean boundary layer temperature [see Eq. (49.24)].

**Fourier's law:** The law that relates the heat flow to the temperature gradient in a body [see Eq. (49.2)].

**Gray surface:** A surface whose monochromatic emissivity is independent of the wavelength.

**Heat transfer coefficient:** A quantity that determines the heat flux due to convective heat transfer between a surface and a fluid that is moving over it [see Eqs. (49.12) and (49.26)].

**Newton's law of cooling:** The relation that determines the heat flux due to convection between a surface and moving fluid [see Eq. (2.10)].

**Nucleate boiling:** The stage in (pool) boiling of a fluid at which bubbles are formed at favored sites on the heating surface and detached from it.

**Participating medium:** A medium that absorbs, emits, and scatters the radiation that passes through it.

**Phonons:** A term that refers to the quantized lattice vibrations in a solid. Phonons are the energy carriers in a dielectric solid and are responsible for conduction of heat and electricity.

**Planck's law:** The law that relates the emissive power of a blackbody at a given wavelength to its temperature [see Eq. (49.29)].

**Promoters:** Substances that are introduced into a pure vapor or onto a condensing surface to facilitate dropwise condensation on a surface.

**Stefan-Boltzmann law:** The relation that determines the heat flux due to radiation from a surface [see Eq. (49.30)].

**Thermal conductivity:** A property that determines the ability of a substance to allow the flow of heat through it by conduction [see Eq. (49.2)].

**Thermal diffusivity:** A material property that governs the rate of propagation of heat in transient processes of conduction [see Eq. (49.6)].

## References

- Bayazitoglu, Y. and Ozisik, M. N. 1988. *Elements of Heat Transfer*. McGraw-Hill, New York.
- Bejan, A. 1984. *Convection Heat Transfer*. John Wiley & Sons, New York.
- Carslaw, H. S. and Jaeger, J. C. 1959. *Conduction of Heat in Solids*. Clarendon, Oxford.
- Cebeci, T. and Bradshaw, P. 1984. *Physical and Computational Aspects of Convective Heat Transfer*. Springer-Verlag, New York.
- Chapman, A. J. 1987. *Heat Transfer*. Macmillan, New York.
- Kreith, F. and Bohn, M. S. 1986. *Principles of Heat Transfer*. Harper & Row, New York.
- Modest, M. 1993. *Radiative Heat Transfer*. McGraw-Hill, New York.
- Ozisik, M. N. 1993. *Heat Conduction*, 2nd ed. John Wiley & Sons, New York.
- Siegel, R. and Howell, J. R. 1981. *Thermal Radiation Heat Transfer*. Hemisphere, New York.
- Ziman, J. M. 1960. *Electrons and Phonons*. Oxford University Press, London.

## Further Information

- ASME Journal of Heat Transfer*. Published quarterly by the American Society of Mechanical Engineers.
- International Journal of Heat and Mass Transfer*. Published monthly by Pergamon Press, Oxford; contains articles in various languages with summaries in English, French, German, or Russian.
- AIAA Journal of Thermophysics and Heat Transfer*. Published quarterly by the American Institute of Aeronautics and Astronautics.

Ohadi, M. M. "Heat Exchangers"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

**50.1 Heat Exchanger Types**

Transfer Processes • Number of Fluids • The Degree of Surface Compactness • Construction Features • Flow Arrangements • Heat Transfer Regimes

**50.2 Shell-and-Tube Heat Exchangers**

General Features • Major Components

**50.3 Compact Heat Exchangers**

Definition • Characteristics • Applications

**50.4 Design of Heat Exchangers**

Basic Design • Thermal-hydraulic Design • The LMTD Method • Effectiveness and Number of Transfer Units (NTU) Method

**M. M. Ohadi**

*University of Maryland*

A **heat exchanger** is a device that is used to transfer heat between two or more fluids that are at different temperatures. Heat exchangers are essential elements in a wide range of systems, including the human body, automobiles, computers, power plants, and comfort heating/cooling equipment. In the chemical and process industries, heat exchangers are used to concentrate, sterilize, distill, pasteurize, fractionate, crystallize, or control the fluid flow and chemical reaction rates. With the recent global promotion of energy efficiency and protection of the environment, the role of heat exchangers in efficient utilization of energy has become increasingly important, particularly for energy intensive industries, such as electric power generation, petrochemical, air conditioning/refrigeration, cryogenics, food, and manufacturing.

The Carnot efficiency for an ideal heat engine that receives  $Q_H$  amount of heat from a high-temperature reservoir at temperature  $T_H$  and rejects  $Q_L$  amount of heat to a reservoir at temperature  $T_L$  is

$$\eta = 1 - \frac{T_L}{T_H} = 1 - \frac{Q_L}{Q_H} \quad (50.1)$$

Equation (50.1) represents the maximum possible efficiency for a heat engine operating between a low- and a high-temperature reservoir. The Carnot efficiency is a good reference against which the performance of an actual heat engine can be compared. From the Carnot efficiency, it is clear that heat exchangers play a direct role in the overall efficiency of thermal machinery and equipment. By employing a more effective heat exchanger in a thermal cycle, one can get a higher

$T_H$  and lower  $T_L$ , resulting in higher efficiencies for the cycle as a whole.

In this chapter, we will first outline the basic classifications of heat exchangers and the relevant terminology used. Next, essential features and fundamental design aspects of **shell-and-tube** and **compact heat exchangers** will be described. Recent advances in thermal performance of heat exchangers will be discussed in the last section of this chapter.

## 50.1 Heat Exchanger Types

---

Heat exchangers can be classified in many different ways. Shah [1981] provides a comprehensive and detailed description of the various categories, the associated terminology, and the practical applications of each type. As indicated there, heat exchangers can be broadly classified according to the six categories described below.

### Transfer Processes

In this category, heat exchangers are classified into direct contact and indirect contact types. In the direct type, the heat exchanging streams (hot and cold streams) come into direct contact and exchange heat before they are separated. Such heat exchangers are more common in applications where both heat and mass transfer are present. A familiar example is evaporative cooling towers used in power plants and in comfort cooling for desert environments. In an indirect contact heat exchanger, the hot and cold fluids are separated by a solid, impervious wall representing the heat transfer surface. The conventional shell-and-tube heat exchangers fall in this category.

### Number of Fluids

In many applications, the two-fluid heat exchangers are the most common type. However, in certain applications, such as cryogenics and the chemical and process industries, multifluid heat exchangers are also common.

### The Degree of Surface Compactness

In this category, heat exchangers are classified according to the amount of heat transfer surface area per unit volume incorporated into the heat exchanger. This is represented by the heat transfer area-to-volume ratio ( $\beta = A/V$ ). For example, for compact heat exchangers  $\beta \approx 700 \text{ m}^2/\text{m}^3$  or greater, for **laminar flow heat exchangers**  $\beta \approx 3000 \text{ m}^2/\text{m}^3$ , and for **micro heat exchangers**  $\beta \approx 10\,000$  or greater.

### Construction Features

In this category, heat exchangers are often divided into four major types: tubular, plate type, extended surface, and regenerative types. There are other types of heat exchangers with unique construction that may not fall in these four categories. However, they are not commonly used and

are specific to specialized applications.

## **Flow Arrangements**

Heat exchangers can be classified according to the manner in which the fluid flows through the tube and shell sides. The two broad types are single pass and multipass. When the fluid flows through the full length of the heat exchanger without any turns, it is considered to have made one pass. A heat exchanger is considered single pass when the hot and cold fluids make one pass in the heat exchanger. Within the single pass category, the common flow arrangements for the hot and cold fluids are parallel flow, counterflow, cross-flow, split flow, and divided flow. The multipass heat exchangers are classified according to the type of construction. The common types include extended surface, shell-and-tube, and plate heat exchangers.

## **Heat Transfer Regimes**

Exchange of thermal energy in a heat exchanger from the hot fluid to the exchanging wall to the cold fluid can employ one or more modes of heat transfer. For example, in gas-to-gas heat exchangers, heat transfer on both the hot and cold fluids is by single phase convection. In the water-cooled steam condensers, single phase convection takes place in one side and condensation in the other side. Yet, in certain cryogenic heat exchangers, condensation takes place on one side and evaporation on the other side. Basic components and essential features of the two most commonly used heat exchangers—namely, shell-and-tube and plate-fin heat exchangers—will be discussed in the following sections.

## **50.2 Shell-and-Tube Heat Exchangers**

---

### **General Features**

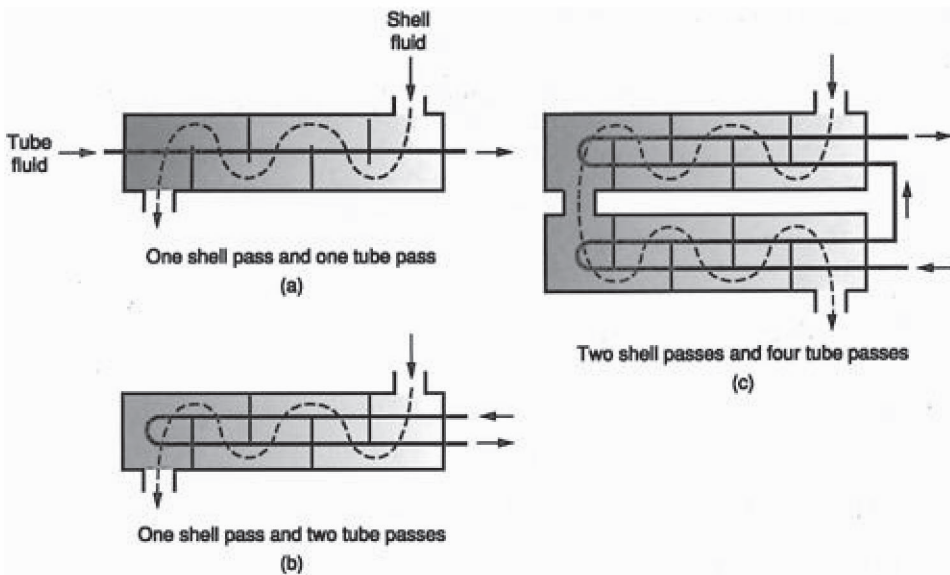
Among the various heat exchanger types, the shell-and-tube heat exchangers are the most commonly used in a number of industries, including process, chemical, power, and refrigerating/air conditioning. In this type of heat exchanger, one fluid flows inside the tubes while the other fluid is forced through the shell side and over the tubes in a cross-flow arrangement.

The shell-and-tube heat exchangers have several main advantages that, as a whole, have contributed to their nearly universal acceptance for a wide range of applications. They can be custom designed for almost any capacity, working fluid type, operating pressure, and temperature conditions. The type of material used to construct the heat exchanger can be any of the most commonly used materials. Shell-and-tube heat exchangers yield a relatively large surface area density. Methods for cleaning and other maintenance, such as periodic replacement of gaskets and tubes, are well established and easily done. Their broad worldwide use over the years has resulted in good methods for design and fabrication.

The size of a shell-and-tube heat exchanger can vary over a wide range from compact to supergiant configurations. Common examples of shell-and-tube heat exchangers are steam

generators, condensers, evaporators, feed water heaters, oil coolers in the power and refrigeration industries, and process heat exchangers in the petroleum refining and chemical industries. Classification of shell-and-tube heat exchangers is usually according to the number of tube and shell passes. The heat exchangers schematically shown in Figs. 50.1(a), 50.1(b), and 50.1(c) represent, respectively, one shell pass and one tube pass, two shell passes and one tube pass, and one shell and four tube passes.

**Figure 50.1** Examples of shell-and-tube heat exchanger configurations.

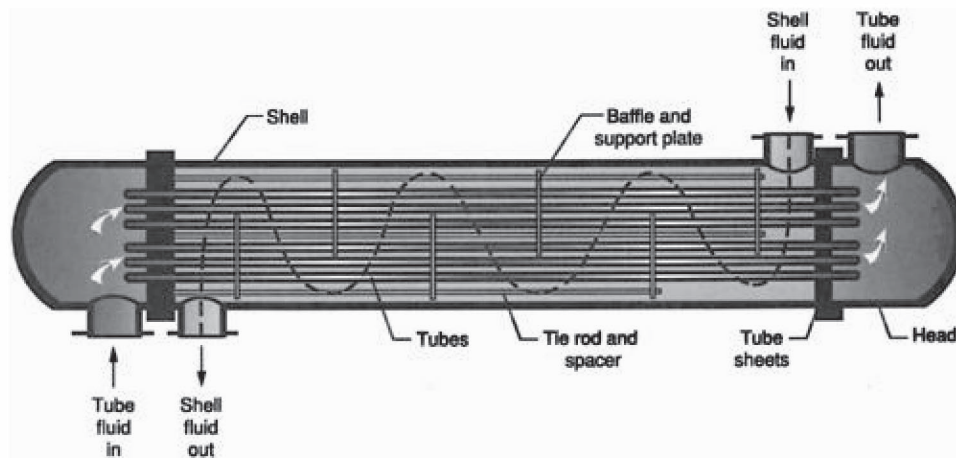


## Major Components

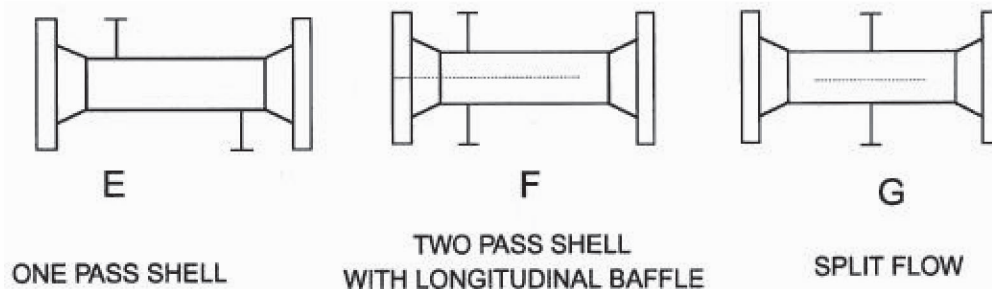
Major components of a shell-and-tube heat exchanger are shown in Fig. 50.2. A variety of different constructions are used in the shell-and-tube heat exchangers. Major types are identifiable by a special notation developed by the Tubular Exchanger Manufacturer's Association (TEMA) in which each heat exchanger is designated by a three-letter combination. The first letter indicates the front-end head type, the second the shell type, and the third letter identifies the rear-end head type. The standard in single shell arrangement type is TEMA E, in which the entry and outlet nozzles are placed at opposite ends. Figure 50.3 shows three sample TEMA configurations.



**Figure 50.2** Major components of a shell-and-tube heat exchanger.



**Figure 50.3** Sample TEMA configurations.



The baffles in a shell-and-tube heat exchanger serve two tasks. First, they direct the flow in the shell side tube bundle approximately at right angles to the tubes so that higher heat transfer coefficients in the shell side can be obtained while reducing thermal stresses on the tubes due to the cross-flow effect. The second, and more important, function of baffles is to provide additional support for the tubes during assembly and operation and minimize flow-induced vibration of the tubes.

The tubes used in a shell-and-tube heat exchanger can be either bare (plain) or of the enhanced type. The enhanced types utilize a combination of additional surface area (e.g., through use of various fin structures) and various mechanisms to increase the heat transfer coefficients at the tube surface. As described in Ohadi [1991] and Webb [1994], some of the recently developed enhanced heat transfer mechanisms can yield in excess of a tenfold increase in the magnitude of the heat transfer coefficients when compared to conventional plain tubes.

The tube sheet in a shell-and-tube heat exchanger is usually a single, round, metallic plate that has been suitably drilled and grooved to accommodate the tubes, the gasket, and other associated accessories. The main purpose of the tube sheet is to support the tubes at the ends. In addition to its mechanical requirements, the tube sheet must be resistant to corrosion from fluids in both the tube and shell sides and must be electrochemically compatible with the tube side materials. The tube sheets are sometimes made from low-carbon steel with a thin layer of corrosion resistant alloy.

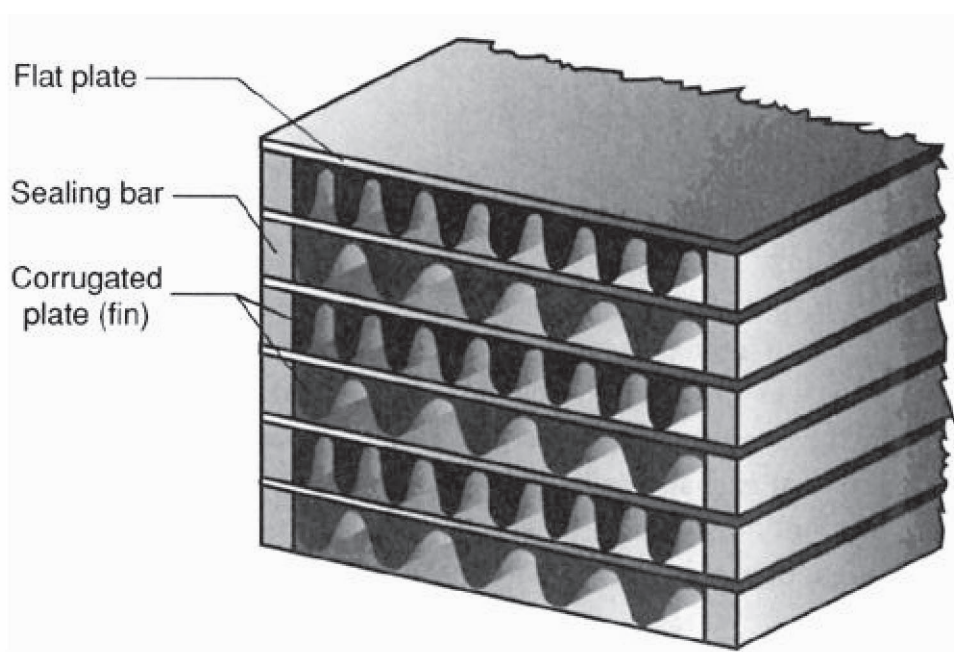
## 50.3 Compact Heat Exchangers

---

### Definition

A heat exchanger is referred to as a compact heat exchanger if it incorporates a heat transfer surface having a surface density above approximately  $700 \text{ m}^2/\text{m}^3$  on at least one of the fluid sides, usually the gas side. Compact heat exchangers are generally of plate-fin type, tube-fin type, tube bundles with small diameter tubes, and regenerative type. In a plate-fin exchanger, corrugated fins are sandwiched between parallel plates as shown in Fig. 50.4. In a tube-fin exchanger, round and rectangular tubes are most commonly used and fins are employed either on the outside, the inside, or on both the outside and inside of the tubes, depending upon the application. Basic flow arrangements of two fluids are single pass cross-flow, counterflow, and multipass cross-counterflow.

**Figure 50.4** A basic plate-fin heat exchanger construction.



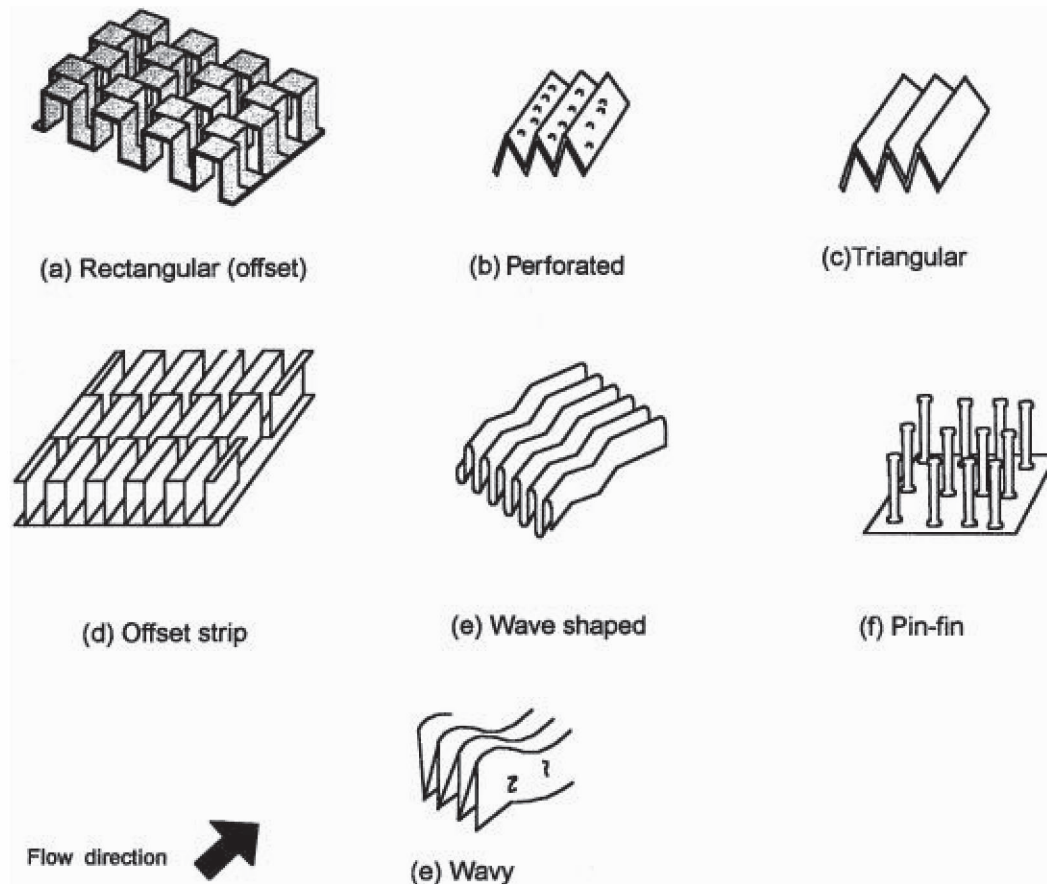
### Characteristics

Some of the subtle characteristics and uniqueness of compact heat exchangers are (1) usually at least one of the fluids is a gas, (2) the fluid must be clean and relatively noncorrosive, (3) fluid pumping power is always of prime importance, and (4) operating pressure and temperatures are somewhat limited compared to shell-and-tube exchangers due to construction features.

The shape of a compact heat exchanger is usually distinguished by having a large frontal area

and short flow lengths. Thus, the proper design of the header for compact heat exchangers is very important for a uniform flow distribution. A variety of surfaces are available for use in compact heat exchangers having different orders of magnitude of surface area densities. Such surfaces could introduce substantial cost, weight, or volume savings as desired by the design. Figure 50.5 shows some sample surfaces and geometries currently employed in compact heat exchangers.

**Figure 50.5** Examples of plate-fin surface geometries.



## Applications

The major heat exchangers in cars and trucks are radiators for engine cooling, heaters for passenger compartment heating, evaporators and condensers for air conditioning, oil coolers for engine oil and transmission oil cooling, and charge air coolers and intercoolers for cooling charge air for increased mass flow rate through the engine.

A variety of compact heat exchangers are needed in space applications where the minimum weight and volume, as well as the absolute reliability and durability, are essential. Very high heat fluxes in low-weight and low-volume heat exchangers can be obtained by employing two-phase flow in offset strip fins. Such heat exchangers will have a significant impact on emerging

applications, such as condensers and evaporators for thermal control of space stations, satellite instrumentation, aircraft avionics packages, thermal control of electronics packages, and thermal control of space life support and astronauts.

In commercial aircraft applications, a variety of primarily compact heat exchangers are used for cabin air conditioning and heating, avionics cooling, deicing, and oil cooling applications.

## 50.4 Design of Heat Exchangers

---

### Basic Design

When designing a heat exchanger, the following requirements must be satisfied for most typical applications:

- Meet the process thermal requirements within the allowable pressure drop penalty

- Withstand the operational environment of the plant, including the mechanical and thermal stress conditions of the process, operational schedule, and resistance to corrosion, erosion, vibration and fouling

- Provide easy access to those components of the heat exchanger that are subject to harsh environments and require periodic repair and replacement

- Should have a reasonable cost (combination of initial and maintenance costs)

- Should be versatile enough to accommodate other applications within the plant, if possible

Broadly speaking, the various design tasks in a new heat exchanger can be classified into two categories: (1) the mechanical design and (2) the thermal-hydraulic design. The mechanical design deals with the proper distribution of thermal and mechanical stresses and the general integrity of the heat exchanger for the conditions specified. The thermal-hydraulic design deals with calculations of heat transfer and pressure drop coefficients and, subsequently, determination of the overall required dimensions of the heat exchanger for specified thermal operating conditions. The mechanical design aspects are beyond the scope of this article and can be found elsewhere [Hewitt *et al.*, 1994]. An introduction to the thermal-hydraulic design criteria will be given in the following.

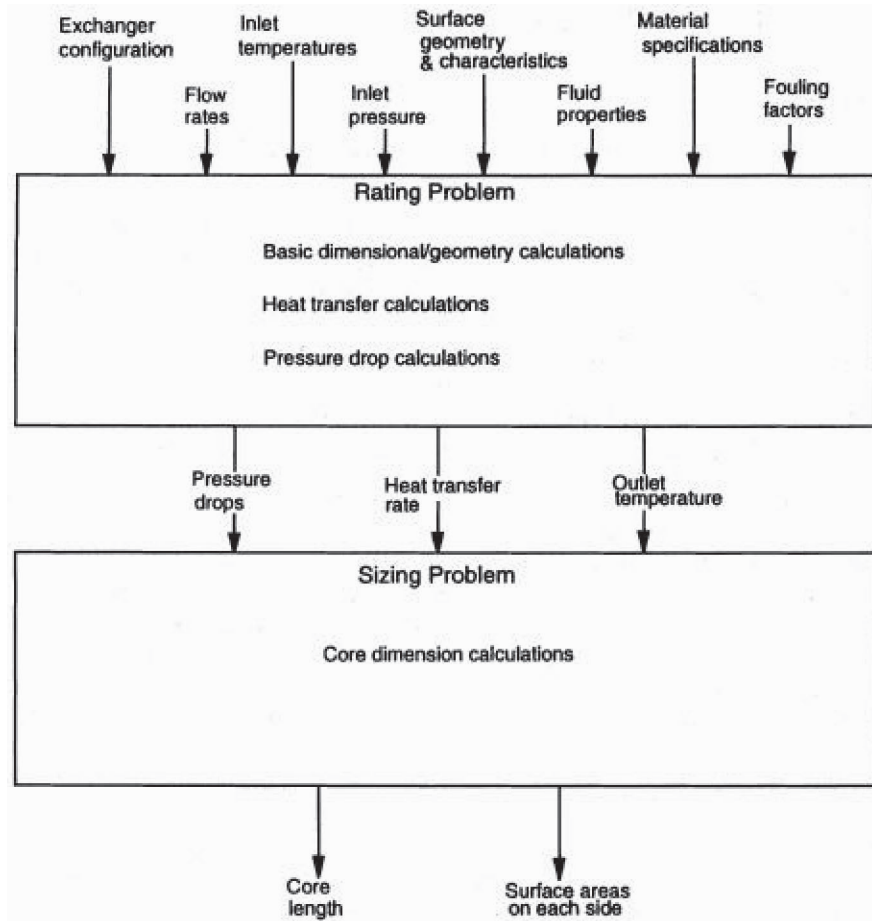
### Thermal-hydraulic Design

As discussed by Shah [1986], the various steps in the thermal-hydraulic design of heat exchangers can be classified into two distinct problems: *the rating problem* (also referred to as *the performance problem*) and *the sizing problem* (also referred to as *the design problem*).

The rating problem involves the determination of the overall heat transfer coefficients and pressure drop characteristics utilizing the following input data: the heat exchanger configuration, flow rates on each side, inlet temperatures and pressure data, thermophysical properties of the fluid fouling factors, and complete details on the materials and surface geometries of each side. The sizing problem utilizes the data provided in the rating problem and determines the overall (core)

dimensions (core length and surface areas) for the heat exchanger. Steps involved in the rating and sizing problems are schematically shown in Fig. 50.6.

**Figure 50.6** Schematic presentation of the heat exchanger rating and sizing problems.



## The LMTD Method

It is a general practice to determine the heat transfer between two fluids in a heat exchanger using the mean quantities and the following defining equation:

$$q = FU_m A \Delta T_m \quad (50.2)$$

in which  $U_m$  and  $\Delta T_m$  represent the mean heat transfer coefficient and effective temperature difference, respectively. The factor  $F$  is a function of flow arrangement and is provided graphically for various heat exchanger configurations [Shah, 1983]. The required surface area in the heat exchanger is then determined utilizing the following equation:

$$A = \frac{q}{FU_m \Delta T_m} \quad (50.3)$$

The above equation is based on several assumptions, such as constant specific heats, constant overall heat transfer coefficients in the heat exchanger, constant fluid properties, absence of heat losses to and from the surroundings, and absence of heat sources in the heat exchanger. The probable errors due to the above assumptions are regarded as the penalty for the simplicity of the method. In practice, correction factors are applied to adjust the calculations for the errors involved.

The overall heat transfer coefficient  $U_m$  is determined by taking into account the various thermal resistances involved in the heat transfer path between the hot and cold fluids:

$$U_m = \frac{1}{\sum R} = \frac{1}{(1/hA)_h + R_{fh} + R_w + R_{fc} + (1/hA)_c} \quad (50.4)$$

in which  $(1/hA)_h$  and  $(1/hA)_c$  respectively represent the convective resistance on the hot and cold sides of the heat exchanger,  $R_{fh}$  and  $R_{fc}$  represent fouling resistance on the hot and cold sides, and  $R_w$  is thermal resistance due to wall tubing. Representative values of  $U$  for common fluids are listed in [Table 50.1](#).

**Table 50.1** Representative Values of the Overall Heat Transfer Coefficient ( $U$ )

Hot Side Fluid	Cold Side Fluid	$U$	
		Btu/(hr-ft <sup>2</sup> -°F)	W/(m <sup>2</sup> K)
Water	Water	200–500	800–2500
Water	Gas	2–10	10–50
Water	Lubricating oil	20–80	160–400
Ammonia	Water (e.g., water-cooled condenser)	150–500	800–2588
Water	Brine	100–200	500–1000
Light organics	Water	75–150	350–750
Medium organics	Water	50–125	240–610
Heavy organics	Water	5–75	25–370
Light organics	Light organics	40–100	200–500
Heavy organics	Heavy organics	10–40	50–200
Heavy organics	Light organics	30–60	150–300
Steam	Water	200–1000	10–6000
Steam	Ammonia	200–700	1000–3500
Freon 12	Water	50–200	300–1000
Steam	Heavy fuel oil	10–40	50–200
Steam	Light fuel oil	30–60	200–400

Steam	Air (air-cooled condenser)	10–40	50–200
Finned tube heat exchanger, water in tubes	Air over finned tubes	5–10	30–60
Finned tube heat exchanger, steam in tubes	Air over tubes	50–700	300–4500

---

To determine the mean temperature difference between the hot and cold fluids over the heat exchanger length ( $\Delta T_m$ ), it is a common practice to use the log mean temperature difference (LMTD), defined as

$$\Delta T_{lm} = \frac{\Delta T_i - \Delta T_o}{\ln(\Delta T_i / \Delta T_o)} \quad (50.5)$$

in which  $\Delta T_i$  and  $\Delta T_o$  represent the temperature difference between the hot and cold fluids at the inlet and outlet of the heat exchanger, respectively.

## Effectiveness and Number of Transfer Units (NTU) Method

The LMTD method for the prediction of heat exchanger performance is useful only when the inlet and exit temperatures for the heat exchanger are known, either because they have been specified by the design or because they have been measured in a test. For situations where calculation of the inlet and outlet temperatures and flow rates are desired, use of the LMTD method requires iterative solution procedures. This can be avoided if the heat exchanger performance is expressed in terms of effectiveness and number of transfer units (NTU), which will be briefly discussed in the following. The heat exchanger effectiveness,  $\varepsilon$ , is defined as the ratio of actual heat transferred to the maximum heat transfer amount which can be transferred in an infinitely long counterflow heat exchanger. In an infinitely long counterflow heat exchanger, with  $(mC_p)_C < (mC_p)_h$ , in which subscripts  $c$  and  $h$  refer to the cold and hot streams, we can write

$$\dot{q}_{\max} = (\dot{m}C_p)_C (T_{h,\text{in}} - T_{c,\text{in}}) \quad (50.6)$$

Similarly, if  $(mC_p)_h < (mC_p)_c$  then  $\dot{q}_{\max}$  in an infinitely long counterflow exchanger is

$$\dot{q}_{\max} = (\dot{m}C_p)_h (T_{h,\text{in}} - T_{c,\text{in}}) \quad (50.7)$$

Now if we write

$$C_{\min} = \min [(\dot{m}c_p)_h, (\dot{m}c_p)_c] \quad (50.8)$$

then the maximum heat transfer in a heat exchanger of any configuration is

$$\dot{q}_{\max} = C_{\min} (T_{h,\text{in}} - T_{c,\text{in}}) \quad (50.9)$$



From this equation, the effectiveness is expressed as

$$\varepsilon = \frac{\dot{q}}{\dot{q}_{\max}} = \frac{(\dot{m}c_p)_h(T_{h,\text{in}} - T_{h,\text{out}})}{C_{\min}(T_{h,\text{in}} - T_{c,\text{in}})} = \frac{(\dot{m}c_p)_c(T_{c,\text{out}} - T_{c,\text{in}})}{C_{\min}(T_{h,\text{in}} - T_{c,\text{in}})} \quad (50.10)$$

The number of transfer units (NTU) is defined for the hot and cold streams as

$$\text{NTU}_h = \frac{AU}{(\dot{m}c_p)_h} \quad (50.11)$$

$$\text{NTU}_c = \frac{AU}{(\dot{m}c_p)_c} \quad (50.12)$$

where  $A$  is the total heat exchanger area and  $U$  is the overall heat transfer coefficient. Similarly,  $\text{NTU}_{\min}$  corresponding to the stream having the minimum  $(\dot{m}c_p)$  is defined as

$$\text{NTU}_{\min} = \frac{AU}{(\dot{m}c_p)_{\min}} \quad (50.13)$$

From there it follows that

$$\varepsilon = \text{NTU}_{\min} \theta \quad (50.14)$$

where  $\theta$  is the ratio between the mean temperature difference  $\Delta T_{\text{mean}}$  and the maximum temperature difference  $\Delta T_{\max}$  :

$$\theta = \frac{\Delta T_{\text{mean}}}{\Delta T_{\max}} \quad (50.15)$$

The effectiveness and NTU relations for various heat exchanger configurations have been developed and are available in literature [Mills, 1992].

## Defining Terms

**Compact heat exchanger:** A heat exchanger that incorporates a high surface area-to-volume density (usually  $700 \text{ m}^2/\text{m}^3$  or higher).

**Heat exchanger:** A device in which heat transfer takes place between two or more fluids that are at different temperatures.

**Laminar flow heat exchanger:** A heat exchanger whose surface area-to-volume density is in the neighborhood of  $3000 \text{ m}^2/\text{m}^3$ .

**LMTD method:** Represents the mean logarithmic temperature difference between the hot and cold fluids in the entire heat exchanger.

**Micro heat exchanger:** A heat exchanger whose surface area-to-volume density is much higher



than compact and laminar flow heat exchangers ( $10\,000\text{ m}^2/\text{m}^3$  or higher).

**NTU method:** Useful in design of heat exchanger when the inlet and outlet fluid temperatures are not known.

**Shell-and-tube heat exchanger:** A heat exchanger in which one fluid flows inside a set of tubes while the other fluid is forced through the shell and over the outside of the tubes in a cross-flow arrangement.

## References

- Guyer, E. C. (Ed.) 1994. *Handbook of Applied Thermal Design*. McGraw-Hill, New York.
- Hewitt, G. F., Shires, G. L., and Bott, T. R. (Eds.) 1994. *Process Heat Transfer*. CRC Press, Boca Raton, FL.
- Kakac, S., Bergles, A. E., and Fernandes, E. O. (Eds.) 1988. *Two-Phase Flow Heat Exchanger: Thermal Hydraulic Fundamentals and Design*. Kluwer Academic, Dordrecht, The Netherlands.
- Mills, A. F. 1992. *Heat Transfer*. Irwin, Homewood, IL.
- Ohadi, M. M. 1991. Electrodynamic enhancement of single-phase and phase-change heat transfer in heat exchangers. *ASHRAE J.* 33(12):42–48.
- Palen, J. W., (Ed.) 1986. *Heat Exchanger Sourcebook*. Hemisphere, New York.
- Rohsenow, W. M., Harnett, J. P., and Ganic, E. N. (Eds.) 1985. *Handbook of Heat Transfer Applications*, 2nd ed. McGraw-Hill, New York.
- Taborek, J., Hewitt, G. F., and Afgan, N. (Eds.) 1983. *Heat Exchanger: Theory and Practice*. Hemisphere/McGraw-Hill, Washington, DC.
- Webb, R. L. 1994. *Principles of Enhanced Heat Transfer*. John Wiley & Sons, New York.

## Further Information

For in-depth treatment of heat exchangers, see the following texts:

- D. Chisholm, Editor, *Heat Exchanger Technology*, Elsevier Applied Science, New York (1988).
- A. P. Fraas, *Heat Exchanger Design*, 2nd ed., Wiley, New York (1989).
- J. P. Gupta, *Fundamentals of Heat Exchanger and Pressure Vessel Technology*, Hemisphere, Washington, DC (1986); reprinted as *Working with Heat Exchangers*, Hemisphere, Washington, DC (1990).
- G. F. Hewitt, Coordinating Editor, *Hemisphere Handbook of Heat Exchanger Design*, Hemisphere, New York (1989).
- S. Kakac, R. K. Shah, and W. Aung, Editors, *Handbook of Single-Phase Convective Heat Transfer*, Wiley, New York (1987).
- S. Yokell, *A Working Guide to Shell-and-Tube Heat Exchangers*, McGraw-Hill, New York (1990).
- R. K. Shah, A. D. Kraus, and D. Metzger, Editors, *Compact Heat Exchangers—A Festschrift for A.L. London*, Hemisphere, Washington, DC (1990).
- T. R. Bott, *Fouling Notebook: A Practical Guide to Minimizing Fouling in Heat*

- Exchangers*, Association of Chemical Engineers, London (1990).
- S. Kakac, Editor, *Boilers, Evaporators, and Condensers*, Wiley, New York (1991).
- Y. Mori, A. E. Sheindlin, and N. H. Afgan, Editors, *High Temperature Heat Exchanger*, Hemisphere, Washington, DC (1986).
- S. Kakac, R. K. Shah, and A. E. Bergles, Editors, *Low Reynolds Number Flow Heat Exchanger*, Hemisphere, Washington, DC (1983).

Hecklinger, R. S. "Combustion"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

### 51.1 Fundamentals of Combustion

### 51.2 Combustion Calculations

#### **Roger S. Hecklinger**

*Roy F. Weston, Inc.*

One of the signal accomplishments of mankind was development of the ability to control combustion. The ability to control combustion not only added to creature comforts, but also provided the means to manufacture ceramics and to smelt metals. Then as now, there is an art to controlling combustion. Over the last 150 years or so, natural laws have been identified and combustion constants have been established that allow us to calculate and predict performance of combustion processes.

**Combustion** is the rapid oxidation of **combustible substances** with release of heat. Oxygen is the sole supporter of combustion. Carbon and hydrogen are by far the most important of the combustible substances. These two elements occur either in a free or combined state in all fuels—solid, liquid, and gaseous. Sulfur is the only other element considered to be combustible. In normal combustion applications, sulfur is a minor constituent with regard to heating value; however, it may be a major concern in design of the combustion furnace, heat transfer surface, and air pollution control equipment that are required to harness the combustion process.

While oxygen is the sole supporter of combustion, the only source of oxygen considered here will be the oxygen in the air around us.

## **51.1 Fundamentals of Combustion**

---

**Table 51.1** displays the elements and compounds that play a part in all of the commonly used combustion processes. The elemental and molecular weights displayed are approximate values that are sufficient for combustion calculations. (For example, the molecular weight of oxygen,  $O_2$ , is listed as 32.0, whereas its molecular weight has been defined more precisely as 31.9988.) Nitrogen is listed as chemical nitrogen,  $N_2$ , with a molecular weight of 28.0, and as atmospheric nitrogen,  $N_{2atm}$ , which is a calculated figure to account for trace constituents of dry air. Water occurs as a vapor in air and in the products of combustion and as a liquid or vapor constituent of some fuels.

**Table 51.1** Elements and Compounds Encountered in Combustion

Substance	Molecular Symbol	Molecular Weight	Form	Density lb per ft <sup>3</sup>
Carbon	C	12.0	Solid	—
Hydrogen	H <sub>2</sub>	2.0	Gas	0.005 3
Sulfur	S	32.1	Solid	—
Carbon monoxide	CO	28.0	Gas	0.078 0
Methane	CH <sub>4</sub>	16.0	Gas	0.042 4
Acetylene	C <sub>2</sub> H <sub>2</sub>	26.0	Gas	0.069 7
Ethylene	C <sub>2</sub> H <sub>4</sub>	28.0	Gas	0.074 6
Ethane	C <sub>2</sub> H <sub>6</sub>	30.1	Gas	0.080 3
Oxygen	O <sub>2</sub>	32.0	Gas	0.084 6
Nitrogen	N <sub>2</sub>	28.0	Gas	0.074 4
Nitrogen-atmos.	N <sub>2atm</sub>	28.2	Gas	0.074 8
Dry air		29.0	Gas	0.076 6
Carbon dioxide	CO <sub>2</sub>	44.0	Gas	0.117 0
Water	H <sub>2</sub> O	18.0	Gas/liquid	0.047 6
Sulfur dioxide	SO <sub>2</sub>	64.1	Gas	0.173 3

A U.S. standard atmosphere of dry air has been defined as a mechanical mixture of 20.947% O<sub>2</sub>, 78.086% N<sub>2</sub>, 0.934% Ar (argon), and 0.033% CO<sub>2</sub> by volume [CRC, 1990]. The percentages of argon and carbon dioxide in air can be combined with chemical nitrogen to develop the following compositions of dry air by volume and by weight that can be used for combustion calculations:

Constituent	% by Volume	% by Weight
Oxygen, O <sub>2</sub>	20.95	23.14
Atmospheric nitrogen, N <sub>2atm</sub>	79.05	76.86

Atmospheric air also contains some water vapor. The level of water vapor in air, or its humidity, is a function of atmospheric conditions. It is measured by wet and dry bulb thermometer readings and a psychrometric chart. If specific data are not known, the American Boiler Manufacturers Association recommends a standard of 0.13 pounds of water per pound of dry air, which corresponds to 60% relative humidity and a dry bulb temperature of 80° F.

Table 51.2 displays the basic chemical reactions of combustion. These reactions result in complete combustion; that is, the elements and compounds unite with all the oxygen with which they are capable of entering into combination. In actuality, combustion is a more complex process in which heat in the combustion chamber causes intermediate reactions leading up to complete combustion. An example of intermediate steps to complete combustion include carbon reaction with oxygen to form carbon monoxide and, later in the combustion process, carbon monoxide reaction with more oxygen to form carbon dioxide. The combined reaction produces precisely the same result as if an atom of carbon combined with a molecule of oxygen to form a molecule of carbon dioxide in the initial reaction. An effectively controlled combustion process results in well less than 0.1% of the carbon in the fuel leaving the combustion chamber as carbon monoxide, with

the remaining 99.9+% of the carbon in the fuel leaving the combustion process as carbon dioxide. It should also be noted with regard to [Table 51.2](#) that some of the sulfur in a fuel may combust to SO<sub>3</sub> rather than SO<sub>2</sub>, with a markedly higher release of heat. However, it is known that only a small portion of the sulfur will combust to SO<sub>3</sub>, and some of the sulfur in fuel may be in the form of pyrites (FeS<sub>2</sub>), which do not combust at all. Therefore, only the SO<sub>2</sub> reaction is given.

**Table 51.2** Chemical Reactions of Combustion

Combustible	Reaction
Carbon	$C + O_2 = CO_2$
Hydrogen	$2H_2 + O_2 = 2H_2O$
Sulfur	$S + O_2 = SO_2$
Carbon monoxide	$2CO + O_2 = 2CO_2$
Methane	$CH_4 + 2O_2 = CO_2 + 2H_2O$
Acetylene	$2C_2H_2 + 5O_2 = 4CO_2 + 2H_2O$
Ethylene	$C_2H_4 + 3O_2 = 2CO_2 + 2H_2O$
Ethane	$2C_2H_6 + 7O_2 = 4CO_2 + 6H_2O$

The desired result of controlled combustion is to release heat for beneficial use. The heat released in combustion of basic combustible substances has been established as displayed in [Table 51.3](#). The heating value of a substance can be expressed either as higher (or gross) heating value or as lower (or net) heating value. The higher heating value takes into account the fact that water vapor formed or evaporated in the process of combustion includes latent heat of vaporization that could be recovered if the products of combustion are reduced in temperature sufficiently to condense the water vapor to liquid water. The lower heating value is predicated on the assumption that the latent heat of vaporization will not be recovered from the products of combustion.

**Table 51.3** Heat of Combustion

Combustible	Molecular Symbol	Heating Value			
		Btu per Pound		Btu per Cubic Foot	
		Gross	Net	Gross	Net
Carbon	C	14 100	14 100	—	—
Hydrogen	H <sub>2</sub>	61 100	51 600	325	275
Sulfur	S	3 980	3 980	—	—
Carbon monoxide	CO	4 350	4 350	322	322
Methane	CH <sub>4</sub>	23 900	21 500	1015	910
Acetylene	C <sub>2</sub> H <sub>2</sub>	21 500	20 800	1500	1450
Ethylene	C <sub>2</sub> H <sub>4</sub>	21 600	20 300	1615	1510
Ethane	C <sub>2</sub> H <sub>6</sub>	22 300	20 400	1790	1640

The usual method to determine the heating value of a combustible substance that is solid or

liquid is to burn it in a **bomb calorimeter** with a known quantity of oxygen and to measure the heat released. Since the bomb calorimeter is cooled to near ambient conditions, the heat recovery measured includes the latent heat of vaporization as the products of combustion are cooled and condensed in the bomb. That is, the bomb calorimeter inherently measures higher heating value (HHV). It has been customary in the U.S. to express heating value as HHV. In Europe and elsewhere, heating value is frequently expressed as the lower heating value (LHV). Heating value can be converted from HHV to LHV if weight decimal percentages of moisture and hydrogen (other than the hydrogen in moisture) in the fuel are known, using the following formula:

$$\text{LHV}_{\text{Btu/lb}} = \text{HHV}_{\text{Btu/lb}} - [\% \text{H}_2\text{O} + (9 \times \% \text{H}_2)] \times (1050 \text{ Btu/lb}) \quad (51.1)$$

The heating value of gaseous fuels is usually determined by measuring the volumetric percentage of the various combustible gaseous constituents and calculating the heating value using the known HHV or LHV for each of the constituents. [Table 51.3](#) displays heat of combustion for various combustible substances on a gross and a net basis for both Btu per pound and Btu per cubic foot at 60°F and 30 inches Hg. A Btu is a British thermal unit, which, in general terms, is the heat required to raise one pound of water one degree Fahrenheit. Test methods for determining heating value are prescribed by the American Society for Testing and Materials (ASTM).

ASTM also publishes methods for determining the **ultimate analysis** of solid and liquid fuels. Solid fuels include coal, lignite, peat, wood, agricultural wastes, and municipal solid waste. Liquid fuels include crude oil, refined oils, kerosene, gasoline, methyl alcohol, and ethyl alcohol. The ultimate analysis of a fuel is developed through measures of carbon, hydrogen, sulfur, nitrogen, ash, and moisture content. Oxygen is normally determined *by difference*; that is, once the percentages of the other components are measured, the remaining material is assumed to be oxygen. For solid fuels it is frequently desirable to determine the **proximate analysis** of the fuel. The procedure for determining the proximate analysis is prescribed by ASTM. The qualities of the fuel measured in percent by weight are moisture, volatile matter, fixed carbon, and ash. This provides an indication of combustion characteristics of a solid fuel. As a solid fuel is heated to combustion, first the moisture in the fuel evaporates, then some of the combustible constituents volatilize (gasify) and combust as a gas with oxygen, and the remaining combustible constituents remain as fixed carbon in a solid state and combust with oxygen to form carbon dioxide. The material remaining after the completion of combustion is the ash. Some fuels, such as anthracite coal and charcoal, have a high percentage of fixed carbon and a low percentage of volatiles and combust with little flame, whereas other fuels such as wood and municipal solid waste have a high percentage of volatiles and a low percentage of fixed carbon and combust with much flame.

[Table 51.4](#) displays ignition temperatures for combustible substances. The ignition temperature is the temperature to which the combustible substance must be raised before it will unite in chemical combustion with oxygen. Thus, the temperature must be reached and oxygen must be present for combustion to take place. Ignition temperatures are not fixed temperatures for a given substance. The actual ignition temperature is influenced by combustion chamber configuration,

oxygen fuel ratio, and the synergistic effect of multiple combustible substances. The ignition temperature of charcoal and the coals is the ignition temperature of their fixed carbon component. The volatile components of charcoal and coals are gasified but not ignited before the ignition temperature is attained.

**Table 51.4** Ignition Temperatures

Combustible	Molecular Symbol	Ignition Temperature, °F
Carbon (fixed)	C	
Charcoal		650
Bituminous coal		765
Anthracite		975
Hydrogen	H <sub>2</sub>	1080
Sulfur	S	470
Carbon monoxide	CO	1170
Methane	CH <sub>4</sub>	1270
Acetylene	C <sub>2</sub> H <sub>2</sub>	700
Ethylene	C <sub>2</sub> H <sub>4</sub>	960
Ethane	C <sub>2</sub> H <sub>6</sub>	1020

The oxygen, nitrogen, and air data displayed in [Table 51.5](#) represent the weight of air theoretically required to completely combust one pound or one cubic foot of a combustible substance. The weight of oxygen required is the ratio of molecular weight of oxygen to molecular weight of the combustion constituent, as displayed in [Table 51.2](#). The weight of nitrogen and air required are calculated from the percent-by-weight constituents of dry air. In actuality, to achieve complete combustion, air in excess of the theoretical requirement is required to increase the likelihood that all of the combustible substances are joined with sufficient oxygen to complete combustion. The level of excess air required in a given combustion process is dependent on the type of fuel, the configuration of the combustion chamber, the nature of the fuel firing equipment, and the effectiveness of mixing combustion air with the fuel. Traditionally, these components have been referred to as the "three T's of combustion"—time, temperature, and turbulence. This is the art of combustion rather than the chemistry of combustion.

**Table 51.5** Theoretical Air Combustion

Combustible	Molecular Symbol	Pounds per Pound of Combustible					
		Required for Combustion			Products of Combustion		
		O <sub>2</sub>	N <sub>2atm</sub>	Air	CO <sub>2</sub>	H <sub>2</sub> O	N <sub>2atm</sub>
Carbon	C	2.67	8.87	11.54	3.67		8.87
Hydrogen	H <sub>2</sub>	8.00	26.57	34.57		9.00	26.57
Sulfur	S	1.00	3.32	4.32	2.00 SO <sub>2</sub>		3.32
Carbon monoxide	CO	0.57	1.89	2.46	1.57		1.89
Methane	CH <sub>4</sub>	4.00	13.29	17.29	2.75	2.25	13.29



Acetylene	C <sub>2</sub> H <sub>2</sub>	3.08	10.23	13.31	3.38	0.70	10.23
Ethylene	C <sub>2</sub> H <sub>4</sub>	3.43	11.39	14.82	3.14	1.29	11.39
Ethane	C <sub>2</sub> H <sub>6</sub>	3.73	12.39	16.12	2.93	1.80	12.39

Table 51.6 displays excess air levels commonly considered for various fuels and types of combustion equipment.

**Table 51.6** Commonly Achieved Excess Air Levels

Fuel	Firing Method	Excess Air, % by Weight
Coal	Pulverized	20
	Fluidized bed	20
	Spreader stoker	30
Fuel oil	Steam-atomizing/register-type burners	10
Natural gas	Register burners	5
Municipal solid waste	Reciprocating grate	80

Excess air serves to dilute and thereby reduce the temperature of the products of combustion. The reduction of temperature tends to reduce the heat energy available for useful work. Therefore, the actual excess air used in the combustion process is a balance between the desire to achieve complete combustion and the need to maximize the heat energy available for useful work.

It is frequently useful to know the temperature attained by combustion. The heat released during combustion heats the products of combustion to a calculable temperature. It must be understood that the calculation procedure presented here assumes complete combustion and that no heat is lost to the surrounding environment. Thus, combustion temperature is useful for comparing one combustion process with another. The heat available for heating the products of combustion is the lower heating value of the fuel. The increase in temperature is the lower heating value divided by the mean specific heat of the products of combustion. The mean specific heat is a function of the constituent products of combustion ( $W_{P.C.}$ ) and the temperature. To approximate the theoretical temperature attainable, one can use a specific heat of 0.55 Btu/lb per degree Fahrenheit for water vapor ( $W_{H_2O}$ ) and 0.28 Btu/lb per degree Fahrenheit for the other gaseous products of combustion ( $W_{P.C.} - W_{H_2O}$ ). Thus, the formula for approximating the temperature attained during combustion is:

$$T_{\text{comb}} = T_{\text{in}} + \frac{\text{LHV}_{\text{Btu/lb}}}{0.55W_{H_2O} + 0.28(W_{P.C.} - W_{H_2O})} \quad (51.2)$$

## 51.2 Combustion Calculations

Typical combustion calculations are provided here for bituminous coal to determine the products

of the combustion process. Similar calculations can be developed for any fuel for which the heating value and the ultimate analysis are known.

Each of the combustible substances combines and completely combusts with oxygen as displayed in Table 51.2. The weight ratio of oxygen to the combustible substance is the ratio of molecular weights. Table 51.5 displays the weight or volume of oxygen theoretically required for complete combustion of one pound of the combustible substance. Sulfur dioxide from combustion of sulfur in fuel is combined with CO<sub>2</sub> in the sample calculation (Table 51.7) as a matter of convenience. If desired, a separate column can be prepared for sulfur dioxide in the products of combustion. Oxygen in the fuel combines with the combustible substances in the fuel, thereby reducing the quantity of air required to achieve complete combustion. The sample calculation uses the weight percentages of oxygen to reduce the theoretical air requirements and the nitrogen in the products of combustion. The decimal percentage of excess air is multiplied by the total theoretical air requirement to establish the weight of excess air and the total air requirement including excess air.

**Table 51.7** Sample Calculation for Bituminous Coal

Ultimate Analysis				Products of Combustion						
Substance	Decimal % by Weight	Theoretical Air, lb/lb		CO <sub>2</sub> incl. SO <sub>2</sub>		H <sub>2</sub> O		N <sub>2</sub>		O <sub>2</sub>
(1)	(2)	(3)	(2) × (3)	(4)	(2) × (4)	(5)	(2) × (5)	(6)	(2) × (6)	
Moisture	0.029					× 1.00	0.03			
Carbon	0.803	× 11.54	9.27	× 3.67	2.95			× 8.57	6.88	
Hydrogen	0.045	× 34.57	1.56			× 9.00	0.41	× 26.57	1.20	
Sulfur	0.015	× 4.32	0.06	× 2.00	0.03			× 3.32	0.05	
Nitrogen	0.014							× 1.00	0.01	
Oxygen	0.028	+ 0.2314	(2) + (3)					× 0.7686	- 0.09	
			- 0.12							
Ash	0.066									
Total	1.000		10.77		2.98		0.44		8.05	
HHV	14 100 Btu/lb									
20% Excess air			2.15					× 0.7686	1.65	× 0.2314
Total dry air			12.92							0.50
Moisture in air		× 0.013	0.17				0.17			
					2.98		0.61		9.70	0.50
$\text{LHV} = 14\,100 - [0.29 + (9 \times 0.045)] \times (1050)$ $= 13\,640 \text{ Btu/lb}$										
$\text{Temperature developed in combustion} = 60 + \frac{13\,640}{(0.61 \times 0.55) + (13.79 - 0.61)0.28}$ $= 3450^{\circ}\text{F}$										
Total products of combustion										
CO <sub>2</sub>	2.98	0.216								
H <sub>2</sub> O	0.61	0.044								
N <sub>2</sub>	9.70	0.704								
O <sub>2</sub>	0.50	0.036								
Total	13.79 lb/lb fuel	1.000 decimal%								

Heating value and ultimate analysis for gaseous fuels are frequently given on a volumetric basis at 60°F and 30 in. Hg. Data on a volumetric basis can be converted to a weight basis by multiplying the volumetric percentages by the density from Table 51.1 to determine an average density in the percent by weight as well as the heating value by weight.

## Defining Terms

**Bomb calorimeter:** A laboratory device for measuring the heat content of solid and liquid fuels. The "bomb" is a pressure-tight container in which a measured quantity of fuel is placed, the bomb is closed, and it is then pressurized with oxygen. The bomb is then submerged in the calorimeter "bucket" containing a measured quantity of water. The fuel in the bomb is ignited electrically to combust the fuel. The heat generated is absorbed by the water in the bucket so that the temperature rise and the quantity of water in the bucket are the measure of the heat released.

**Combustion:** The rapid oxidation of combustible substances with release of heat.

**Combustible substance:** A substance that when heated to its ignition temperature in the presence of oxygen rapidly oxidizes and releases heat. Carbon and hydrogen are by far the most important of the combustible substances.

**Proximate analysis:** Analysis made of a solid fuel that provides a guide to its combustibility. Proximate analysis includes percentages of moisture, volatile matter, fixed carbon, and ash adding to 100%. The higher the percentage of volatile matter, the more flame will be manifested during combustion.

**Ultimate analysis:** Analysis of fuel, including the chemical constituents important in the combustion process. The ultimate analysis usually includes carbon, hydrogen, sulfur, nitrogen, oxygen, ash, and moisture, the percentages adding to 100. Oxygen is normally determined by difference.

## References

- ASTM. Gaseous fuels; coal and coke. *Annual Book of ASTM Standards*. Section 5, Volume 05.05. American Society for Testing and Materials, Philadelphia, PA.
- Lide, D. W., Jr. (ed.) 1990. *CRC Handbook of Chemistry and Physics*. 71st ed. CRC Press, Boca Raton, FL.
- Stultz, S. C. and Kitto, J. B. (eds.) 1992. *Steam—Its Generation and Use*, 40th ed. Babcock & Wilcox, Barberton, OH.

## Further Information

### Handbooks

- Perry, R. H. and Green, D. W. (eds.) 1984. *Perry's Chemical Engineer's Handbook*, 6th ed. McGraw-Hill, New York.
- Chopey, N. P. and Hicks, T. G. (eds.) 1993. *Handbook of Chemical Engineering Calculations*, 2nd ed. McGraw-Hill, New York.
- Avalione, E. A. and Baumeister, T., III. (eds.) 1987. *Marks Standards Handbook for Mechanical Engineers*, 9th ed. McGraw-Hill, New York.
- Reed, R. J. *North American Combustion Handbook: Combustion, Fuels, Stoichiometry, Heat Transfer, Fluid Flow*, 3rd ed. 1988. North American Manufacturing, Chicago, Illinois.

**Societies**

American Institute of Chemical Engineers

345 East 47th Street

New York, NY 10017

American Society of Mechanical Engineers

Fuels & Combustion Technology Division

345 East 47th Street

New York, NY 10017

Goldschmidt, V. W., Wahlberg, C. J. "Air Conditioning"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Air Conditioning

---

## 52.1 Historical Sketch

## 52.2 Comfort

## 52.3 Air Conditioning Process

## 52.4 Representative Cycles

Vapor Compression Cycle • Evaporative Cooling • Absorption Cooling • The Issue of Refrigerants

### Victor W. Goldschmidt

*Purdue University*

### Curtis J. Wahlberg

*Purdue University*

The purpose of this chapter is to introduce the reader to basic concepts in air conditioning, as well as some of its history. Details on equipment types, design, selection, installation, operation, and maintenance are not addressed. The ASHRAE (American Society of Heating, Refrigerating and Air-Conditioning Engineers) handbooks are suggested as a source for further study and reference.

## 52.1 Historical Sketch

---

The term *air conditioning* was probably coined by S. W. Cawrer in 1906, when he spoke of air conditioning as a means of "conditioning" cotton fibers with humid air in North Carolina. Interestingly, an early textbook (*Air Conditioning*, John Wiley, 1908) restricted air conditioning to the process of air humidification. The concept has broadened considerably and is now understood to be the process of cooling and controlling the moisture of air, primarily in occupied spaces.

The history of air conditioning is fascinating. As early as the 1600s, philosophers such as Francis Bacon (1561–1626) deviated from religion, government, and ethics to observe weather and heat. He was driven to state that "Heat and cold are Nature's two hands whereby she chiefly worketh; heat we have in readiness in respect to fire, but for cold we must stay till it cometh or seek it in deep caves and when all is done we cannot obtain it in a great degree." Bacon supposedly died of chills and fever contracted while experimenting on using snow as a preserver of game fowl.

A key pioneer in air conditioning/**refrigeration** was William Cullen. In 1748 he published a

paper on "Cold Produced by Evaporating Fluids," which presented his success in lowering the temperature of water to freezing conditions when evaporating ether. At that time theologians believed that making ice was only a divine function and that Cullen (a physician secretly teaching chemistry in a university hospital) was "trafficking with the devil" with his initiatives. It was probably John Leslie who became the first tangible spark in absorption cooling/refrigeration. A divinity school graduate and tutor to Josiah Wedgewood's family, his inquiries into the nature of heat were seen by later colleagues at the University of Edinburgh to be close to devil worship; they advised him to restrict his efforts to pure mathematics. Fortunately, he did not, and neither did many other brave pioneers who followed in refrigeration and air conditioning.

It could be that the first air conditioning system was the result of concerns that a physician had for the comfort of ailing sailors suffering with critical fever. In the mid-1800s Dr. John Gorrie constructed an open air-cycle refrigeration machine in order to cool two rooms in a hospital in Apalachicola, Florida, and provide his patients with some relief. Although Gorrie died frustrated, broke, and subject to criticism, he did become the first of a line of air conditioning pioneers, which also includes A. Muhl (who held the first patent for cooling residences—in this case with ether compression and expansion), Alfred R. Wolft (1859–1909; who provided comfort conditioning to more than 100 buildings, including the Waldorf Astoria, Carnegie Hall, and St. Patrick's Cathedral), and Willis Carrier (1876–1950; who not only provided the first psychrometric chart, but also set new trends in product development and marketing).

Equally as fascinating as the history of air conditioning are its thermodynamics, product development, and utilization. Some of these will be addressed in the sections that follow.

## 52.2 Comfort

---

The main purpose of air conditioning is human comfort in the built environment. The human body is continuously generating heat, which in turn must be transferred to the environment in order to maintain a constant body temperature. The transfer of heat from the body to the environment depends on the surface conditions of the body (i.e., clothing), as well as the temperature, velocity, and humidity of the surrounding air and the temperature of surrounding surfaces (affecting radiative heat transport to or from the body). Conditions for comfort are obviously influenced by an individual's age, level of activity, and clothing. In general, nominal ranges for comfort include temperature from 20 to 26°C and **dew-point temperature** from 2 to 17°C, with air velocities under 0.25 m/s. [For details see ASHRAE (1991, 1992).] The control of temperature and moisture content is generally given to an air conditioning system.

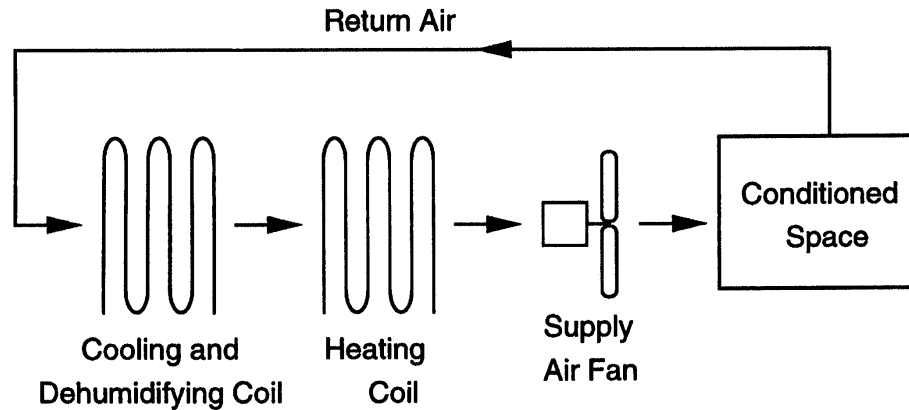
## 52.3 Air Conditioning Process

---

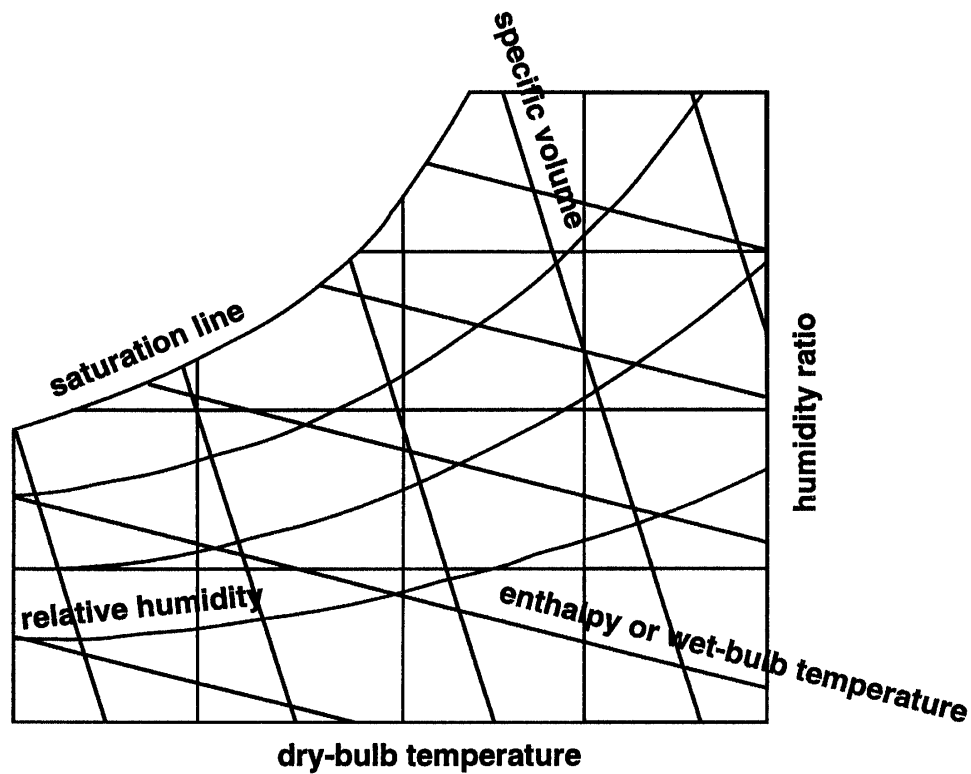
The typical process of air conditioning, exhibited in [Fig. 52.1](#), can best be described on a psychrometric chart. These charts present data for **relative humidity**, **specific volume**, **wet-bulb**

**temperature**, and **enthalpy** for atmospheric air in terms of **dry-bulb temperature** and **humidity ratio**, as shown schematically in Fig. 52.2. [For details see chapter 6 of ASHRAE (1993).]

**Figure 52.1** Typical air conditioning process.



**Figure 52.2** Psychrometric chart schematic.

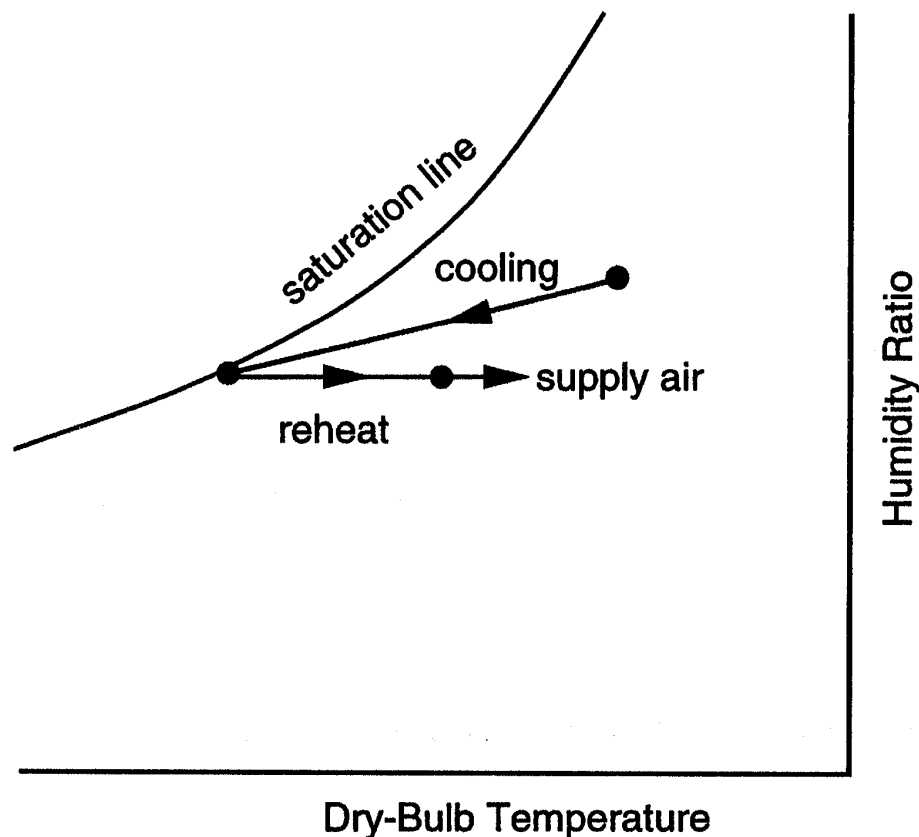


The cooling and dehumidifying coil of Fig. 52.1 brings the moist air down to its dew point then



removes moisture in the form of condensate. Air leaves the coil in conditions close to saturated and at a lower temperature than the room return air (see [Fig. 52.3](#)). The latent cooling is that amount necessary to have the humidity ratio of the room air at the comfort level. In turn, the heating section brings the dry-bulb temperature up to comfort conditions. This kind of system (sometimes called *reheat*) can easily provide the comfort conditions required. Alternatively, the saturated cold air can be mixed with warmer make-up air (or part of the return air), avoiding the need for reheat (which is a form of energy loss).

**Figure 52.3** Cooling and dehumidification with reheat.



## 52.4 Representative Cycles

### Vapor Compression Cycle

[Figure 52.4](#) shows the basic components for a vapor compression cycle used as an air conditioner. Room air flows over the evaporator, hence being cooled (sensible and latent cooling), while the two-phase refrigerant within the coil boils and heats beyond the saturated condition and slightly

into the **superheated** region. The compressor increases the pressure of the refrigerant, which is then cooled slightly beyond saturation in the condenser using the outside ambient air. Through this process, the air conditioner gains thermal energy,  $Q_L$ , from the cooled indoor ambient and rejects thermal energy,  $Q_H$ , to the outdoor environment while work,  $W$ , is done on the cycle through the compressor. The line losses are generally small; hence an energy balance would require that

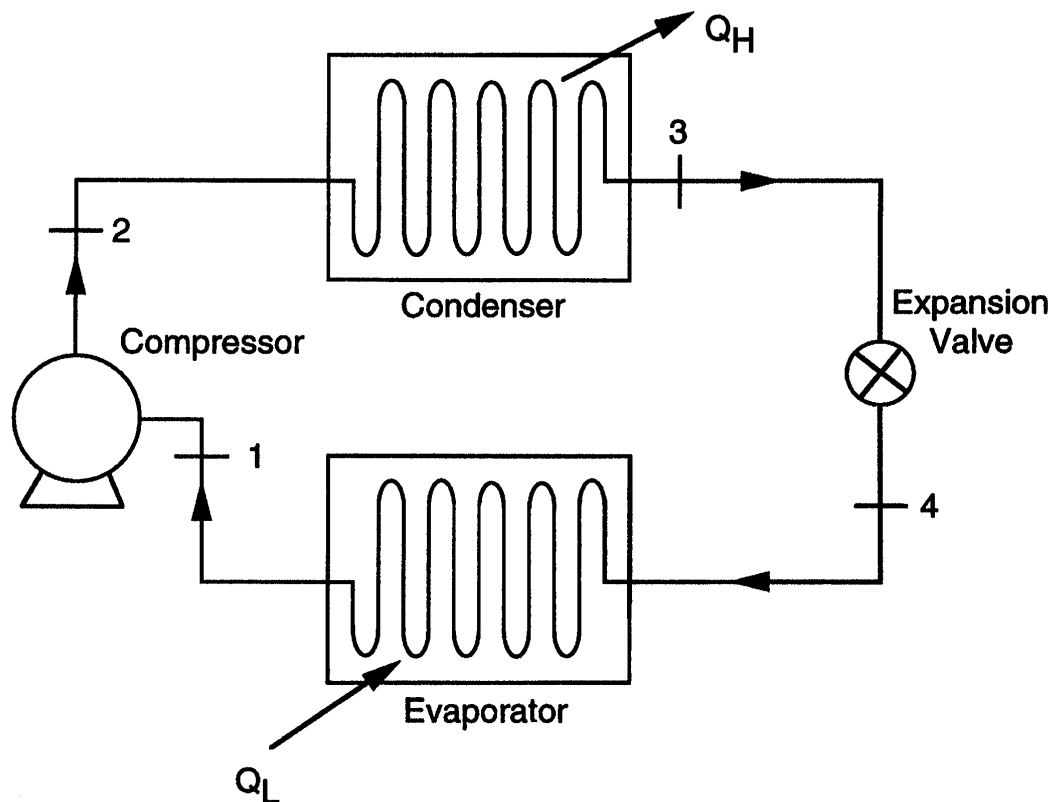
$$W + Q_L = Q_H$$

with the coefficient of performance, COP, defined as

$$COP = Q_L / W$$

The industry has also accepted a similar ratio given by a nondimensionless coefficient of performance, the energy efficiency ratio (EER), given in units of Btu/h of cooling per watt of work.

**Figure 52.4** Vapor compression cycle components.

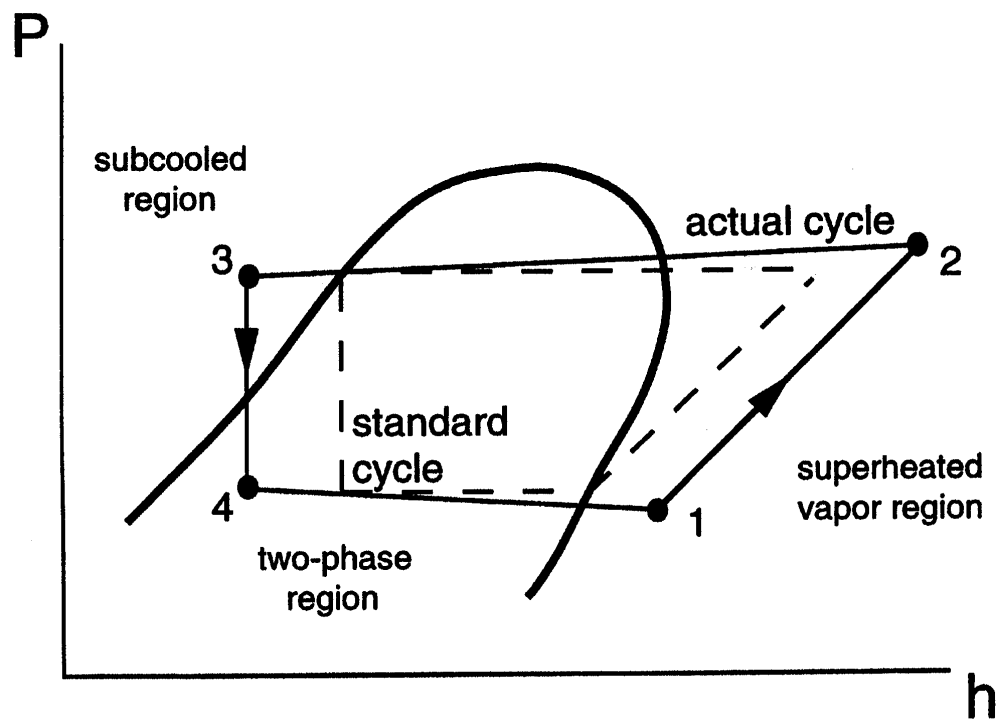


The compressor is generally driven by an electric motor; however, in large systems gas-driven

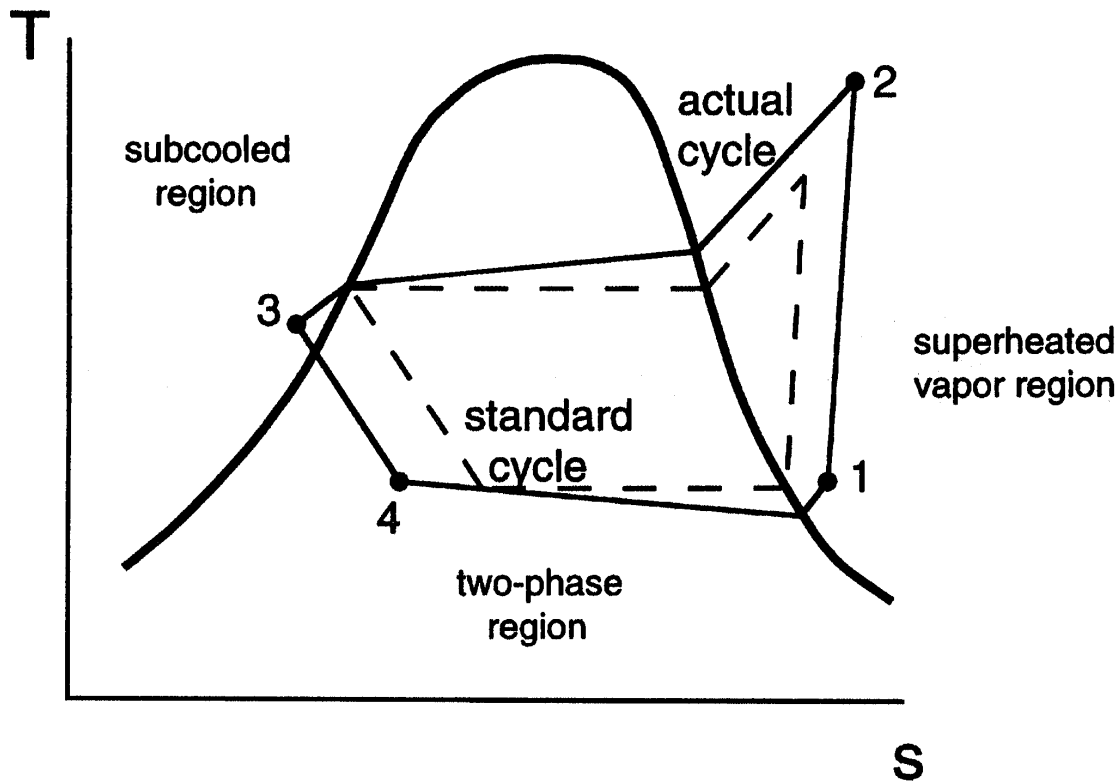
engines are also used. The evaporator might also function as a chiller, cooling a brine in a secondary loop for terminal units that provide cooling in different zones of the building. In a similar manner the condenser might reject its heat through cooling towers, ponds, or ground-coupled sources/sinks. The selection of these other alternatives is dependent on energy sources, building type, cost parameters, and even local preferences.

The corresponding refrigerant  $Ph$  (pressure, enthalpy) and  $Ts$  (temperature, entropy) diagrams for the vapor compression systems are shown in Figs. 52.5 and 52.6, respectively. The slight superheat at the outlet of the evaporator is to ensure that no liquid droplets enter the compressor (suction side accumulators are also used but are not shown in the schematic). The **subcooled** conditions downstream of the condenser are desirable to ensure proper response and control by the expansion valve (normally a thermostatic expansion valve or an electrostatic expansion valve). These valves control superheat by adjusting the mass flow rate of the refrigerant. In some cases fixed-area throttling devices might be used (such as capillary tubes or fixed-area orifices).

**Figure 52.5**  $Ph$  diagram: vapor compression cycle.



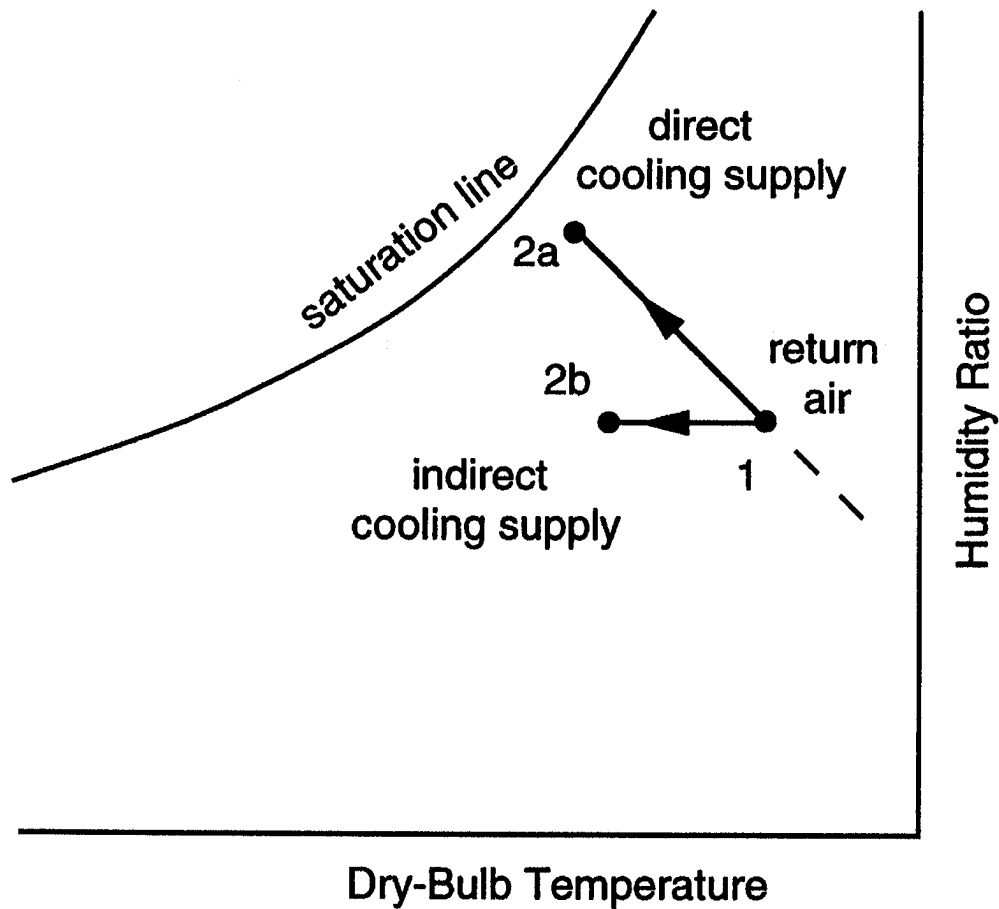
**Figure 52.6**  $Ts$  diagram: vapor compression cycle.



## Evaporative Cooling

Evaporative cooling is a cost-effective alternative for dry climates. Water, when sprayed and evaporated, will lead to a lowering of the dry-bulb temperature in the air. The cooling may be *direct* (the water is added to the air being conditioned) or *indirect*. In the case of indirect cooling a secondary airstream (cooled by evaporation of a spray) flows through a heat exchanger, removing thermal energy from the conditioned ambient return air. The corresponding processes are shown in Fig. 52.7 [For details see [ASHRAE, 1991](#), chapter 46].

**Figure 52.7** Evaporative cooling

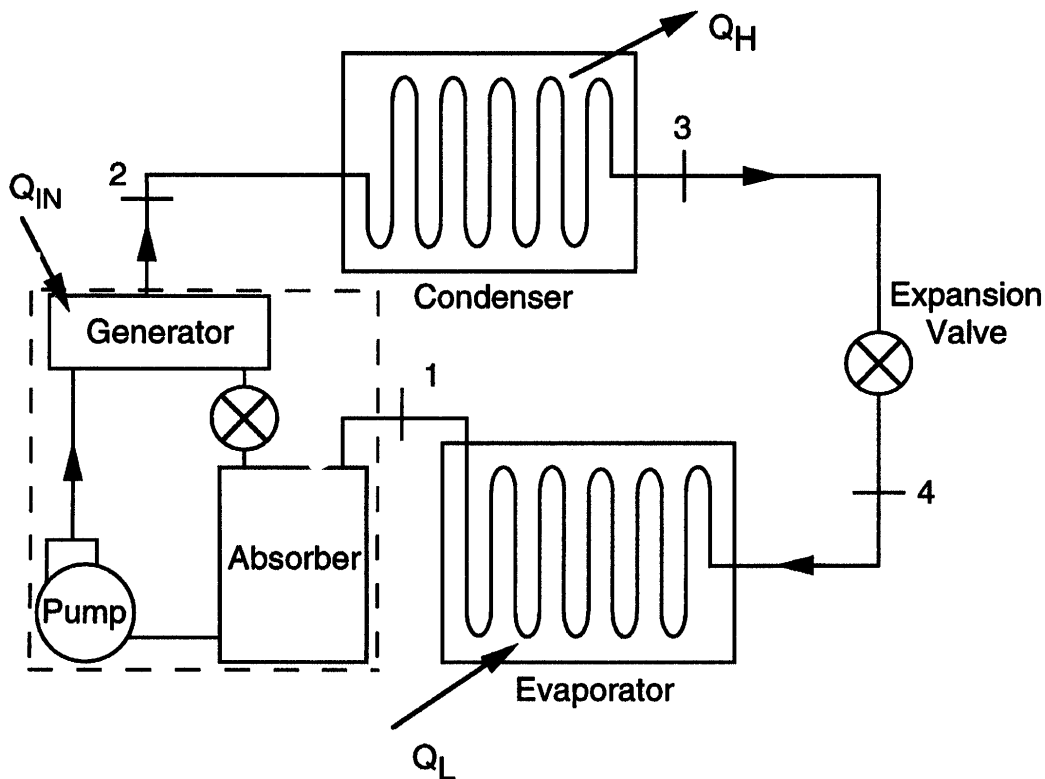


## Absorption Cooling

Ferdinand Carre invented the absorption system in the mid-1800s. In lieu of a compressor, as used in vapor compression systems, an absorption system uses three general steps, as diagrammed in [Fig. 52.8](#). First, the refrigerant leaving the evaporator (in a vapor state) is absorbed in a liquid. This operation takes place in an *absorber*. Heat is removed during this step. Second, the pressure of the solution is increased through a liquid pump. Third, the high-pressure solution is heated in order to release the high-pressure refrigerant in a vapor state and transfer it to the condenser. This third step takes place in a *generator*. The remaining solution (now with a very low concentration of refrigerant) flows through a throttling valve and back to the absorber. Two features of the absorption system make it preferable over a vapor-compression system. The compressor is exchanged with a liquid pump consuming much lower energy and a heat source. (The heat source drives off the refrigerant vapor from the high-pressure liquid.) Furthermore, absorption cooling lends itself to the use of waste heat, solar energy, or similar sources. Carre's original patent called

for ammonia ( $\text{NH}_3$ ) as the refrigerant and water ( $\text{H}_2\text{O}$ ) as the transport medium. Other systems include water–lithium bromide and water–lithium chloride systems, where water is used as the refrigerant (hence, limited to air conditioning applications where the refrigerant temperature remains above freezing). For more details see chapter 17 of Stoecker and Jones [1982].

**Figure 52.8** Absorption cooling cycle components.

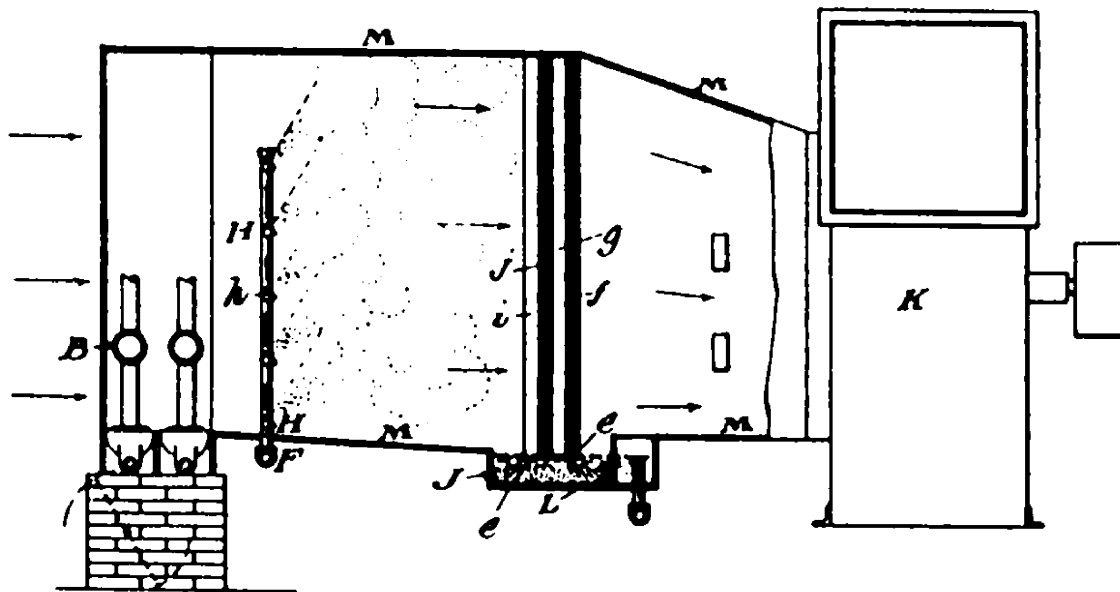


## The Issue of Refrigerants

Refrigerants originally selected for their safety, cost, stability, compatibility with materials, and excellent performance (such as R12 for refrigerators and automobile air conditioners and R22 for residential air conditioners) have recently been the subject of major controversy.

Data of ozone concentration in the stratosphere suggested a depletion, and theoretical models placed part of the culpability on chlorofluorocarbons (such as R12). Not all researchers agree with these conclusions, which are based on complex mathematical models with relatively simplistic approximations (among these are the values and nature of the turbulent transport coefficients) and assumptions (such as the effect of other constituents and dynamics in the atmosphere). The disagreement is also based on the limited data and recognition that there are major temporal and spatial scales that demand long-time data and trends for proper inferences [Anderson *et al.*, 1990].

Similar models for the potential greenhouse effect of refrigerants have led to the inference that refrigerants such as R22 might have an undesirably high "global warming potential." The data on "global warming" are subject to even more controversy than those for ozone depletion. There are long-scale effects related to the orbital motion of our planet that might mask any global warming effects attributed to refrigerants like R22. Nonetheless, there is currently a strong effort to replace R22 as well as R12. This fact is having an impact on the air conditioning industry, with potential major effects in years to come [Goldschmidt, 1992].



---

*Willis H. Carter*

#808,897

The object of the invention is to provide an efficient practical apparatus of simple construction which will thoroughly separate all solid impurities, floating particles, and noxious materials from the air either with or without altering its temperature and humidity.

© 1998 by CRC PRESS LLC

also very useful in industrial processing where strict environmental control enhanced production quality.

The Carrier Engineering Corporation was formed in 1915 and both industrial and residential air conditioning and refrigeration equipment is still being produced today bearing the Carrier name. (© 1993, DewRay Products, Inc. Used with permission.)

## Defining Terms

**Dew-point temperature:** Temperature at which a mixture of dry air and water vapor becomes saturated while cooled at constant pressure from an unsaturated state.

**Dry-bulb temperature:** Conventional thermometer temperature.

**Enthalpy:** Intrinsic property of a substance determined by the summation of its internal energy and the product of its pressure and volume (denoted  $h = u + Pv$  ).

**Humidification:** Process of adding moisture to an airstream (dehumidification: removal of moisture).

**Humidity ratio:** Ratio of water vapor to dry air in a mixture.

**Refrigeration:** Act of maintaining a low-temperature region of finite size at a selected temperature by removing heat from it.

**Relative humidity:** Ratio of partial pressure of the vapor in a mixture to the saturation pressure of the vapor at the same dry-bulb temperature of the mixture.

**Specific volume:** Volume per unit mass of a substance (generally denoted  $v$ ).

**Subcooled:** Conditions where a liquid is cooled below the saturation temperature at a given pressure.

**Superheated:** Conditions where a vapor is heated above the saturation temperature at a given pressure.

**Wet-bulb temperature:** Temperature from a conventional thermometer where the bulb is covered with a wetted wick.

## References

Anderson, M. K., Cox, J. E., and Goldschmidt, V. W. 1990. A perspective on the status of the environmental problem with CFC refrigerants. *Proceedings of the Congress Cold '90*. Buenos Aires, Argentina, 12–8 October.

ASHRAE. 1991. *1991 ASHRAE Handbook HVAC Applications*. ASHRAE, Atlanta, GA.

ASHRAE. 1992. *Thermal Environmental Conditions for Human Occupancy*. ANSI/ASHRAE Standard 55-1992. ASHRAE, Atlanta, GA.

ASHRAE. 1993. *1993 ASHRAE Handbook Fundamentals*. ASHRAE, Atlanta, GA.

Goldschmidt, V. W. 1992. Status and technical aspects related to CFC replacements. *International Congress Cold '92*. Buenos Aires, Argentina, 7–9 September.

Stoecker, W. F. and Jones, J. W. 1982. *Refrigeration and Air Conditioning*. McGraw-Hill, New York.



Barron, R. F. "Refrigeration and Cryogenics"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Refrigeration and Cryogenics

## 53.1 Desiccant Cooling

## 53.2 Heat Pumps

## 53.3 Cryogenics

Physical Properties of Cryogenic Liquids • Cryogenic Refrigeration Systems

### Randall F. Barron

Louisiana Tech University

*Refrigeration* involves the production and utilization of temperatures below normal ambient temperature. The general goal of conventional refrigeration is not only to achieve low temperatures, but also to maintain a space or material at subambient temperatures for an extended time. Some examples of the utilization of refrigeration include (a) cooling of residences (air conditioning), (b) ice making, (c) cold storage of foods, and (d) condensation of volatile vapors in the petroleum and chemical industries.

**Cryogenics** is a special field of refrigeration that involves the production and utilization of temperatures below  $-150^{\circ}\text{C}$  or 123 K. This dividing line was chosen between conventional refrigeration and cryogenics because the refrigerants used in air conditioning, ice making, food cold storage, etc., all condense at temperatures much higher than  $-150^{\circ}\text{C}$ . Typical cryogenic fluids (listed in [Table 53.1](#)) include methane, nitrogen, oxygen, argon, neon, hydrogen, and helium. Some examples of the utilization of cryogenic temperatures are (a) production of liquefied gases, (b) maintaining low temperatures required for superconducting systems, (c) treatment of metals to improve the physical properties of the material, and (d) destruction of defective tissue (cryosurgery).

**Table 53.1** Properties of Cryogenic Liquids

Cryogenic Liquid	NBP <sup>1</sup> ,K	TP <sup>2</sup> ,K	Density at NBP,kg/m <sup>3</sup>	Latent Heat,kJ/kg	Specific Heat,kJ/kg-K	Viscosity, $\mu\text{Pa} \cdot \text{s}$
Helium-3	3.19	<sup>3</sup>	58.9	8.49	4.61	1.62
Helium-4	4.214	<sup>4</sup>	124.8	20.90	4.48	3.56
Hydrogen	20.27	13.9	70.8	445.6	9.68	13.2
Neon	27.09	24.54	1206.0	85.9	1.82	130
Nitrogen	77.36	63.2	807.3	50.4	2.05	158
Air	78.8	—	874	205	1.96	168
Fluorine	85.24	53.5	1507	166.3	1.54	244

Argon	87.28	83.8	1394	161.9	1.136	252
Oxygen	90.18	54.4	1141	213	1.695	190
Methane	111.7	88.7	424.1	511.5	3.461	118

<sup>1</sup>Normal boiling point (boiling point at 1 atm pressure).

<sup>2</sup>Triple point (approximately the freezing point).

<sup>3</sup>He-3 has no triple point.

<sup>4</sup>Lambda-point temperature = 2.171 K.

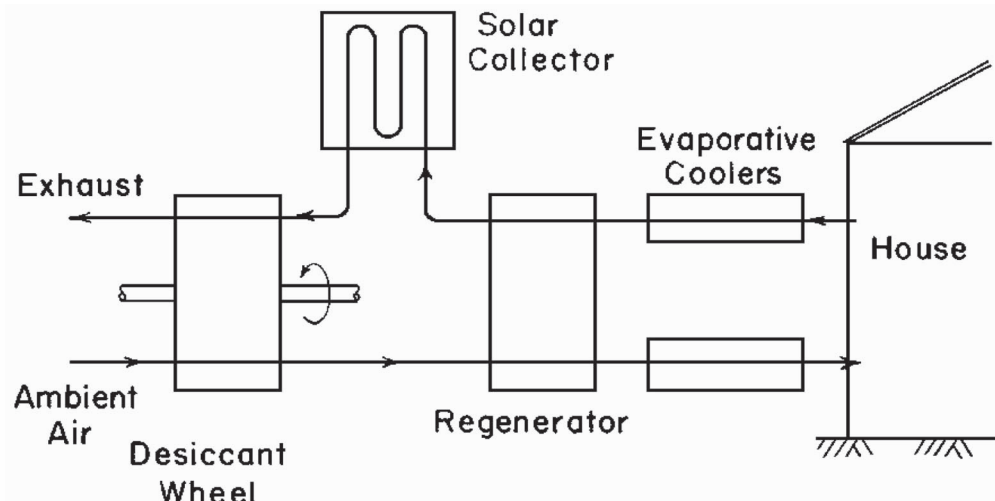
Source: Johnson, V. J. (Ed.) 1960. *A Compendium of the Properties of Materials at Low Temperature, Part I*. WADD Tech Rep. 60-56. U.S. Government Printing Office, Washington, D.C.

## 53.1 Desiccant Cooling

**Desiccant** cooling has been used in connection with solar cooling systems because of the ability of the desiccant system to achieve both cooling and dehumidification [Duffie and Beckman, 1991]. Lof [1955] described a **closed-cycle** desiccant cooling system using liquid triethylene glycol as the drying agent. In this system, glycol was sprayed into an absorber to absorb moisture from the air from the building. The stream then flowed through a heat exchanger to a stripping column, where the glycol was mixed with a stream of solar-heated air. The high-temperature air removed water from the glycol, which then returned to the absorber. This system, using steam as the heating medium, has been used in hospitals and similar large installations.

An **open-cycle** desiccant cooling system is shown in Fig. 53.1 [Nelson *et al.*, 1978]. The supply air stream is dehumidified in a desiccant regenerative heat exchanger (wheel), then cooled in a heat exchanger and an evaporative cooler before being introduced into the space to be cooled. The return-air stream from the space is evaporatively cooled before returning through the heat exchanger to a solar collector, where the air is heated. The hot air flows back through the desiccant "wheel," where moisture is removed from the desiccant before the air stream is exhausted to the atmosphere.

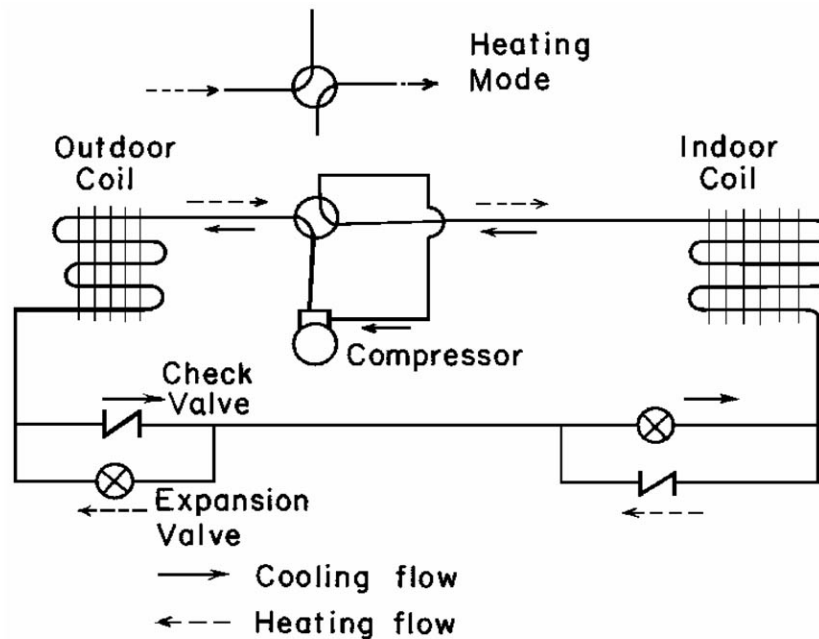
**Figure 53.1** Solar desiccant cooler. The desiccant wheel acts to dehumidify the ambient air entering the system.



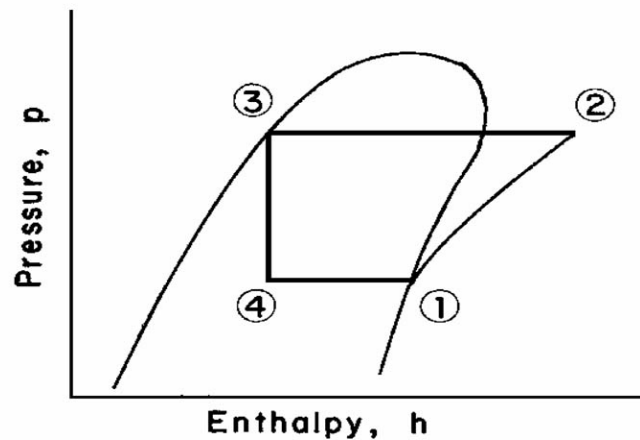
## 53.2 Heat Pumps

Heat pumps are systems that can provide either cooling or heating for a building [McQuiston and Parker, 1994]. The schematic of a typical heat pump is shown in Fig. 53.2. The heat pump cycle is shown on the pressure-enthalpy plane in Fig. 53.3.

**Figure 53.2** Heat pump. The flow directions for the cooling mode are shown as solid arrows; the flow directions for the heating mode are shown as dashed arrows.



**Figure 53.3** Pressure-enthalpy diagram for a heat pump: (1) compressor inlet; (2) compressor outlet; (3) outlet of the outdoor coil (*cooling mode*) or outlet of the indoor coil (*heating mode*); and (4) outlet of the expansion valve.



The dimensionless **coefficient of performance (COP)** for a heat pump operating in the cooling mode is given by:

$$\text{COP}(C) = \frac{Q_A}{-W} = \frac{h_1 - h_3}{h_2 - h_1} \quad (53.1)$$

where  $Q_A$  is the energy absorbed from the region to be cooled,  $W$  is the net work input (negative work) to the heat pump, and  $h$  is the enthalpy of the working fluid at the corresponding points in the cycle. For an ideal heat refrigerator (Carnot refrigerator), the COP is

$$\text{COP}(C)_{\text{Car}} = \frac{T_L}{T_H - T_L} \quad (53.2)$$

where  $T_H$  is the high temperature (absolute) in the cycle and  $T_L$  is the low temperature (absolute) in the cycle. The COP for an actual heat pump is on the order of 1/3 to 1/2 of the ideal COP [Kreider and Rabl, 1994].

For a heat pump operating in the heating mode, the COP is defined by

$$\text{COP}(H) = \frac{Q_H}{-W} = \frac{h_2 - h_3}{h_2 - h_1} \quad (53.3)$$

where  $Q_H$  is the heat added to the region being heated. For a Carnot heat pump the ideal COP is

$$\text{COP}(H)_{\text{Car}} = \frac{T_H}{T_H - T_L} \quad (53.4)$$

In the U.S. the COP for heat pump systems is often given in units of Btu/W-h, and this ratio is called the **energy efficiency ratio (EER)**, where  $\text{EER} = 3.412 \text{ COP}$ .

The capacity of a heat pump is strongly dependent on the outside air temperature. When the outdoor temperature falls much below water freezing temperature, it is necessary to provide supplemental energy in the form of electrical resistance heating.

**Example.** A heat pump operating in the heating mode accepts energy from ambient air at 5°C (278.2 K) and transfers 8 kW (27300 Btu/h) to the inside of a residence at 22°C (295.2 K). The actual COP is 40% of the ideal or Carnot COP. Determine the required power input to the heat pump and the EER.

**Solution.** The Carnot COP for the heating mode is calculated from Eq. (53.4):

$$\text{COP}(H)_{\text{Car}} = \frac{T_H}{T_H - T_L} = \frac{295.2}{295.2 - 278.2} = 17.36$$

The actual COP is 40% of the ideal COP, or

$$\text{COP}(H) = (0.40)(17.36) = 6.95 = Q_H/(-W)$$

The power input to the heat pump is

$$-W = \frac{Q_H}{\text{COP}(H)} = \frac{8.00}{6.95} = 1.152 \text{ kW}$$

The energy efficiency rating is found as follows:

$$\text{EER} = 3.412 \text{ COP}(H) = (3.412)(6.95) = 23.7 \text{ Btu/W-h}$$

## 53.3 Cryogenics

---

The field of cryogenic refrigeration generally involves temperatures below  $-150^\circ\text{C}$  or 123 K [Scott, 1959]. This dividing line between conventional refrigeration and cryogenics was selected because the normal boiling points of ammonia, hydrogen sulfide, and other conventional refrigerants lie above  $-150^\circ\text{C}$ , whereas the normal boiling points of cryogenics, such as liquid helium, hydrogen, oxygen, and nitrogen all lie below  $-150^\circ\text{C}$ .

### Physical Properties of Cryogenic Liquids

The physical properties of several cryogenic liquids at the **normal boiling point** (NBP) are listed in Table 53.1. All of the liquids are clear, colorless, and odorless, with the exception of liquid oxygen (pale blue color) and liquid fluorine (straw-yellow color). Liquid nitrogen, liquid oxygen, liquid argon, and liquid neon are generally obtained by separating air into its constituent components in an air distillation system [Barron, 1985].

Hydrogen can exist in two distinct molecular forms; ortho-hydrogen and para-hydrogen. The equilibrium mixture at high temperatures (above room temperature) is 75% o-H<sub>2</sub> and 25% p-H<sub>2</sub>, which is called *normal hydrogen*. At NBP (20.27 K) equilibrium hydrogen is 99.8% p-H<sub>2</sub> and 0.2% o-H<sub>2</sub>.

Neither helium-4 nor helium-3 exhibits a **triple point** (coexistence of the solid, liquid, and vapor phases); however, helium-4 has two liquid phases: liquid helium-I, which is a normal liquid, and liquid helium-II, which exhibits superfluidity. Liquid helium-II exists below the **lambda point** (2.171 K).

### Cryogenic Refrigeration Systems

The coefficient of performance (COP) of the thermodynamically ideal refrigerator (Carnot refrigerator) is given by

$$\text{COP}_{\text{Car}} = \frac{T_L}{T_H - T_L} \quad (53.5)$$

where  $T_L$  = low temperature in the cycle and  $T_H$  = high temperature in the cycle. The figure of merit (FOM) for any refrigerator or cryocooler is given by

$$\text{FOM} = \frac{\text{COP}_{\text{act}}}{\text{COP}_{\text{Car}}} \quad (53.6)$$

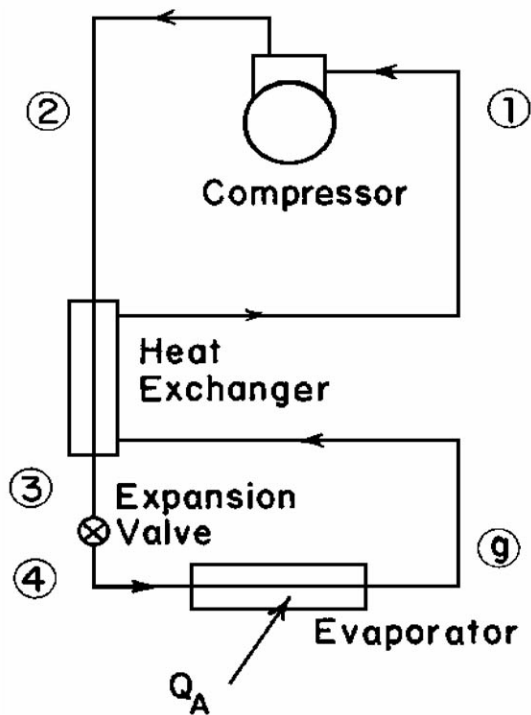
### Joule-Thomson Refrigerator

A schematic of the Joule-Thomson **cryocooler** is shown in Fig. 53.4. The refrigeration effect for this refrigerator is given by

$$Q_A/m = (h_1 - h_2) - (1 - e)(h_1 - h_g) \quad (53.7)$$

where  $e$  = heat exchanger effectiveness,  $h_1$  = fluid enthalpy at temperature  $T_2$  and pressure  $p_1$ , and  $h_g$  = saturated vapor enthalpy. Nitrogen is typically used as the working fluid for the temperature range between 65 and 115 K. Microminiature J-T cryocoolers using nitrogen-methane gas mixtures have been manufactured for operation at 65 K to cool electronic components [Little, 1990].

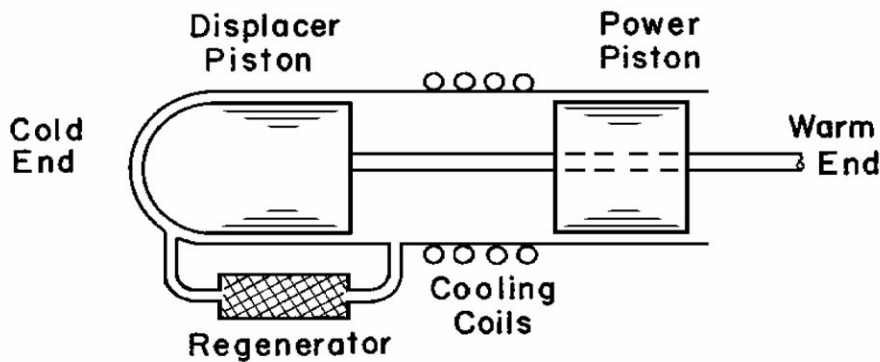
**Figure 53.4** Joule-Thomson cryocooler.



## Stirling Refrigerator

The Stirling cryocooler consists of a cylinder enclosing a power piston and a displacer piston, as shown in Fig. 53.5. The two chambers are connected through a regenerator (heat exchanger) that is one of the more critical components of the refrigerator. The detailed description of the Stirling system is given by Walker [1983]. Ideally, the Stirling refrigerator has a FOM of unity; however, actual Stirling cryocoolers have FOM values of approximately 0.36.

**Figure 53.5** Stirling cryocooler. The power piston compresses the working fluid (typically, helium gas), while the displacer piston (operating 90° out of phase) moves the working fluid from the warm space through the regenerator to the cold space and back again. Heat is absorbed from the cryogenic region at the cold end, and cooling water in the cooling coils removes the heat at the warm end.



## Vuilleumier Refrigerator

The Vuilleumier (VM) cryocooler, first patented by Rudolph Vuilleumier in 1918 in the U.S., is similar to the Stirling cryocooler, except that the VM cryocooler uses a thermal compression process instead of the mechanical compression process used in the Stirling system [Timmerhaus and Flynn, 1989]. A schematic of the VM refrigerator is shown in Fig. 53.6. The COP for an ideal VM refrigerator is given by:

$$\text{COP} = \frac{Q_A}{Q_H} = \frac{T_L(T_H - T_0)}{T_H(T_0 - T_L)} \quad (53.8)$$

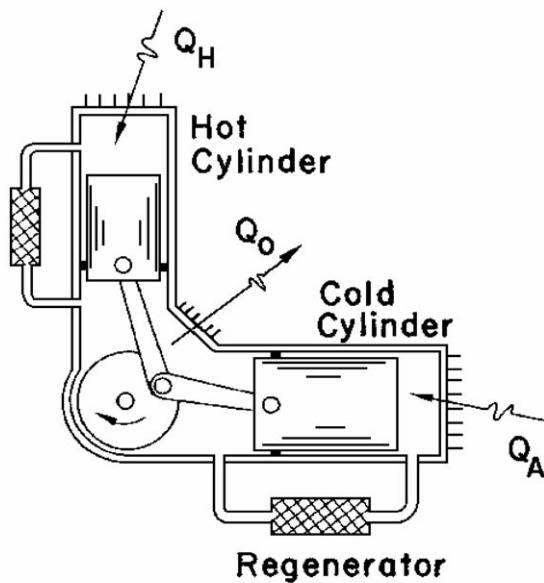
where  $T_0$  is the intermediate temperature (absolute) in the cycle.

## Gifford-McMahon Refrigerator

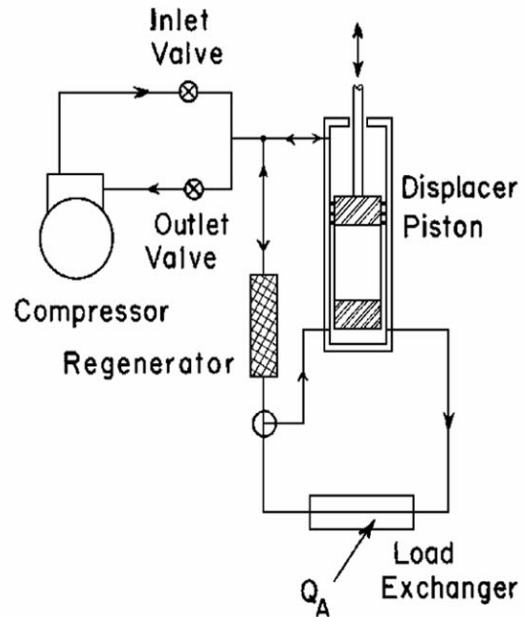
A schematic of the Gifford-McMahon (GM) cryocooler, developed by W. E. Gifford and H. O. McMahon [1960], is shown in Fig. 53.7. This system has valves and seals that operate at ambient temperature, so low-temperature valve and seal problems are eliminated. The GM refrigerator is well suited for multistaging, where refrigeration is provided at more than one temperature level. For example, in systems in which thermal shields are placed around a region to be maintained at cryogenic temperatures, refrigeration may be provided for both the thermal shields (at 77 K, for example) and the low-temperature region (at 20 K, for example) by a single refrigerator.



**Figure 53.6** Vuilleumier refrigerator.



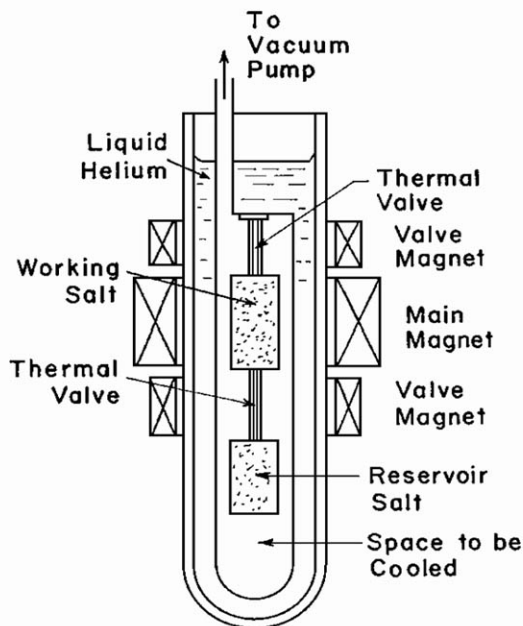
**Figure 53.7** Gifford-McMahon refrigerator.



### Magnetic Refrigerators

Cooling through the use of the **adiabatic demagnetization** process with a paramagnetic salt is one technique used to maintain temperatures below 1 K. The refrigerator consists of a paramagnetic salt such as iron ammonium alum, which is the cooling medium for the refrigerator, along with a solenoid to control the magnetic field around the working salt, as shown in Fig. 53.8. Thermal valves are used to control the heat transfer to and from the working salt. Ideally, the FOM for the magnetic refrigerator would be unity; however, real magnetic refrigerators have a FOM in the range from 0.40 to 0.80.

**Figure 53.8** Magnetic refrigerator. The main magnet controls the magnetic field within the working paramagnetic salt, while the valve magnets switch the thermal valves (strips of lead) from the "off" state (superconducting) to the "on" state (normal).



## Dilution Refrigerator

The He-3/He-4 dilution refrigerator is widely used to maintain temperatures in the range between 0.005 and 1 K. The basis for the operation of this refrigerator is the phase separation of mixtures of normal He-3 and superfluid He-4, discovered by K. G. Walters and W. M. Fairbank [1956], that occurs at temperatures below 0.86 K. The refrigeration effect of the dilution refrigerator is determined from

$$Q_A = n_3(h_m - h_i) \quad (53.9)$$

where  $n_3$  is the molar flow rate of He-3, and  $h_m$  and  $h_i$  are the enthalpies of the streams leaving and entering the mixing chamber in the refrigerator. These enthalpies can be calculated from the following, for temperatures below about 0.04 K [Radebaugh, 1967]:

$$h_m = (94 \text{ J/mol-K}^2)(T_m)^2 \quad (53.10)$$

$$h_i = (12 \text{ J/mol-K}^2)(T_i)^2 \quad (53.11)$$

where  $T_m$  is the temperature of the fluid leaving the mixing chamber and  $T_i$  is the temperature of the fluid entering the mixing chamber.

## Defining Terms

**Adiabatic demagnetization:** The process of reduction of the magnetic field around a material while allowing negligible transfer of heat. This process results in a decrease in temperature of the material.

**Closed cycle:** A system in which the working fluid is recirculated.

**Coefficient of performance (COP):** The ratio of the useful heat load to the required work input for a heat pump or a refrigerator. The COP gives a measure of the effectiveness of the system.

**Cryocooler:** A refrigerator operating at cryogenic temperatures.

**Cryogenics:** The science and technology involving very low temperatures.

**Desiccant:** A material that has a high affinity for water.

**Energy efficiency ratio (EER):** The coefficient of performance of a heat pump or refrigerator expressed in units of Btu/W-h.

**Lambda point:** The temperature at which liquid helium-4 changes from a normal fluid to a superfluid.

**Normal boiling point (NBP):** The boiling point of a liquid at standard atmospheric pressure (101.325 kPa).

**Open cycle:** A system in which the working fluid is exhausted to the atmosphere and not recirculated.

**Superfluidity:** The physical phenomenon for liquid helium-4 in which the liquid can flow with

zero pressure drop through a tube heated at one end and cooled at the other end.

**Triple point:** The state at which all three phases (solid, liquid, and vapor) coexist in thermodynamic equilibrium.

## References

- Barron, R. F. 1985. *Cryogenic Systems*, 2nd ed. Oxford University Press, New York.
- Duffie, J. A. and Beckman, W. A. 1991. *Solar Engineering of Thermal Processes*, 2nd ed., p. 606–609. John Wiley & Sons, New York.
- Gifford, W. E. and McMahon, H. O. 1960. A new low-temperature gas expansion cycle—Part II. In *Advances in Cryogenic Engineering*, vol. 5, p. 368–32. Plenum Press, New York.
- Johnson, V. J. (Ed.) 1960. *A Compendium of the Properties of Materials at Low Temperature, Part I*. WADD Tech. Rep. 60-56. U.S. Government Printing Office, Washington, DC.
- Kreider, J. F. and Rabl, A. 1994. *Heating and Cooling of Buildings*, p. 428. McGraw-Hill, New York.
- Little, W. A. 1990. Advances in Joule-Thomson cooling. In *Advances in Cryogenic Engineering*, vol. 35, p. 1305–1314. Plenum Press, New York.
- Lof, G.O.G. 1955. House heating and cooling with solar energy. In *Solar Energy Research*, p. 33. University of Wisconsin Press, Madison, WI.
- McQuiston, F. C. and Parker, J. D. 1994. *Heating, Ventilating, and Air Conditioning*, 4th ed., p. 624–626. John Wiley & Sons, New York.
- Nelson, J. S., Beckman, W. A., Mitchell, J. W., Duffie, J. A., and Close, D. J. 1978. Simulations of the performance of open cycle desiccant cooling systems. *Solar Energy*. 21:273.
- Radebaugh, R. 1967. *Thermodynamic Properties of He<sup>3</sup>-He<sup>4</sup> Solutions*. NBS Tech. Note 362. U.S. Government Printing Office, Washington, DC.
- Scott, R. B. 1959. *Cryogenic Engineering*, p. 1. Van Nostrand, Princeton, NJ.
- Timmerhaus, K. D. and Flynn, T. M. 1989. *Cryogenic Process Engineering*, p. 156–159. Plenum Press, New York.
- Walker, G. 1983. *Cryocoolers*, part 1, p. 280–282. Plenum Press, New York.
- Walters, K. G. and Fairbank, W. M. 1956. *Phys. Rev.* 103:262.

## Further Information

*Solar Energy Technology Handbook*, W. C. Dickerson and P. N. Cheremisinoff, eds., Marcel Dekker, Inc. (1980), gives additional information on solar desiccant cooling systems.

The *ASHRAE Handbook, Refrigeration Volume* and the *ASHRAE Handbook, HVAC Systems and Equipment Volume* are excellent sources for heat pump design information and for examples of the application of heat pumps. These volumes are published by the American Society of Heating, Refrigerating and Air-Conditioning Engineers, 1791 Tullie Circle N.E., Atlanta, GA 30329.

One of the most extensive sources of information on cryogenic systems is the series of volumes *Advances in Cryogenic Engineering* published by Plenum Press. These volumes include the papers presented at the Cryogenic Engineering Conferences over the past four decades.

Matthys, E. F. "Heat Transfer to Non-Newtonian Fluids"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Heat Transfer to Non-Newtonian Fluids

---

### 54.1 The Fluids

### 54.2 Friction

### 54.3 Heat Transfer

Laminar Regime • Turbulent Regime

### 54.4 Instrumentation and Equipment

### Eric F. Matthys

*University of California, Santa Barbara*

Most fluids used in industrial applications are **non-Newtonian**. The very name tells us that we lump in this category all fluids except one type, the *Newtonian* ones. Newtonian fluids are those that obey Newton's law relating shear stress and shear rate with a simple material property (the viscosity) dependent on basic thermodynamics variables such as temperature and pressure, but independent of flow parameters such as shear rate and time. These fluids constitute therefore only one category of fluids: the "simpler" ones. Non-Newtonian fluids are then defined as being *all the other ones!* One might therefore reasonably assume that the state of knowledge on non-Newtonian fluid heat transfer is much more extensive than that of Newtonian fluid heat transfer, but that is not the case. A major reason why this is so is that it is usually much more difficult to study non-Newtonian fluids because of their complex nature and complex interactions with the flow field. As a result, the majority of the information on non-Newtonian fluids is very empirical and consists primarily of simple friction or heat transfer correlations where the constants are determined by best-fit of experimental data.

In general, many engineers dealing with fluid mechanics or heat transfer think automatically only of Newtonian fluids because these are usually the only fluids covered in such undergraduate engineering classes. The average engineer has therefore typically never been exposed to non-Newtonian fluids before his or her first encounter with them on the job. This has important consequences as far as engineering practice, because serious mistakes can easily be made if one is not aware of the bizarre behavior of some of these fluids. Practically speaking, when the fluid is not a very simple or one-component liquid or gas such as water, oil, air, nitrogen, and so on, there is a very good chance that it may exhibit some non-Newtonian properties. Fluids such as suspensions of particles and fibers, slurries, **polymer** solutions and melts, **surfactants**, paints, foodstuffs, body fluids, soaps, inks, organic materials, adhesives, etc., may all exhibit non-Newtonian behavior. It behooves the engineer then to be aware of the potential problems and to decide on an informed basis whether to use simpler (but perhaps leading to large errors) Newtonian fluid correlations or to use the more complex information (maybe) available on

non-Newtonian fluids.

Given the breadth and complexity of the field, the nonspecialist reader will be better served—in this author's opinion—by a chapter that focuses on giving a general idea of the basic issues and difficulties involved and on providing some useful references. Mentioning only a few equations is indeed more likely to lead to inadvertent inappropriate usage than to be very helpful. The reader is instead referred to two excellent and readily available handbook articles which provide numerous correlations: one by Irvine and Karni [1987] for **purely viscous fluids** and one by Cho and Hartnett [1985], which emphasizes more **viscoelastic fluids**. Some other references will also be given hereafter.

## 54.1 The Fluids

---

It should be noted first that one needs only to address the issue of *moving* fluids here, because the very definition of a non-Newtonian substance implies that we are dealing with a fluid undergoing flow. Accordingly, only convective heat transfer needs to be discussed here, because both conduction and radiation are normally unaffected and require only that appropriate material properties such as thermal conductivity and emissivity be known, as in the case of Newtonian fluids. Another point to consider is that the definition of a non-Newtonian fluid covers only the *shear* viscosity of the fluid, but many non-Newtonian fluids will also show a complex *extensional* flow behavior. Some non-Newtonian fluids, perhaps the most difficult to deal with, are also viscoelastic.

The first step in trying to predict the heat transfer behavior of a non-Newtonian fluid is therefore to determine its nature and type. For simplicity, one often classifies these fluids as purely viscous versus viscoelastic, the former lacking the elastic characteristics of the latter. Their viscosity can either increase or decrease both with shear rate and with time, and some fluids may also exhibit a yield stress. Viscoelastic fluids in general exhibit memory effects, but some may show particularly large time-dependent effects in addition to shear rate–related variations (e.g., some surfactant solutions).

In general, the heat transfer to purely viscous fluids can be quantified by relatively simple modified Newtonian fluid correlations, and with some information on the viscosity one can then predict reasonably well the heat transfer and friction. These fluids have been studied early on, and many flow configurations have been investigated. Reviews and handbooks, even somewhat older ones, may provide most of the state-of-the-art information. For viscoelastic fluids, on the other hand, and especially in turbulent flow, much less information is available; and this is an area of active current research. Some of these fluids exhibit a fascinating property—that of reducing greatly the friction and heat transfer under turbulent flow conditions. These phenomena are called **drag and heat transfer reductions**. For these fluids, using a Newtonian-like approach to predict heat transfer may well result in very large errors. This field is evolving rapidly and much of the current information may need to be found in technical journals.

It is relatively easy to determine whether a fluid might be viscoelastic with some simple qualitative experiments such as recoil, finger-dip and fiber-pull, rod climbing, die swell, tubeless siphon, etc. One could also pretty much guess that most fluids that are not viscoelastic but are

nevertheless mixtures, solutions, or suspensions may likely be purely viscous non-Newtonian fluids. Beyond such simple classification, however, some data on the viscous or viscoelastic nature of the fluid will be necessary to predict heat transfer or friction quantitatively. This information has to be acquired from some **rheology** experiments, although rough estimates can at times be found in the literature. One should be especially cautioned, however, against using viscosity models beyond their range of applicability, as is often done for the ubiquitous power-law model, for example.

It cannot be emphasized enough for the engineer who is not yet familiar with non-Newtonian fluids that one should be particularly wary of relying on one's experience with Newtonian fluids or one's intuition when tackling the more complex non-Newtonian fluids. The latter's behavior may indeed be very surprising. For example, a few parts per million of a polymer in solution in water could well reduce the heat transfer by a factor of 10. Accordingly, the reader is strongly advised to consult some of the additional material referred to hereafter to develop enough understanding of the issues to avoid such problems.

## 54.2 Friction

---

The reader is referred to **Chapter 35** for fluid mechanics information on these fluids. An approach similar—although simpler—to that discussed here for heat transfer can usually be followed for friction. It is important to note, however, that the usual strong coupling between friction and heat transfer for Newtonian fluids—which enables one to predict heat transfer if one knows friction and vice versa for these fluids—may not necessarily hold for non-Newtonian fluids, another source of trouble and error.

## 54.3 Heat Transfer

---

As mentioned earlier, one is usually reduced to using simple empirical correlations to predict the heat transfer for these fluids. In addition to the handbooks listed above, the reader is also referred to a number of reviews for additional information: Metzner [1965], Skelland [1967], Dimant and Poreh [1976], Shenoy and Mashelkar [1982], Hartnett and Kostic [1989], and Matthys [1991].

As in the case of Newtonian fluids, one may use the **Nusselt number** to quantify the convective heat transfer coefficient. In the case of a purely viscous non-Newtonian fluid, the Nusselt number will usually be a function of modified **Reynolds** and **Prandtl numbers**. For viscoelastic non-Newtonian fluids, it will be in addition a function of another nondimensional number that may be related to the elasticity of the fluid through a relaxation time (e.g., a **Weissenberg number**) or to the extensional viscosity of the fluid or some other parameter. It is crucial to distinguish between the several types of Reynolds (and corresponding Prandtl) numbers that are used for the prediction of friction and heat transfer for these fluids. (Note that such a distinction is not always made clearly in the literature, and one should proceed cautiously.) Often *generalized* Reynolds and Prandtl numbers are used for laminar flow, whereas *apparent* numbers may be more suitable for turbulent flow. (For applied engineering work, one may also often use *solvent-based* numbers.) This issue is beyond the scope of this article, and good discussions of these parameters



can be found in the reviews listed above.

Since the main effect of the non-Newtonian character of the fluid on the heat transfer results from interactions between fluid and flow field through large variations in viscosity or elasticity (sometimes over several orders of magnitude), the variations in other thermophysical parameters (e.g., thermal conductivity and specific heat capacity) are often of much lesser importance. One may then often assume that these properties vary only with temperature in a given manner for a specific non-Newtonian fluid or even that they are those of the Newtonian solvent (e.g., for low-concentration solutions). Naturally, for highly concentrated suspensions of fibers or particles, for example, one has to use appropriate properties for the actual fluid, and there are procedures available to predict such properties based on the concentration of the suspended material. For more complex or lesser-known fluids, one may well be forced to measure these properties directly. It is useful to remember that for viscoelastic fluids, in particular, the large reductions in heat transfer seen in turbulent flow are not related to modifications of the "static" thermophysical properties, but rather to fluid/turbulence interactions, a dynamic process. This explains why these fluids do not normally show such reductions in heat transfer in the laminar regime. The distinction between laminar and turbulent flows is therefore even more crucial to make here than for Newtonian fluids. Note that the transition between the two regimes takes place under relatively similar conditions for both Newtonian and non-Newtonian fluids, and that the usual Newtonian fluid criteria for transition are often used.

## Laminar Regime

Heat transfer in the laminar regime can be fairly readily predicted for non-Newtonian fluids because of the simple nature of the flow and the absence of significant elastic effects. A good early discussion on this issue for some fluids is provided in Skelland [1967]. Generally, one simply modifies the Nusselt number for a Newtonian fluid with a coefficient that involves a parameter reflecting the change in velocity profile (e.g., a power-law exponent  $n$ ) and possibly the aspect ratio if it is internal flow in a noncircular duct [Irvine and Karni, 1987; Hartnett and Kostic, 1989]. For heat transfer in the entrance region, one has to introduce another parameter such as a Graetz number to account for the temperature profile development (as in the case of Newtonian fluids). Also as in the case of Newtonian fluids, the laminar thermal entrance region can be fairly long for high Reynolds numbers. Overall, the change in Nusselt number introduced by the non-Newtonian nature of the fluid is often relatively moderate in the laminar regime, and, if nothing else is available, the use of a Newtonian value may not be too far off in first approximation for fluids with a moderate power-law exponent, for example. For external flow, simple empirical corrections to Newtonian expressions can also be used [e.g., Irvine and Karni, 1987]. For free and mixed convection, relatively little information is available, and the reader is referred in particular to the review by Shenoy and Mashelkar [1982].

## Turbulent Regime

For this regime it becomes necessary to pay particular attention to the distinction between purely viscous and viscoelastic fluids. Heat transfer to many purely viscous fluids can be adequately

predicted by the use of simple relations involving the Nusselt number and apparent Reynolds and Prandtl numbers or by taking advantage of the analogy between friction and heat transfer for these fluids [Metzner, 1965; Cho and Hartnett, 1985]. The thermal entrance region for these fluids is generally similar to that of Newtonian fluids and very short (e.g., about 20 to 50 diameters for a tube).

For viscoelastic fluids, on the other hand, the situation is very different. Some suspensions of very long fibers may exhibit such a nature, as will polymer solutions and many complex fluids. Dilute solutions of a polymer or surfactant in tube flow, in particular, are of great interest. These fluids may indeed exhibit dramatic friction and heat transfer reductions with respect to the solvent alone at the same Reynolds number. This effect is very large, even with very small (e.g., sub-percent level) traces of polymer or surfactant additives. The level of drag and heat transfer reductions at a given Reynolds number will generally increase with concentration of drag-reducing additive in the solvent. Interestingly, there is a minimum below which the drag and heat transfer can no longer be reduced, the so-called drag and heat transfer reduction asymptotes [e.g., Matthys, 1991]. One advantageous feature of this **asymptotic regime** is that the concentration no longer plays a role there in determining the friction and heat transfer, which also means that the Nusselt number is no longer a function of an additional parameter besides the Reynolds and Prandtl numbers. Good correlations exist for this regime—in fact, surprisingly robust ones. The asymptotes are apparently the same for all drag-reducing fluids.

In the region between Newtonian and asymptotic behavior, the friction and heat transfer are functions of the concentration (i.e., viscoelasticity) of the fluid as well as of the diameter of the pipe. At this time, there is no universal technique that allows us to predict the heat transfer in this region based on simple material properties measurement. The diameter effect in particular is very challenging, and, even though some scaling methods have been proposed, there has not been enough experimental data made available to establish unambiguously the limits of their applicability. Note also that for a given concentration and diameter, the level of friction and heat transfer reduction in this regime will still depend on the Reynolds number (i.e., shear stress or shear rate), and a fluid may exhibit Newtonian behavior at low Reynolds number but then progressively revert to asymptotic behavior at high Reynolds number. In some cases there may be an onset shear stress that has to be exceeded before any drag or heat transfer reduction behavior is observed.

Interestingly, the heat transfer is usually reduced proportionally more than the friction for drag-reducing fluids, (e.g., 10 times versus 5 times for a typical asymptotic solution). For polymer solutions, however, the heat transfer reduction will also decrease faster than the drag reduction as the polymer is degraded mechanically. It should also be noted that the entrance lengths for these fluids are much longer than for Newtonian fluids. For polymer solutions in tubes, for example, the thermal entrance length may easily be several hundred diameters long (compared to 20 or so for Newtonian or purely viscous fluids). The hydrodynamic entrance length, on the other hand, may reach "only" 100 diameters. These apparent discrepancies between heat transfer and friction are often attributed to some "uncoupling" between the two and a breakdown of the classic Newtonian Reynolds or Colburn analogies. The very long thermal entrance lengths, in particular, should be kept in mind when designing or analyzing heat exchangers for these fluids, as it is likely that the

heat transfer may never reach fully developed conditions in many exchangers. Expressions exist that give the heat transfer in the entrance region as a function of distance [e.g., [Matthys, 1991](#)]. In appropriate nondimensional terms, the entrance region heat transfer development is independent of the Reynolds number in many cases. Interestingly, and contrary to the common belief that the entrance lengths are uncoupled for polymer solutions, recent studies have shown a very good coupling between the hydrodynamic and thermal developments for drag-reducing surfactant solutions [[Gasljevic and Matthys, 1994](#)].

Another important issue in studying or using viscoelastic fluids is that of **degradation**. Indeed, the polymeric fluids, for example, being macromolecular in nature, may be very susceptible to mechanical degradation. This means that—when subjected to flow in tubes and especially through pumps and filters—some macromolecules will be permanently broken and that the drag and heat transfer reductions will then be lost, partially at first and completely later on. This can be often seen as an increase in friction or heat transfer at high Reynolds number. Such degradation can also be caused by thermal or chemical processes. A consequence is that polymer solutions are not well suited for recirculating flows. Surfactant solutions, on the other hand, are much less susceptible to permanent degradation, although the drag and heat transfer reductions can be eliminated completely (but reversibly) under high shear stress (e.g., at high Reynolds number or in hydraulic components). They also possess the remarkable ability to reconstitute very rapidly after being subjected to high shear (e.g., in a centrifugal pump). In that respect they are very attractive and, even though little understood at this time, will undoubtedly become used more in the future. Many applications of such fluids are possible that would capitalize on their remarkable drag-reducing properties. We are studying presently, for example, the use of surfactant solutions as energy conservation agents in hydronic heating and cooling systems, a very promising application indeed.

## 54.4 Instrumentation and Equipment

---

Much of the equipment encountered or needed when dealing with non-Newtonian fluids is similar to that used with Newtonian ones, although, of course, some additional equipment may be needed for the quantification of the non-Newtonian nature of the fluid. Typically, this latter work requires the use of rheometers, perhaps over a wide range of temperatures. Numerous such devices exist and are discussed in books on rheology.

As far as basic instrumentation, a few words of caution are in order. Pressure measurements, for instance, can be complicated by viscoelastic pressure hole errors. These can be minimized by paying particular attention to the shape and uniformity of the tap holes or by using differential pressure measurements whenever possible. Flow measurements may be particularly troublesome, and one should not rely automatically on flow meters designed for water—such as orifice meters, turbine meters, etc.—which can give large errors for non-Newtonian fluids unless a careful calibration has been conducted for the specific fluid and flow conditions. Positive displacement devices are more suitable. Ultrasonic flow meters may also be used depending on their built-in velocity profile calibration. Velocity measurements with typical LDV systems should be readily possible, but the use of hot-wire anemometry would probably lead to large errors in many cases if appropriate corrections are not introduced.

For temperature measurements, thermocouples, RTDs, thermistors, thermometers, and so on are all generally suitable under steady state conditions. Note that the flow field may be affected, however, and that heat transfer may also be reduced, which may mean, for example, that a longer time might be needed before a steady state measurement is achieved. Indeed, unsteady measurements may be more difficult. When drag-reducing flows are involved, the temperature gradients in a tube, say, may also be much larger than for Newtonian fluids at a given heat flux because of the reduced convective heat transfer coefficients. More extensive mixing may then be necessary for bulk temperature measurements. Note also that—as discussed above—the thermal entrance length may be very long for these fluids and that fully developed conditions may never be reached in practical situations.

The performance of some heat exchangers may be significantly impacted by the fluids, especially drag-reducing fluids [e.g., [Gasljevic and Matthys, 1993](#)], and caution should be exercised there as well. Mixing devices may also perform differently with non-Newtonian fluids. Centrifugal pumps in most cases will work appropriately, even with improved efficiency at times, but may also degrade the fluid if excessive shear stress is applied. Pump and valve flow curves do not appear to be changed dramatically in most cases. Naturally, issues of erosion, corrosion, fouling, disposal, etc. should also be investigated for the fluid of interest.

## Defining Terms

**Asymptotic regime:** The regime of maximum drag and heat transfer reductions that can be achieved by drag-reducing fluids.

**Degradation:** A modification imparted to the fluid by mechanical, thermal, or chemical effects that changes its properties (e.g., a reduction in viscosity or drag-reducing capability).

**Heat transfer (and drag) reduction:** A sometimes dramatic decrease in heat transfer (and friction) in turbulent flow due to the viscoelastic nature of the fluid.

**Non-Newtonian fluid:** A fluid that does not obey Newton's law of simple proportionality between shear stress and shear rate. Its viscosity is then also a function of shear rate, time, etc.

**Nusselt number:** A nondimensional number involving the convective heat transfer coefficient. It is a measure of the fluid capability to transfer heat to a surface under flow.

**Polymer:** A material constituted of molecules made of repeating units (monomers).

**Prandtl number:** A nondimensional number comparing the diffusivity of momentum and heat. It reflects the relative ease of transport of momentum (friction) versus energy at the molecular level.

**Purely viscous non-Newtonian fluid:** A non-Newtonian fluid that does not exhibit elasticity.

**Reynolds number:** A nondimensional number involving the flow velocity and the fluid viscosity. It is a measure of the flow rate and level of turbulence in the flow.

**Rheology:** The study of flow and deformation of matter (generally associated with measurements of viscosity and other material properties for *non-Newtonian* fluids).

**Surfactant:** A material leading to modified (e.g., reduced) surface tension effects because of dual hydrophilic and hydrophobic nature.

**Viscoelastic non-Newtonian fluid:** A fluid that—in addition to variations of viscosity with shear rate—also exhibits an elastic character (i.e., has a memory, shows recoil, etc.).

**Weissenberg number:** A nondimensional number reflecting the viscoelasticity of a fluid through its relaxation time (similar to the Deborah number).

## References

- Cho, Y. I. and Hartnett, J. P. 1985. Non-Newtonian fluids. In *Handbook of Heat Transfer Applications*, ed. W. M. Rohsenow, J. P. Hartnett, and E. N. Ganic, pp. 2.1–2.50. McGraw-Hill, New York.
- Dimant, Y. and Poreh, M. 1976. Heat transfer in flows with drag reduction. In *Advances in Heat Transfer*, vol. 12, pp. 77–113. Academic Press, New York.
- Gasljevic, K. and Matthys, E. F. 1993. Effect of drag-reducing surfactant additives on heat exchangers. In *Developments in Non-Newtonian Flows*, ed. D. Siginer, vol. AMD-175, pp. 101–108. ASME, Washington, DC.
- Gasljevic, K. and Matthys, E. F. 1994. Hydrodynamic and thermal field development in the pipe entry region for turbulent flow of drag-reducing surfactant solutions. In *Developments in Non-Newtonian Flows II*, ed. D. Siginer. Vol. FED-206, pp. 51–61. ASME, Washington, DC.
- Hartnett, J. P. and Kostic, M. 1989. Heat transfer to Newtonian and non-Newtonian fluids in rectangular ducts. In *Advances in Heat Transfer*, vol. 19, pp. 247–356. Academic Press, New York.
- Irvine, T. F. and Karni, J. 1987. Non-Newtonian fluid flow and heat transfer. In *Handbook of Single-Phase Convective Heat Transfer*, ed. S. Kakac, R. K. Shah, and W. Aung, pp. 20.1–20.57. John Wiley & Sons, New York.
- Matthys, E. F. 1991. Heat transfer, drag reduction, and fluid characterization for turbulent flow of polymer solutions: Recent results and research needs. *J. Non-Newtonian Fluid Mech.* 38:313–342.
- Metzner, A. B. 1965. Heat transfer in non-Newtonian fluids. In *Advances in Heat Transfer*, vol. 2, pp. 357–397. Academic Press, New York.
- Shenoy, A. V. and Mashelkar, A. R. 1982. Thermal convection in non-Newtonian fluids. In *Advances in Heat Transfer*, vol. 15, pp. 143–225. Academic Press, New York.
- Skelland, A. H. P. 1967. *Non-Newtonian Flow and Heat Transfer*, John Wiley & Sons, New York.

## Further Information

The reader is also referred to journals such as *International Journal of Heat and Mass Transfer*, *Journal of Heat Transfer*, *Journal of Non-Newtonian Fluid Mechanics*, and *Journal of Rheology*. Symposia covering this subject are held at conferences by ASME, AIChE, and the Society of Rheology. Numerous excellent research papers in this field have also been published recently by—among others—the groups led by Y. Cho, A. Ghajar, J. Hartnett, T. Irvine, R. Sabersky, A. Steiff, H. Usui, and J. Zakin. The reader is also welcome to contact the author for additional information.

Swanson, L. W. "Heat Pipes"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

### 55.1 Heat Pipe Container, Working Fluid, and Wick Structures

### 55.2 Heat Transfer Limitations

### 55.3 Effective Thermal Conductivity and Heat Pipe Temperature Difference

### 55.4 Application of Heat Pipes

## Larry W. Swanson

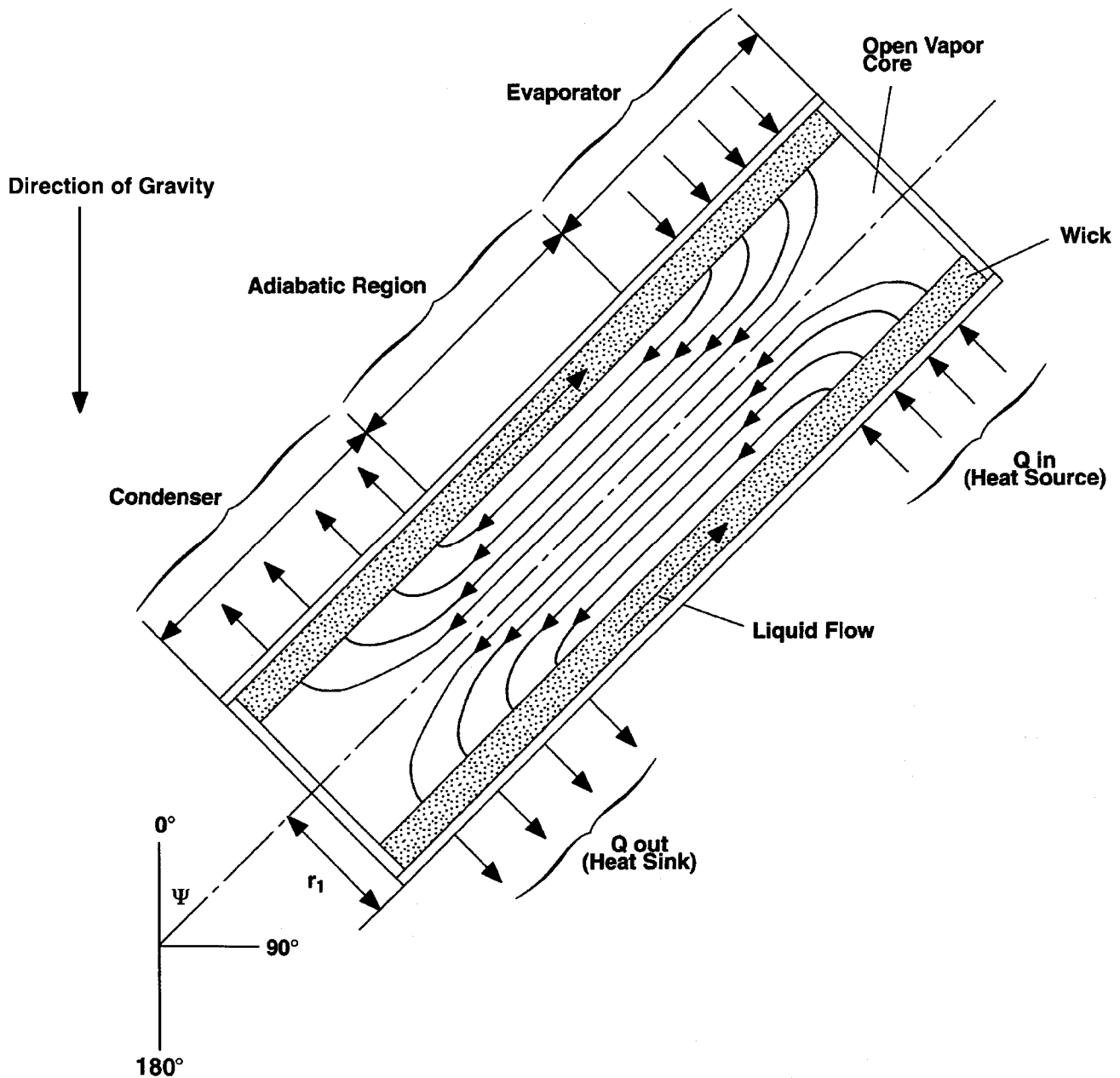
*Heat Transfer Research Institute College Station, Texas*

The heat pipe is a vapor-liquid phase-change device that transfers heat from a hot reservoir to a cold reservoir using **capillary forces** generated by a **wick**, or porous material, and a working fluid. Originally conceived by Gaugler in 1944, the operational characteristics of heat pipes were not widely publicized until 1964, when Grover and his colleagues at Los Alamos Scientific Laboratory independently reinvented the concept. Since then many types of heat pipes have been developed and used by a wide variety of industries.

Figure 55.1 shows a schematic of a heat pipe aligned at angle  $\psi$  relative to the vertical axis (gravity vector). The heat pipe is composed of a container lined with a wick that is filled with liquid near its saturation temperature. The vapor-liquid interface, usually found near the inner edge of the wick, separates the liquid in the wick from an open vapor core. Heat flowing into the evaporator is transferred through the container to the liquid-filled wicking material, causing the liquid to evaporate and vapor to flow into the open-core portion of the evaporator. The capillary forces generated by the evaporating interface increase the pressure difference between the vapor and liquid. The vapor in the open core flows out of the evaporator through the adiabatic region (insulated region) and into the condenser. The vapor then condenses, generating capillary forces similar, although much lesser in magnitude, to those in the evaporator. The heat released in the condenser passes through the wet wicking material and container out into the cold reservoir. The condensed liquid is then pumped—by the liquid pressure difference due to the net capillary force between the evaporator and condenser—out of the condenser back into the evaporator. Proper selection and design of the pipe container, working fluid, and wick structure are essential to the successful operation of a heat pipe. The **heat transfer limitations**, **effective thermal conductivity**, and axial temperature difference define the operational characteristics of the heat pipe.



**Figure 55.1** Schematic of a typical heat pipe.



## 55.1 Heat Pipe Container, Working Fluid, and Wick Structures

The container, working fluid, and wick structure of a heat pipe determine its operational characteristics. One of the most important considerations in choosing the material for the heat pipe container and wick is its compatibility with the working fluid. Degradation of the container or wick and contamination of the working fluid due to chemical reaction can seriously impair heat pipe performance. For example, noncondensable gas created during a chemical reaction eventually can accumulate near the end of the condenser, decreasing the condensation surface

area. This reduces the ability of the heat pipe to transfer heat to the external heat sink. The material and geometry of the heat pipe container also must have high burst strength, low weight, high thermal conductivity, and low porosity.

Using the proper working fluid for a given application is another critical element of proper heat pipe operation. The working fluid must have good thermal stability properties at the specified operational temperature and pressure. The operational temperature range of the working fluid has to lie between its triple point and its critical point for liquid to exist in the wicking material. The **wettability** of the working fluid contributes to its capillary pumping and priming capability. High-surface-tension fluids are commonly used in heat pipes because they provide the capillary pumping and wetting characteristics necessary for proper operation. Other desirable thermophysical properties include a high liquid thermal conductivity, high latent heat of vaporization, low liquid viscosity, and a low vapor viscosity.

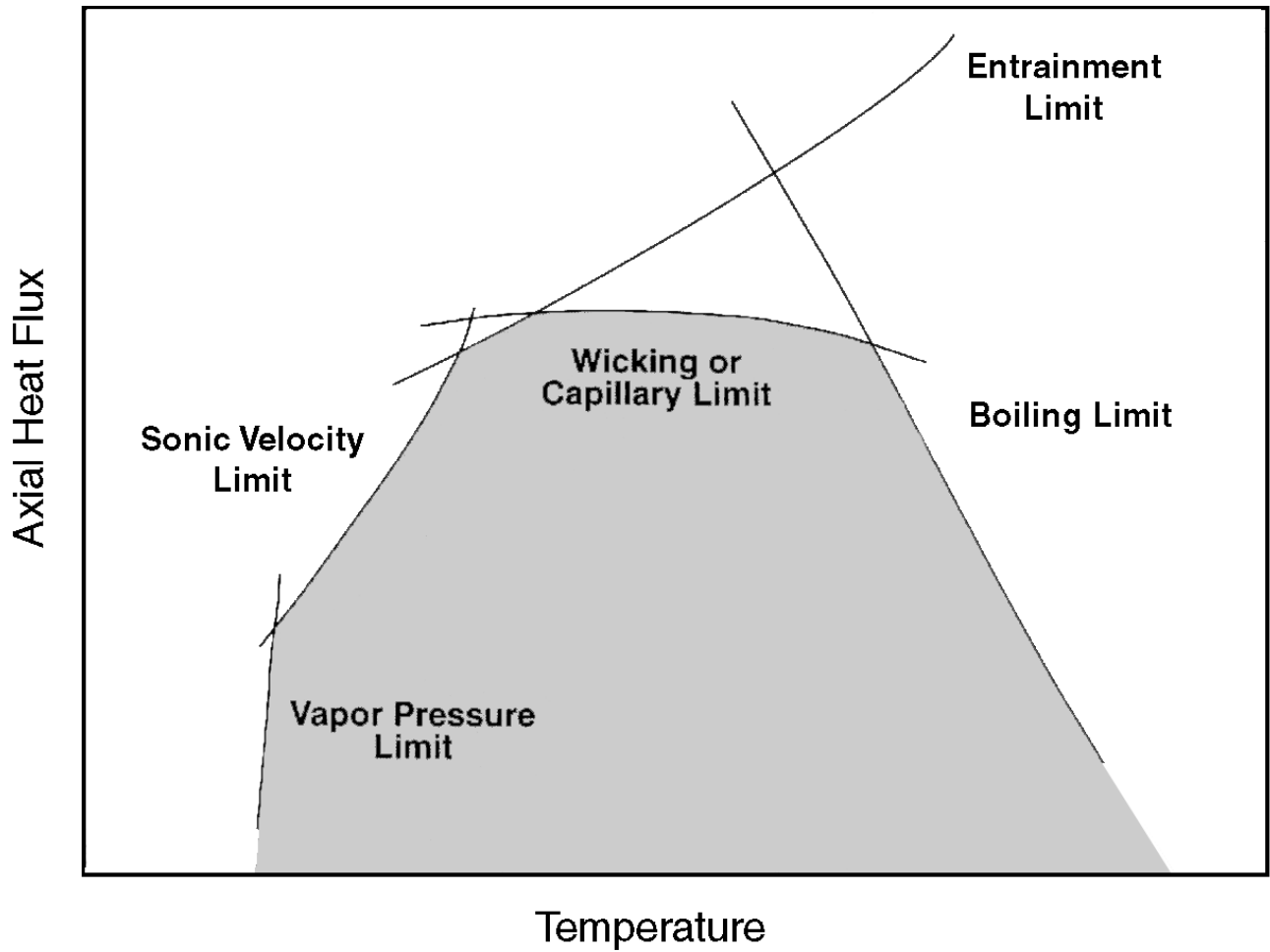
The wick structure and working fluid generate the capillary forces required to (a) pump liquid from the condenser to the evaporator and (b) keep liquid evenly distributed in the wicking material. Heat pipe wicks can be classified as either homogeneous wicks or composite wicks. Homogeneous wicks are composed of a single material and configuration. The most common types of homogeneous wicks include wrapped screen, sintered metal, axial groove, annular, crescent, and arterial. Composite wicks are composed of two or more materials and configurations. The most common types of composite wicks include variable screen mesh, screen-covered groove, screen slab with grooves, and screen tunnel with grooves. Regardless of the wick configuration, the desired material properties and structural characteristics of heat pipe wick structures are a high thermal conductivity, high wick porosity, small capillary radius, and high wick permeability. The container, wick structure, and working fluid are used to determine the heat transfer limitations of heat pipes.

## 55.2 Heat Transfer Limitations

---

Heat pipes undergo various heat transfer limitations depending on the working fluid, the wick structure, the dimensions of the heat pipe, and the heat pipe operational temperature. [Figure 55.2](#) gives a qualitative description of the various heat transfer limitations, which include vapor-pressure, sonic, entrainment, capillary, and boiling limitations. The composite curve enclosing the shaded region in [Fig. 55.2](#) gives the maximum heat transfer rate of the heat pipe as a function of the operational temperature. The figure shows that, as the operational temperature increases, the maximum heat transfer rate of the heat pipe is limited by various physical phenomena. As long as the operational heat transfer rate falls within the shaded region, the heat pipe will function properly.

**Figure 55.2** Heat transfer limitations in heat pipes.



The vapor-pressure limitation (or viscous limitation) in heat pipes develops when the pressure drop in the vapor core reaches the same order of magnitude as the vapor pressure in the evaporator. Under these conditions the pressure drop due to flow through the vapor core creates an extremely low vapor pressure in the condenser, preventing vapor from flowing into the condenser. A general expression for the vapor-pressure limitation is given by Dunn and Reay [1982]:

$$Q_{vp,max} = \frac{\pi r_v^4 h_{fg} \rho_{v,e} P_{v,e}}{12 \mu_{v,e} l_{eff}} \quad (55.1)$$

where  $r_v$  is the cross-sectional radius of the vapor core (m),  $h_{fg}$  is the latent heat of vaporization (J/kg),  $\rho_{v,e}$  is the vapor density in the evaporator ( $\text{kg/m}^3$ ),  $P_{v,e}$  is the vapor pressure in the evaporator (Pa), and  $\mu_{v,e}$  is the vapor viscosity in the evaporator ( $\text{N s/m}^2$ ). The value  $l_{eff}$  is the effective length of the heat pipe (m), equal to  $l_{eff} = 0.5(l_e + 2l_a + l_c)$ . The vapor-pressure limitation can occur during the start-up of heat pipes at the lower end of the working fluid temperature range.

The sonic limitation also can occur in heat pipes during start-up at low temperatures. The low temperature produces a low vapor density, thereby reducing the speed of sound in the vapor core. Thus, a sufficiently high mass flow rate in the vapor core can cause sonic flow conditions and generate a shock wave that chokes the flow and restricts the pipe's ability to transfer heat to the condenser. Dunn and Reay [1982] give an expression for the sonic limitation that agrees very well with experimental data:

$$Q_{s,\max} = 0.474 A_v h_{fg} (\rho_v P_v)^{1/2} \quad (55.2)$$

where  $A_v$  is the cross-sectional area of the vapor core ( $\text{m}^2$ ). The sonic limitation should be avoided because large temperature gradients occur in heat pipes under choked-flow conditions.

The entrainment limitation in heat pipes develops when the vapor mass flow is large enough to shear droplets of liquid off the wick surface, causing dry-out in the evaporator. A conservative estimate of the maximum heat transfer rate due to entrainment of liquid droplets has been given by Dunn and Reay [1976] as

$$Q_{e,\max} = A_v h_{fg} \left[ \frac{\rho_v \sigma_l}{2 r_{c,\text{ave}}} \right]^{1/2} \quad (55.3)$$

where  $\sigma_l$  is the surface tension ( $\text{N/m}$ ) and  $r_{c,\text{ave}}$  is the average capillary radius of the wick ( $\text{m}$ ).

The capillary limitation in heat pipes occurs when the net capillary force generated by the vapor-liquid interfaces in the evaporator and condenser are not large enough to overcome the frictional pressure losses due to fluid motion. This causes the heat pipe evaporator to dry out and shuts down the transfer of heat from the evaporator to the condenser. For most heat pipes the maximum heat transfer rate due to the capillary limitation can be expressed as [Chi, 1976]

$$Q_{c,\max} = \left[ \frac{\rho_l \sigma_l h_{fg}}{\mu_l} \right] \left[ \frac{A_w K}{l_{\text{eff}}} \right] \left( \frac{2}{r_{c,e}} - \left[ \frac{\rho_l}{\sigma_l} \right] g L_t \cos \psi \right) \quad (55.4)$$

where  $K$  is the wick permeability ( $\text{m}^2$ ),  $A_w$  is the wick cross-sectional area ( $\text{m}^2$ ),  $\rho_l$  is the liquid density ( $\text{m}^3$ ),  $\mu_l$  is the liquid viscosity ( $\text{Ns/m}^2$ ),  $r_{c,e}$  is the wick capillary radius in the evaporator ( $\text{m}$ ),  $g$  is the acceleration due to gravity ( $9.8 \text{ m/s}^2$ ), and  $L_t$  is the total length of the pipe ( $\text{m}$ ). For most practical operating conditions, this limitation can be used to determine maximum heat transfer rate in heat pipes.

The boiling limitation in heat pipes occurs when the degree of liquid superheat in the evaporator is large enough to cause the nucleation of vapor bubbles on the surface of the wick or the container. Boiling is usually undesirable in heat pipes because local hot spots can develop in the wick, obstructing the flow of liquid in the evaporator. An expression for the boiling limitation is [Chi, 1976]

$$Q_{b,\max} = \frac{2\pi L_{\text{eff}} k_{\text{eff}} T_v}{h_{fg} \rho_l \ln(r_i/r_v)} \left( \frac{2\sigma_l}{r_n} - \frac{2\sigma_l}{r_{c,e}} \right) \quad (55.5)$$

where  $k_{\text{eff}}$  is the effective thermal conductivity of the composite wick and working fluid ( $\text{W/m K}$ ),  $T_v$  is the vapor saturation temperature ( $\text{K}$ ),  $r_i$  is the inner container radius ( $\text{m}$ ), and  $r_n$  is the nucleation radius (equal to  $2.00 \cdot 10^{-6}$  m in the absence of noncondensable gas).

## 55.3 Effective Thermal Conductivity and Heat Pipe Temperature Difference

---

One key attribute of the heat pipe is that it can transfer a large amount of heat while maintaining nearly isothermal conditions. The temperature difference between the external surfaces of the evaporator and the condenser can be determined from the following expression,

$$\Delta T = R_t Q \quad (55.6)$$

where  $R_t$  is the total thermal resistance ( $\text{K/W}$ ) and  $Q$  is the heat transfer rate ( $\text{W}$ ). [Figure 55.3](#) shows the thermal resistance network for a typical heat pipe and the associated thermal resistances. In most cases the total thermal resistance can be approximated by

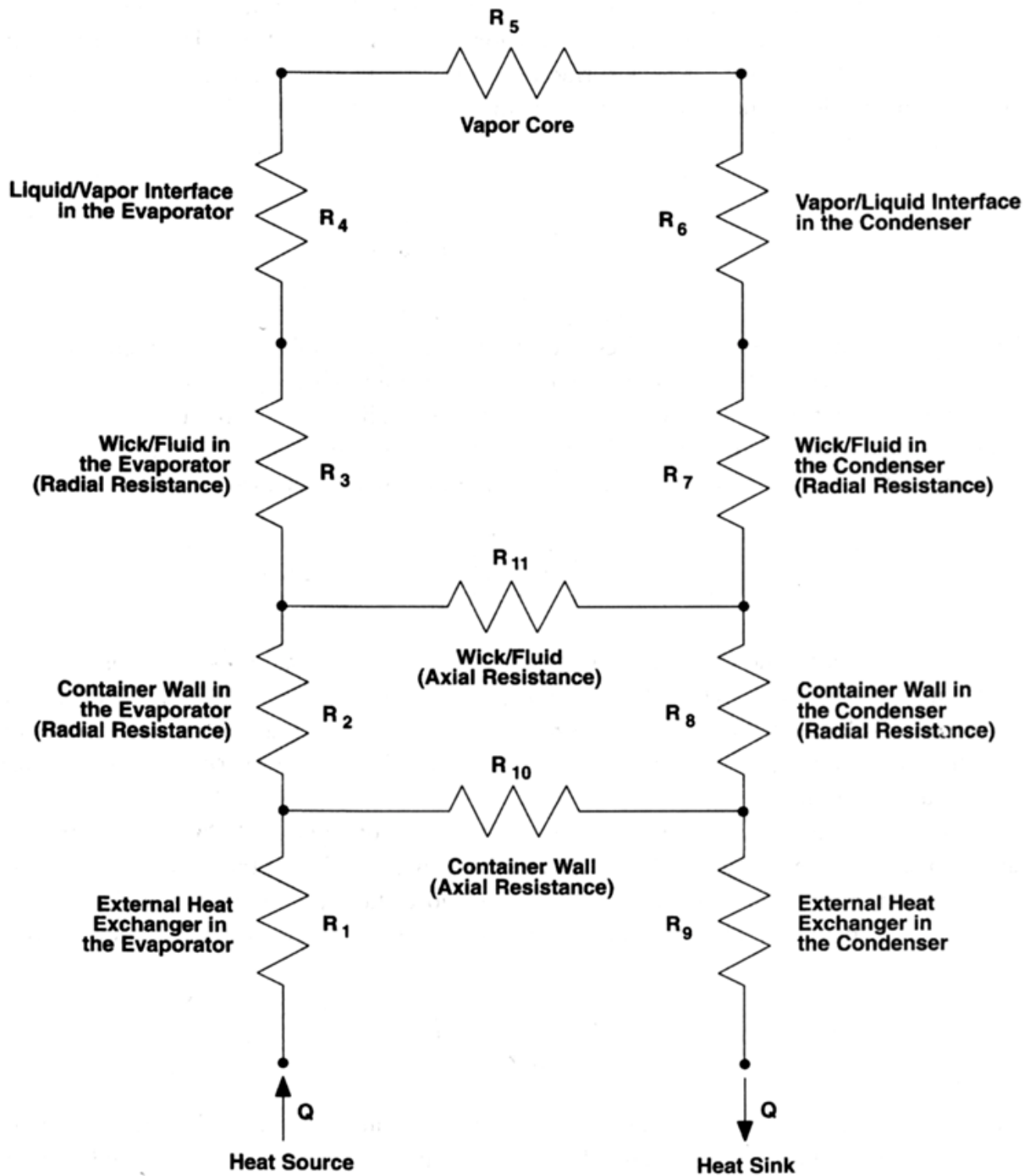
$$R_t = R_1 + R_2 + R_3 + R_5 + R_7 + R_8 + R_9 \quad (55.7)$$

The effective thermal conductivity of the heat pipe is defined as the heat transfer rate divided by the temperature difference between the heat source and heat sink,

$$k_{\text{eff}} = \frac{L_t}{R_t A_t} \quad (55.8)$$

where  $A_t$  is the overall cross-sectional area of the pipe ( $\text{m}^2$ ). Under normal operating conditions, the total thermal resistance is relatively small, making the external surface temperature in the evaporator approximately equal to that in the condenser. Thus, the effective thermal conductivity in a heat pipe can be very large (at least an order of magnitude larger than that of aluminum).

**Figure 55.3** Thermal resistance network in a heat pipe.



## 55.4 Application of Heat Pipes

---

Heat pipes have been applied to a wide variety of thermal processes and technologies. It would be an impossible task to list all of the applications of heat pipes; therefore, only a few important industrial applications are given here. In the aerospace industry, heat pipes have been used successfully in controlling the temperature of vehicles, instruments, and space suits. Cryogenic heat pipes have been applied in (a) the electronics industry for cooling various devices (e.g., infrared sensors, parametric amplifiers), and (b) the medical field for cryogenic eye and tumor surgery. Heat pipes have been employed to keep the Alaskan tundra frozen below the Alaskan pipeline. Other cooling applications include (1) turbine blades, generators, and motors; (2) nuclear and isotope reactors; and (3) heat collection from exhaust gases, and solar and geothermal energy.

In general, heat pipes have advantages over many traditional heat-exchange devices when (a) heat has to be transferred isothermally over relatively short distances, (b) low weight is essential (the heat pipe is a passive pumping device and therefore does not require a pump), (c) fast thermal-response times are required, and (d) low maintenance is mandatory.

### Defining Terms

**Capillary force:** The force caused by a curved vapor-liquid interface. The interfacial curvature is dependent on the surface tension of the liquid, the contact angle between the liquid wick structure, the vapor pressure, and the liquid pressure.

**Effective thermal conductivity:** The heat transfer rate divided by the temperature difference between the evaporator and condenser outer surfaces.

**Heat transfer limitations:** Limitations on the axial heat transfer capacity imposed by different physical phenomena (e.g., vapor-pressure, sonic, entrainment, capillary, and boiling limitations).

**Wettability:** The ability of a liquid to spread itself over a surface. A wetting liquid spreads over a surface, whereas a nonwetting liquid forms droplets on a surface.

**Wick:** A porous material used to generate the capillary forces that circulate fluid in a heat pipe.

### References

- Chi, S. W. 1976. *Heat Pipe Theory and Practice*. Hemisphere, Washington, DC.  
Dunn, P. D. and Reay, D. A. 1982. *Heat Pipes*, 3rd ed. Pergamon Press, Oxford, UK.

### Further Information

Recent developments in heat pipe research and technology can be found in the proceedings from a number of technical conferences: (1) the International Heat Pipe Conference, (2) the National Heat Transfer Conference, (3) the ASME Winter Annual Meeting, and (4) the AIAA Thermophysics Conference.

Books particularly useful for the design of heat pipes include (1) *Heat Pipe Design Handbook* by Brennan and Kroliczek, available from B&K Engineering in Maryland; (2) *The Heat Pipe* by Chisholm, available from Mills and Boon Limited in England; and (3) *Heat Pipes: Construction*

*and Application* by Terpstra and Van Veen, available from Elsevier Applied Science in New York.

An additional book particularly strong in heat pipe theory is *The Principles of Heat Pipes* by Ivanovskii, Sorokin, and Yagodkin, available from Clarendon Press in England.

Any further questions can be addressed to the author at (409)260-6230 (phone) or 409.260.6249 (fax).



Timmerhaus, K. "Separation Processes"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



This night scene is of the Shell Oil higher olefins refinery in Geismar, Louisiana. The refinery is Shell's second world-scale higher olefins plant utilizing their proprietary technology. Higher olefins are versatile chemical intermediates used in the manufacture of detergents, lubricants, and a wide range of other products. (Photo courtesy of Shell Oil Company.)

# VIII

## Separation Processes

---

**Klaus Timmerhaus**

*University of Colorado*

- 56 **Distillation** *J. R. Fair*  
Separation Specification • Required Basic Data • Index of Separation Difficulty • Required Stages • Column Dimensions • Column Auxiliaries • Batch Distillation
- 57 **Absorption and Stripping** *W. M. Edwards and J. R. Fair*  
Absorber-Stripper Systems • Absorber-Stripper Design Diagrams • Key Design Assumptions • Physical Data Requirements • Absorber and Stripper Design Equations • Absorption and Stripping Efficiency
- 58 **Extraction** *V. Van Brunt*  
Representative Extraction Processes • Solvent Characteristics and Solvent Screening • Extraction Equilibria • Extraction Staging and Equipment
- 59 **Adsorption** *S. Sircar*  
Adsorbent Materials • Adsorption Equilibria • Heat of Adsorption • Thermodynamic Selectivity of Adsorption • Adsorption Kinetics • Adsorption Column Dynamics • Adsorptive Separation Processes and Design
- 60 **Crystallization and Evaporation** *R. C. Bennett*  
Methods of Creating Supersaturation • Reasons for the Use of Crystallization • Solubility Relations • Product Characteristics • Impurities Influencing the Product • Kinds of Crystallization Processes • Calculation of Yield in a Crystallization Process • Mathematical Models of Continuous Crystallization • Equipment Designs • Evaporation
- 61 **Membrane Separation** *D. G. Woods, D. R. B. Walker, and W. J. Koros*  
Dialysis • Reverse Osmosis • Gas and Vapor Separations • Asymmetric Membranes • Membrane Stability and Fouling • Module Design Considerations
- 62 **Fluid-Particle Separation** *S.-H. Chiang and D. He*  
Equipment • Fundamental Concept • Design Principles • Economics
- 63 **Other Separation Processes** *W. C. Corder and S. P. Hanson*  
Sublimation • Diffusional Separations • Adsorptive Bubble Separation • Dielectrophoresis • Electrodialysis

SEPARATION TECHNOLOGY IS A KEY CONCEPT used in the chemical, petroleum, petrochemical, rubber, pulp, pharmaceutical, food, mineral, and other, related process industries. It is also fundamental to most aspects of environmental control. Since separation technology generally requires a major investment in both capital and operating costs, the impact on corporate profitability is considerable in those processes requiring such technology. Similarly, the growth of new industries based, for example, on biotechnology and electronics will require the use of proven separation technology or the development of additional separation techniques.

Recent progress in the understanding of phase-equilibrium thermodynamics and transport phenomena, as well as the development of new equipment and the need for products of higher

purity, has greatly augmented and diversified the field of separation technology. New contacting devices that minimize pressure drops and maximize the reliability of scale-up, new membrane systems that can be used in a variety of old and new applications, new adsorbents that improve the adsorption selectivity, and new solvents for use in extraction and absorption are among the more recent advances that have led to changes in the traditionally accepted methods for innovating, designing, and operating separation processes. This traditional approach in the past decade has also been modified to include the impact of chemical processing on the environment, a sensitivity to the hazards associated with the production of dangerous and toxic materials, and a growing awareness of the limited availability of many raw materials. In addition, the advent of the computer has substantially altered the methods by which separation processes are conceived, designed, and optimized.

Coverage in Section VIII begins with distillation (**Chapter 56**), the most widely used industrial method for separating liquid mixtures. Even though this separation process is one of the largest consumers of energy, it is still the separation choice in the chemical and petrochemical industries because of its apparent simplicity. Absorption and stripping as outlined in **Chapter 57** have found useful application in a number of industries, particularly in natural and synthetic gas processing, but also in environmentally important scrubbing operations. In the latter, for example, strippers are employed for removing environmentally sensitive trace components from liquid streams.

General features of separation by extraction are covered in **Chapter 58**. This process is an indirect separation that relies on the ease of separating a chemical from a solvent rather than from its original feed. In such cases, extraction could provide both capital and operating cost savings compared to the conventional separation approach of distillation. **Chapter 59** discusses the principles and applications of adsorption in the chemical, petrochemical, biochemical, and environmental industry for the separation and purification of fluid mixtures. Various types of adsorbent materials as well as information on equilibrium data are covered in this review.

The use of crystallization to obtain high-purity solids is covered in **Chapter 60**. Its wide use is attributed to the highly purified and usable form in which the product can be obtained from relatively impure solutions, often by a single processing step. Separation by dialysis or reverse osmosis, on the other hand, operates on a molecular level to obtain selective passage of one or more of the components in the feed stream. Recent advances in membrane-based separation has made it one of the most rapidly growing areas in process technology. Details are outlined in **Chapter 61**.

Fluid-particle separation, discussed in **Chapter 62**, involves the removal and collection of matter or particles in a dispersed or colloidal state in suspension. Areas of fluid-particle separation considered in the discussion include screening, thickening/sedimentation, filtration, cycloning, centrifugation, flotation, and membrane filtration. Their analysis generally requires a combined knowledge of fluid mechanics, particle mechanics, solution chemistry, and surface/interface chemistry.

Several other separation technologies are summarized in **Chapter 63**. Here emphasis has been confined to sublimation, diffusional separation, adsorptive-bubble separation, dielectrophoresis, and electrodialysis. Generally, such separation technologies are used only when more traditional separation techniques involve higher capital and operating costs or lower levels of separation performance.

Fair, J. R. "Distillation"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

- 56.1 Separation Specification
- 56.2 Required Basic Data
- 56.3 Index of Separation Difficulty
- 56.4 Required Stages
- 56.5 Column Dimensions
- 56.6 Column Auxiliaries
- 56.7 Batch Distillation

**James R. Fair**

*University of Texas, Austin*

Distillation is a method of separation that is based on the difference in composition between a liquid mixture and the vapor formed from it. The composition difference is due to differing effective vapor pressures, or volatilities, of the components of the liquid mixture. When such a difference does not exist, as at an azeotropic point, separation by distillation is not possible. Distillation as normally practiced involves condensation of the vaporized material, usually in multiple vaporization/condensation operations. It thus differs from evaporation, which is usually applied to separate a liquid from a solid, but which can be applied to simple liquid concentration operations. In the chemical and petroleum industries, distillation is one of the largest consumers of energy among the several processing operations employed.

Distillation is the most widely used industrial method of separating liquid mixtures and is at the heart of the separation processes in many chemical and petroleum plants. The most elementary form of the method is simple distillation, in which the liquid is brought to boiling and the vapor formed is separated and condensed to form a product. In some cases, a normally gaseous mixture is condensed by using refrigeration to obtain a liquid that is amenable to distillation separation—the separation of air is a prime example. If the process is continuous with respect to feed and product flows, it is called flash distillation. If the feed mixture is available as an isolated batch of material, the process is a form of **batch distillation** and the compositions of the collected vapor and residual liquid are thus time-dependent. The term *fractional distillation* (which may be contracted to **fractionation**) was originally applied to the collection of separate fractions of condensed vapor, each fraction being segregated. Currently, the term is applied to distillation processes in general, where an effort is made to separate an original mixture into several components by means of distillation. When the vapors are enriched by contact with counterflowing liquid **reflux**, the process is called rectification. When fractional distillation is accomplished with a continuous feed of material and continuous removal of product fractions, the



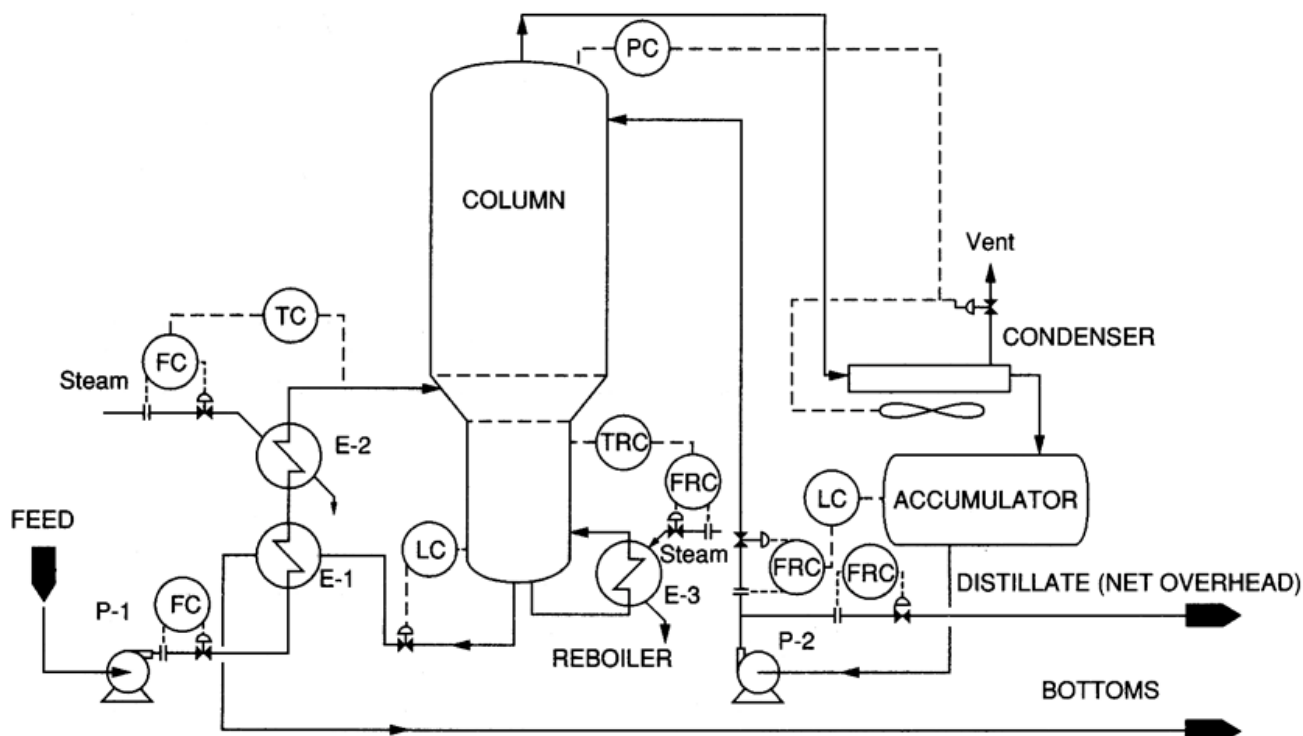
process is called **continuous distillation**. When steam is added to the vapors to reduce the partial pressures of the components to be separated, the term *steam distillation* is used.

Most distillations conducted commercially operate continuously, with a more volatile fraction recovered as **distillate** and a less volatile fraction recovered as **bottoms** (or residue). If a portion of the distillate is condensed and returned to the process to enrich the vapors, the liquid is called **reflux**. The apparatus in which the enrichment occurs is usually a vertical, cylindrical vessel called a still or distillation column. This apparatus normally contains internal devices for effecting vapor-liquid contact. The devices may be categorized as plates or packings. In some cases a **batch distillation** is preferred, where the feed charge is placed in a stillpot and heat is added. The vaporized mixture is condensed and analyzed. Clearly, this method of operation is in a time-variant domain, unlike continuous distillation.

Distillation has been practiced in one form or another for centuries. It was of fundamental importance to the alchemists and has been in use for more than 2000 years. Because it is a process involving vaporization of a liquid, energy must be supplied if it is to function.

A representative continuous distillation system is shown in Fig. 56.1. The design of such a system, or the analysis of an existing system, follows a fixed procedure. The steps in this procedure are described in the following sections, with primary emphasis on the distillation column. Such a procedure also provides an overall summary of the technology associated with distillation column design and operation.

**Figure 56.1** Flow diagram of a representative distillation system. For the case shown, the column has two diameters, commensurate with the changing flows of liquid and vapor in the column. Many columns have a single diameter. The condenser is air-cooled, but water is also often used as a coolant.



## 56.1 Separation Specification

---

If a mixture is to be separated by distillation, it is crucial that the mixture composition be defined carefully. Serious problems can arise when some components of the mixture are disregarded, or not known to exist. Next, the degree of required separation must be defined. Often there is a key component of the feed to which the separation specification can be related. The specification may deal with composition or recovery, or both. For example, for a benzene/toluene feed mixture, the specification may be a minimum of 95% of the benzene in the feed to be taken overhead with the distillate (the percent recovery), and the minimum purity of benzene in the distillate to be 99.5%. Simply stated, one must determine in advance the degree of separation desired.

## 56.2 Required Basic Data

---

A key parameter that is related to the ease (or difficulty) of making the separation is known as the **relative volatility**. This is a ratio of the effective vapor pressures of the components to be separated. It might be thought of as an index of the difficulty of making the separation. It is determined from the following relationship:

$$\alpha_{AB} = \frac{(\gamma_A^L P_A^o)}{(\gamma_B^L P_B^o)} \quad (56.1)$$

where components  $A$  and  $B$  are those between which the specification of separation is written,  $\gamma^L$  is a thermodynamic parameter called the liquid phase activity coefficient, and  $P^o$  is the vapor pressure. In Eq. (56.1), component  $A$  is more volatile than component  $B$  (i.e.,  $A$  is more likely to be distilled overhead than is  $B$ ). When liquid mixtures behave ideally (no excess energy effects when the mixture is formed), the activity coefficients have a value of 1.0. The vapor pressures of compounds are readily available from various handbooks. When the key components have a relative volatility of 1.10 or less, separation by distillation is difficult and quite expensive. For cases where  $\alpha_{AB}$  is 1.50 or greater, the separation is relatively easy.

The determination of relative volatility involves the area of solution thermodynamics, and for distillation a subarea is called vapor-liquid equilibrium (VLE). There are many approaches to determining VLE: direct measurement, correlation of the data of others, extrapolation of correlations beyond the areas of measurement, and prediction of equilibrium data when no measurements have been made.

## 56.3 Index of Separation Difficulty

---

After the needed equilibrium data have been obtained, the next step is to determine the difficulty (or ease) of separation, using some index that is reliable and generally understood. The most used index is based on the required number of **theoretical stages**. For rigorous calculations, the stage count results from a detailed analysis of flow and composition changes throughout the column. For a simplified analysis, a basic parameter is the minimum number of theoretical stages:

$$N_{\min} = \frac{\ln[(y_i/y_j)^D (x_j/x_i)^B]}{\ln \alpha_{ij,\text{avg}}} \quad (56.2)$$



where  $i$  and  $j$  are the components between which the specified separation is to be made. The value of  $N_{\min}$  can serve as a general criterion of separation difficulty. For example, if the key components  $i$  and  $j$  have close boiling points and a relative volatility of 1.40, and if  $i$  is to have a purity of 99.0% minimum in the distillate and 3.0% maximum in the bottoms, and  $j$  is to have a maximum purity of 4.0% in the distillate and a minimum purity of 99.5% in the bottoms, then

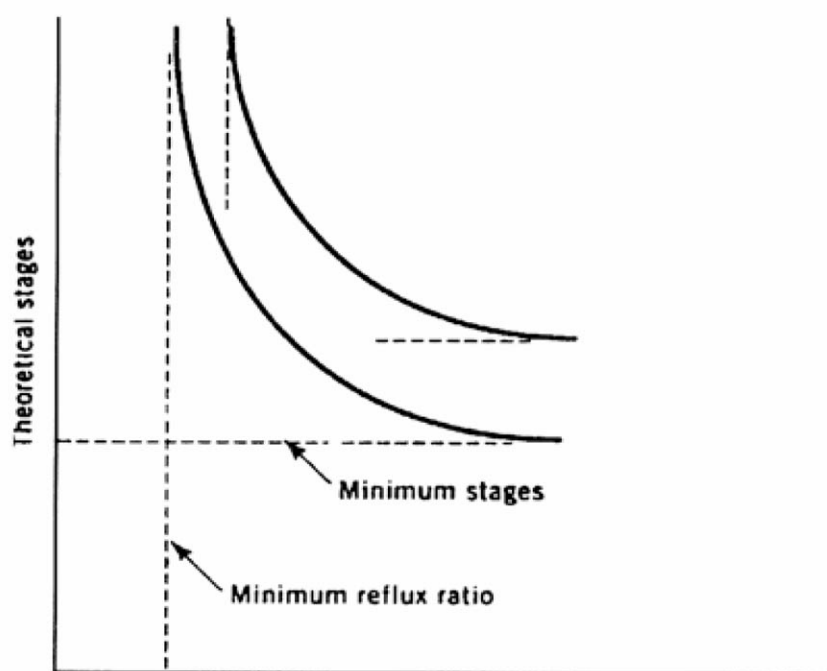
$$N_{\min} = \frac{\ln[(99.0/4.0)/(3.0/99.5)]}{\ln 1.40} = 19.9 \text{ stages} \quad (56.3)$$

As a very rough rule of thumb, the actual number of theoretical stages required will be about twice the minimum number of stages. Equation (56.3) is suitable for making quick comparisons between the design conditions of overhead and bottoms compositions, and of relative volatility.

## 56.4 Required Stages

The distillation column is a vertical, cylindrical vessel into which are placed contacting devices that cause the rising vapor and descending liquid to come into intimate contact. Such contact is necessary if the column is to make an efficient separation. As indicated in Fig. 56.1, a return of some of the distillate product is necessary for the column to work efficiently. This flow stream is called the reflux. The ratio of this reflux flow to the flow of distillate is called the reflux ratio. The designer has the option of varying this ratio, and in so doing can influence the required number of theoretical stages, as shown in Fig. 56.2. The asymptote for stages is  $N_{\min}$  as described above. The asymptote for reflux ratio is the minimum reflux ratio  $R_{\min}$ , a parameter related to the design compositions and the relative volatility. There is some optimum reflux ratio that relates to the economic conditions that exist for the design. Often the optimum reflux ratio is taken as about 1.3 times the minimum reflux ratio.

**Figure 56.2** General relationship between stages and reflux. A given curve represents a specific separation requirement.



The actual number of stages required for the column is based on stage efficiency:

$$N_{\text{act}} = N_t / E_{\text{oc}} \quad (56.4)$$

where  $N_t$ , the number of theoretical stages, can be computed from

$$\frac{N_t - N_{\min}}{N_t + 1} = 0.75 - 0.75 \left( \frac{R - R_{\min}}{R + 1} \right)^{0.5668} \quad (56.5)$$

In Eq. (56.5),  $N_{\min}$  and  $R_{\min}$  are taken from a plot such as Fig. 56.2 or, more likely, from rigorous calculations using modern computer programs. The reflux ratio  $R$  is a design variable, as mentioned earlier. The value of the overall efficiency  $E_{\text{oc}}$  for use in Eq. (56.4) must be obtained from experience or from predictive mathematical models. It is often in the range of 0.60 to 0.80 (60 to 80%).

## 56.5 Column Dimensions

---

The height of the column is a function of  $N_{\text{act}}$ , the number of actual stages required for the separation. The height of each stage is a function of the geometry of the contacting device that is used. If the device is represented by a flat, horizontal sheet metal "plate" containing perforations, then the vertical spacing of these plates will determine the required height of the column. For example, if 40 theoretical stages are needed, and the efficiency of such a plate is 80%, then 50 actual stages (plates) are required. If these plates must be spaced at 0.61 m (24 in.) intervals, then the column must be at least 30 m (98 ft) high to accommodate the vapor and liquid flows. Considering other needs for internal devices, the column might have an actual height of perhaps 35 m.

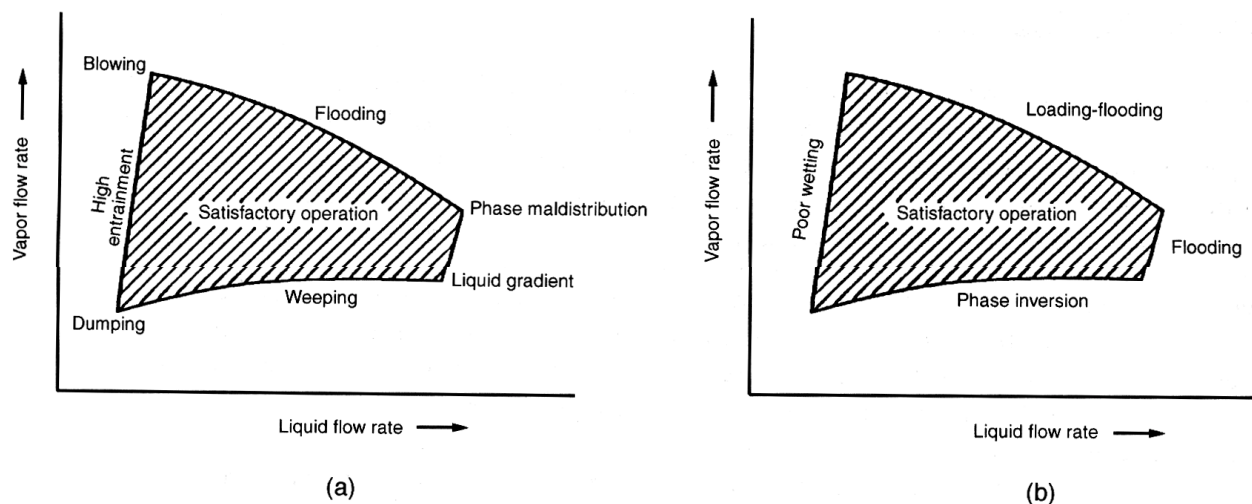
If, instead of plates, a type of internal device called *packing* is used, some equivalent dimension must be used. The required height is obtained from

$$Z = (N_t)(HETP) \quad (56.6)$$

where  $HETP$  is the height equivalent to a theoretical plate (stage) and its value often must be obtained from the manufacturer or vendor of the packing device. There are many different sizes and shapes of packing used commercially.

If the height of the column is determined from the needed stages, the diameter must be determined from the required flow rates of vapor and liquid in the column. These rates are based on the total amount of feed mixture to be processed, as well as the amount of reflux that is returned to the column. In turn, the device (plate or packing) selected for contacting must have the capacity to handle the required flow rates. A schematic diagram of the factors that influence column diameter is shown as Fig. 56.3. The device must be able to handle the vapor flow without tending to carry the liquid upward. Likewise, the device must be able to handle the liquid flows without choking the column because of a buildup of liquid. Correlations are available for assigning numbers to the scales of Fig. 56.3, and an optimum design will place the design condition well within the shaded area.

**Figure 56.3** Generalized operating diagram for a distillation column, showing ranges of operation as a function of relative vapor and liquid flows. (a) Plate column. (b) Packed column.



## 56.6 Column Auxiliaries

Figure 56.1 shows that a distillation system contains more than just the column. Of the several auxiliary components, two heat exchangers are of particular importance—the reboiler and the condenser. The reboiler at the base receives the net heat input to the column (often in the form of condensing low pressure steam) and establishes the amount of vapor to be handled by the column. Much of the heat added at the reboiler is removed in the overhead condenser, rejecting the heat directly to the atmosphere (as shown in Fig. 56.1) or to a stream of cooling water. While the heat duties of the condenser and reboiler are often about the same, the overall heat balance depends also on the heat contents of the streams entering and leaving the column and on the amount of heat losses to the surroundings.

## 56.7 Batch Distillation

Although most commercial distillations are run continuously, there are certain applications where batch distillation is the method of choice. For this method, a charge, or batch, of the initial mixture is placed in a vessel, where it can be heated and distilled over a period of time. Compositions of the charge and product thus vary with time, which is not the case for continuous distillation. Particular cases where batch distillation may be preferred are (1) semiworks operations producing interim amounts of product in equipment that is used for multiple purposes, (2) distillations of specialty chemicals where contamination can be a problem (the batch equipment can be cleaned or sterilized between batch runs), (3) operations involving wide swings in feed compositions and product specifications, where batch operating conditions can be adjusted to meet the varying needs, and (4) laboratory distillations where separability is being investigated without concern over the scale-up to commercial continuous operations.

Batch distillations are generally more expensive than their continuous counterparts, in terms of cost per unit of product. Close supervision and computer control are required, the equipment is

more complex if several products are to be recovered, and total throughput is limited by the needs for changing operating modes and for recharging the system with feed material.

## Defining Terms

**Batch distillation:** A distillation operation in which the feed mixture is charged to a vessel, heated to boiling, and the vapor condensed into different fractions over a period of time (the batch cycle).

**Bottoms:** The product of the distillation that is relatively nonvolatile and that flows from the bottom of the column (also called *residue*).

**Continuous distillation:** A distillation operation in which the feed mixture is charged continuously to the distillation column, with the products withdrawn continuously with invariant compositions.

**Distillate:** The product of the distillation that is relatively volatile and that flows from the top of the column (also called *net overhead*).

**Feed:** The mixture to be distilled.

**Fractionation:** A contraction of *fractional distillation*, a term used loosely to denote a distillation operation that provides two or more products (fractions), normally by means of reflux and a plurality of theoretical stages.

**Reflux:** Liquid that is derived from the distillate product and fed back to the top of the column for the purpose of enhancing the separation.

**Relative volatility:** A ratio of effective vapor pressures of the components of a mixture to be distilled. An important index of ease of separation.

**Theoretical stage:** A conceptual term which refers to the "ideal" situation in which the vapor leaving a defined section of a column is in thermodynamic equilibrium with the liquid leaving the same section.

## Reference

Fair, J. R. 1984. Liquid-gas systems. In *Perry's Chemical Engineers' Handbook*, 6th ed., ed. R. H. Perry and D. Green. McGraw-Hill, New York.

## Further Information

Kister, H. Z. 1992. *Distillation—Design*. McGraw-Hill, New York.

Rousseau, R. W. (Ed.) 1987. *Handbook of Separation Process Technology*. John Wiley & Sons, New York.

Schweitzer, P. A. (Ed.) 1988. *Handbook of Separation Techniques for Chemical Engineers*, 2nd ed. McGraw-Hill, New York.

Seader, J. D. 1984. Distillation. In *Perry's Chemical Engineer's Handbook*, 6th ed., ed. R. H. Perry and D. Green. McGraw-Hill, New York.

Walas, S. M. 1985. *Phase Equilibria in Chemical Engineering*. Butterworths, Reading, MA.

Edwards, W. M., Fair, J. R. "Absorption and Stripping"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Absorption and Stripping

---

- 57.1 Absorber-Stripper Systems
- 57.2 Absorber-Stripper Design Diagrams
- 57.3 Key Design Assumptions
- 57.4 Physical Data Requirements
- 57.5 Absorber and Stripper Design Equations
- 57.6 Absorption and Stripping Efficiency

**William M. Edwards**

*Consultant*

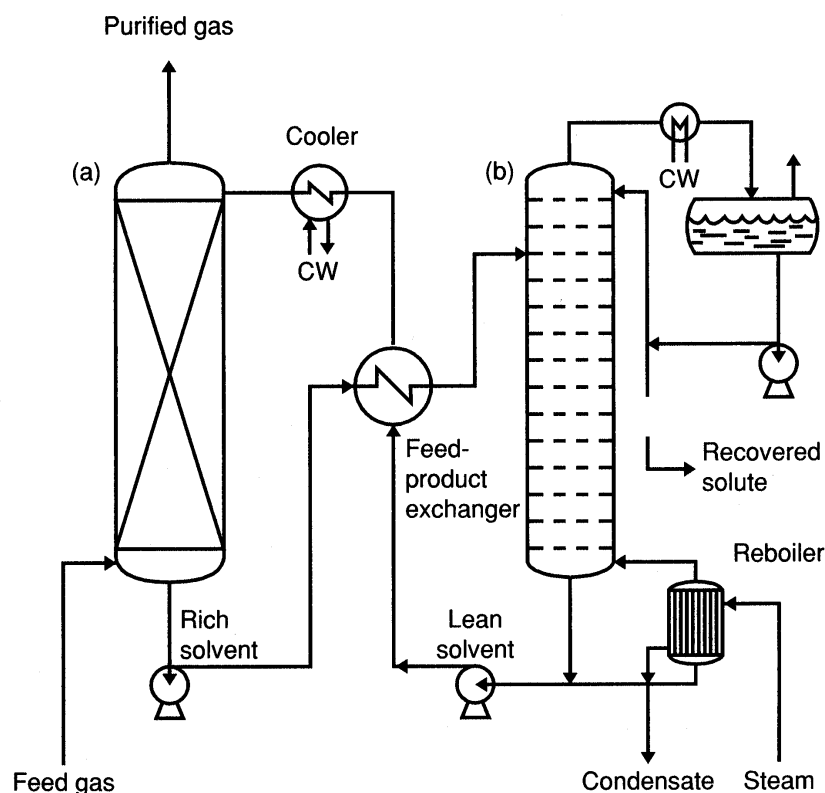
**James R. Fair**

*University of Texas*

**Absorption** and **stripping** are two chemical process operations that normally are employed in tandem for removing a minor component (**solute**) from an incoming process gas stream and for subsequently recovering that same component in a more concentrated form (see [Fig. 57.1](#)).

**Absorbers** often are employed for removing environmentally sensitive trace components from air or other gas streams. **Strippers** are employed for removing such components from liquid streams. Examples of each include the absorption of xylenes from air using a nonvolatile solvent and the stripping of volatile organic compounds (VOCs) from water with air.

**Figure 57.1** Absorber-stripper system: (a) absorber, (b) stripper.



## 57.1 Absorber-Stripper Systems

A schematic of an absorber-stripper operation is shown in [Fig. 57.1](#). The solute-containing feed gas enters the bottom of the **absorption tower**, where it is contacted by the **solvent**. The solvent flows countercurrently downward through gas-liquid contacting devices, which may be either packing materials or cross-flow trays. The intimate contacting of gas and liquid permits transfer of the solute from the gas phase to the liquid (solvent) phase, thereby effecting a purification of the gas.

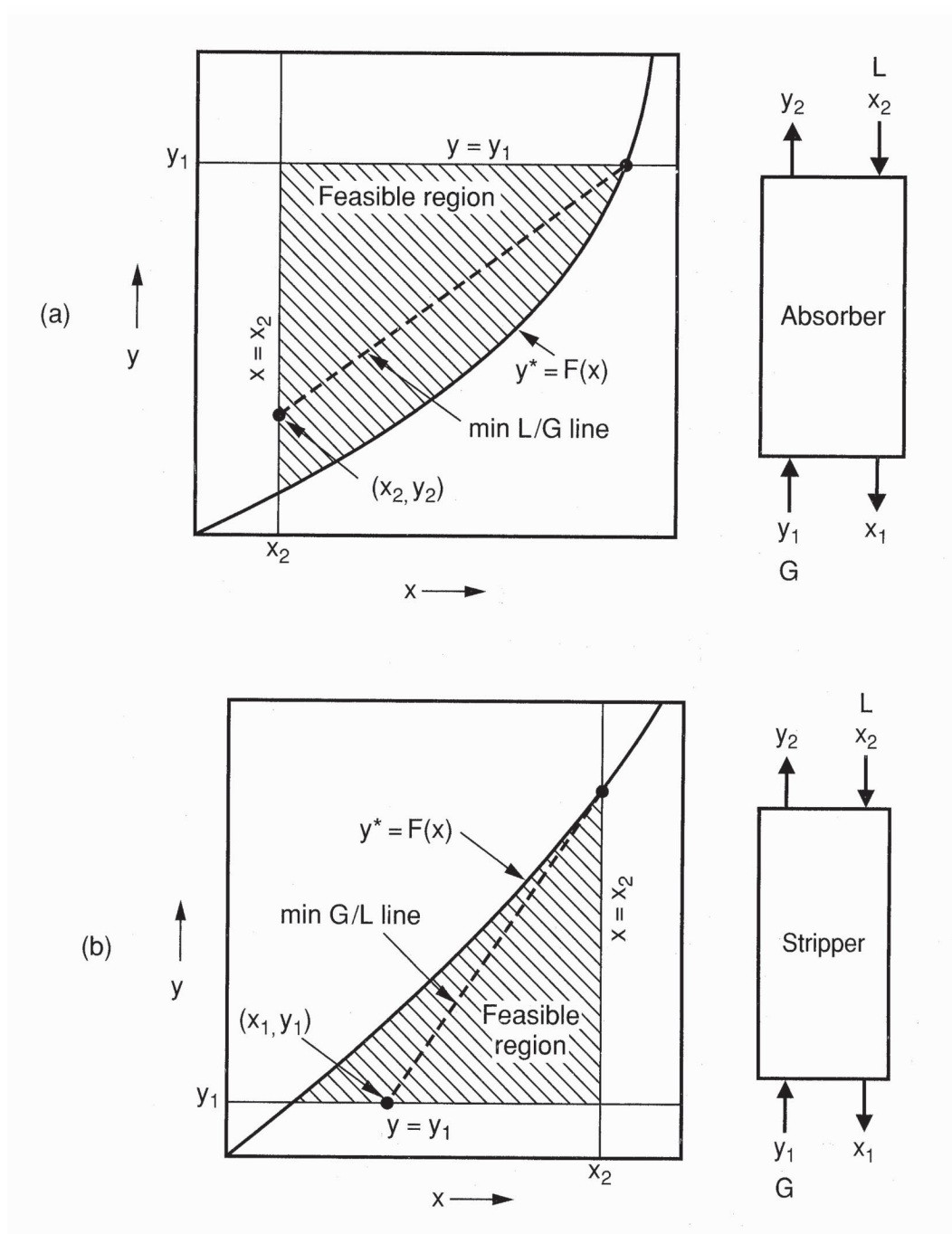
The solute-containing solvent leaves the bottom of the absorption tower and is pumped to the top of the **stripping tower**, where it flows downward over contacting devices under conditions permitting transfer of the solute back to a gas phase that differs from that present in the absorber. The stripped solvent is then recycled back to the absorber for reuse. However, a portion of the downflowing solvent is vaporized in a heat exchanger (the *reboiler*) to provide a stripping vapor, which serves to remove the solute from the downflowing solvent. Most of the stripped (nearly solute-free) solvent is recycled back to the absorber for reuse.

The overhead solvent and solute vapors leaving the stripper are condensed or otherwise removed and can be sent to storage or processed further. The concentration of the solute in this effluent stream is relatively high by now, making it easier to process in downstream units.

## 57.2 Absorber-Stripper Design Diagrams

Example design diagrams for an absorber-stripper combination are shown in Figs. 57.2(a) and 57.2(b). In these diagrams  $y$  is the concentration of solute in the gas (or vapor) phase and  $x$  is the concentration of solute in the solvent liquid phase. These concentrations are expressed as mole fractions. The subscript "1" signifies the concentration of solute at the bottom of the tower, and "2" signifies the concentration of solute at the top of the tower. The total gas flow rate is  $G$ , and  $L$  is the total liquid flow rate (both in moles per hour).

**Figure 57.2** Design diagrams (a) for absorption, (b) for stripping.





The key curves in the figures are the ones labeled  $y^* = F(x)$ . These curves define those solute concentrations in the gas and liquid phases that are in thermodynamic equilibrium with each other. When such concentrations are reached there can be no further transfer of solute between the gas and liquid phases. The dashed lines shown in Fig. 57.2 are known as the *operating lines*. These lines track the material balance of the system all along the length of the tower.

### 57.3 Key Design Assumptions

---

The design of absorbers and strippers involves two steps: (1) estimating the number of theoretical equilibrium stages or transfer units needed to satisfy process design requirements, and (2) estimating the ability of the contacting devices to bring the phases to equilibrium (called the *efficiency* of the devices). In the absence of actual pilot plant data the efficiency is the most difficult parameter to specify. Real-world systems always operate at much less than 100% of theoretical efficiency.

In most environmental applications one can adopt a set of assumptions that in some respects makes estimation of the number of theoretical stages easier and in other respects adds complications. The assumptions are (1) that the solute concentrations are very low, (2) that both the operating and equilibrium lines are straight, and (3) that there are no significant heat effects.

The first assumption makes the tower design calculations easier, but the equilibrium curve  $y^* = F(x)$  is more difficult to determine because solutes at very low concentrations tend to behave nonideally in liquid mixtures; that is, their activity coefficients deviate materially from unity. There is further discussion on this point in the next section.

### 57.4 Physical Data Requirements

---

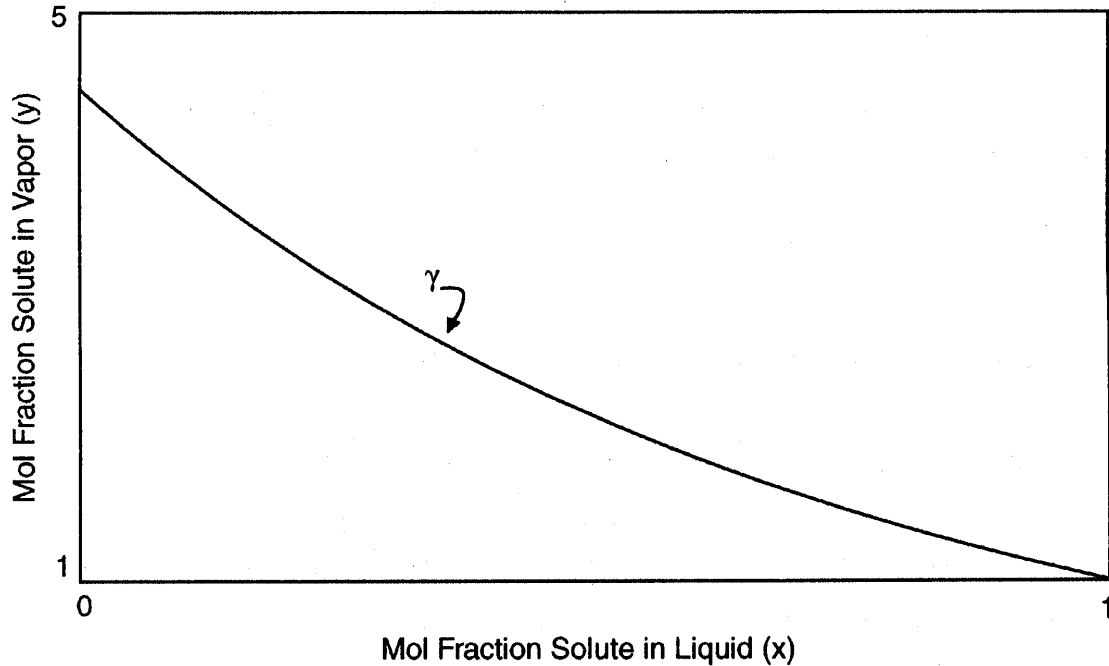
The most difficult physical data to locate when considering a new system design are the  $y^* = F(x)$  data. Moreover, the right-hand function  $F(x)$  rarely is available for new systems.

$$y^* = F(x) = \gamma^L (P_o / P_t) x = K x \quad (57.1)$$

where  $P_o$  is the vapor pressure of pure solute at the system temperature,  $P_t$  is the total pressure, and  $\gamma^L$  is the liquid-phase activity coefficient (a function of liquid concentration  $x$ ).

For a binary system the activity coefficients generally vary as shown in Fig. 57.3. For the case shown the values of  $\gamma^L$  are unity or greater. Note also that the component in large concentration behaves ideally ( $\gamma^L = 1$ ) and the component approaching infinite dilution has its maximum deviation (largest value of  $\gamma^L$ ). In most environmental applications the systems are working at or near infinite dilution.

**Figure 57.3** Activity coefficient behavior of solute.



Fortunately, there is a systematic methodology for estimating the function  $F(x)$  for new systems that is known as the UNIFAC method [Reid *et al.*, 1987], and a computer program is available for making these estimates.

## 57.5 Absorber and Stripper Design Equations

For packed towers the design equation for absorption is

$$N_{og} = (1 - 1/A)^{-1} \ln[(1 - 1/A)(y_1 - mx_2)/(y_2 - mx_2) + 1/A] \quad (57.2)$$

where  $N_{og}$  = theoretical number of overall gas transfer units,  $A = L/mG$ , and  $m$  = slope of the equilibrium curve  $y^* = F(x)$  near the operating point.

For plate towers the design equation for absorption is

$$N = N_{og} (1 - 1/A) / \ln(A) \quad (57.3)$$

where  $N$  = theoretical number of plates (or stages) in the tower.

Plate towers are often employed for stripping. When the liquid feed is dilute and the operating and equilibrium curves are straight lines, the stripper design equation is

$$N = \ln[(1 - A)(x_2 - y_1/m)/(x_1 - y_1/m) + A] / \ln(1/A) \quad (57.4)$$

## 57.6 Absorption and Stripping Efficiency

---

Computations of the number of theoretical plates in a plate absorber or stripper assume that the vapor leaving each plate is in equilibrium with the liquid on the plate. In actual practice a condition of complete equilibrium can never exist because interphase mass transfer requires that a finite concentration driving force differential must exist before any transfer between phases can take place. This leads to a definition of overall plate efficiency,

$$E = 100N/N_{\text{actual}} \quad (57.5)$$

which can be correlated with system design variables. For packed towers the efficiency is measured in terms of the height of an overall transfer unit  $H_{\text{og}}$  so that

$$Z = N_{\text{og}} H_{\text{og}} \quad (57.6)$$

where  $Z$  is the total required height of tower packing. The terms  $E$  and  $H_{\text{og}}$  are best determined experimentally, and data do exist that can be of help in this regard [Fair, 1984].

### Defining Terms

**Absorption:** Process for removing a minor component (or solute) from a gas stream using a liquid solvent.

**Absorption tower (or absorber):** Usually a metal column containing packing materials or plates and designed to improve gas-liquid contacting efficiency.

**Solute:** Component to be removed from the gas entering the absorption tower.

**Solvent:** Liquid employed for removing solute from the incoming solute-rich gas stream.

**Stripping:** Process for separating absorbed solute from solute-rich liquid solvent, thereby regenerating the solvent for reuse in the absorption tower.

**Stripping tower (or stripper):** Usually a metal column containing perforated distillation trays and having a steam reboiler for providing heat to drive the solute out of the rich solvent.

### References

- Fair, J. R. 1984. Liquid-gas systems. In *Perry's Chemical Engineers' Handbook*, 6th ed., p. 18-1–18-45. McGraw-Hill, New York.
- Reid, R. C., Prausnitz, J. M., and Poling, B. E. 1987. Fluid phase equilibria in multicomponent systems. In *The Properties of Gases and Liquids*, 4th ed., p. 314–332. McGraw-Hill, New York.

### Further Information

Edwards, W. M. 1984. Mass transfer and gas absorption. In *Perry's Chemical Engineers'*

- Handbook*, 6th ed., p. 14-1–14-40. McGraw-Hill, New York.
- Palmer, D. A. 1987. *Handbook of Applied Thermodynamics*. CRC, Boca Raton, FL.
- Sandler, S. I. 1989. *Chemical and Engineering Thermodynamics*. John Wiley & Sons, New York.
- Sherwood, T. K., Pigford, R. L., and Wilke, C. R. 1975. *Mass Transfer*. McGraw-Hill, New York.

Van Brunt, V. "Extraction"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

- 58.1 Representative Extraction Processes
- 58.2 Solvent Characteristics and Solvent Screening
- 58.3 Extraction Equilibria
- 58.4 Extraction Staging and Equipment

### Vincent Van Brunt

*University of South Carolina*

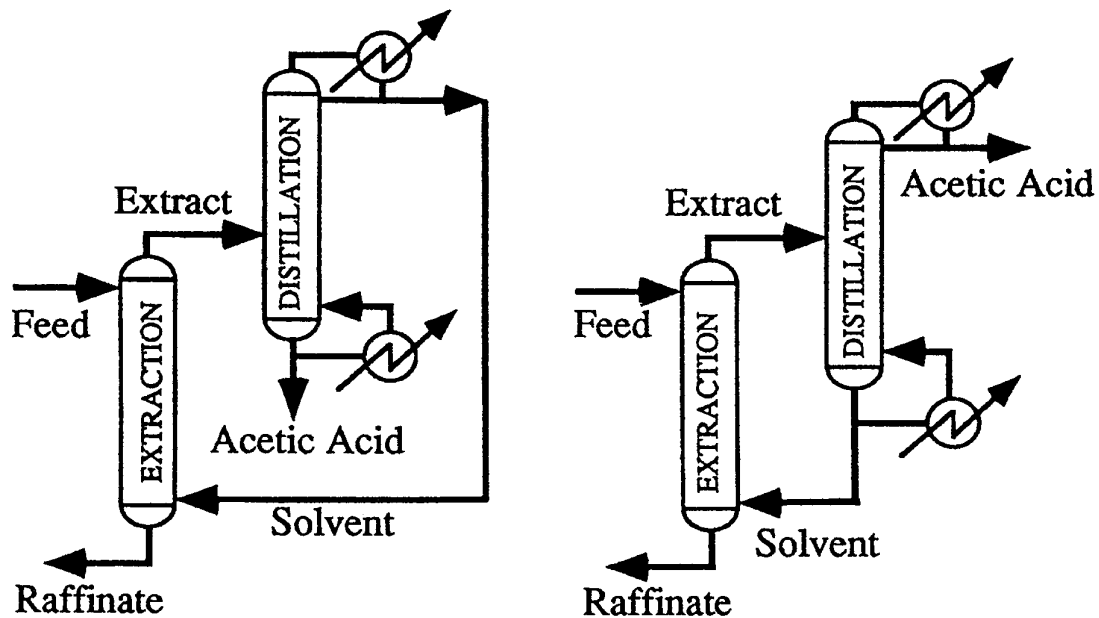
If a mixture cannot be easily separated using a direct separation such as evaporation or distillation, alternative indirect separation processes are considered. Extraction is an indirect separation that relies on the ease of separating a chemical from a solvent compared to that from its original feed.

As an example, consider the separation of acetic acid from water. Although the acid boils 18 degrees higher than water, secondary bonding effects in the liquid phase and nonideal chemical effects in the vapor phase reduce the relative volatility of the water to acetic acid to approximately 1.1. Recovery of the acid from water requires that all the water be vaporized and be in the distillate. The low relative volatility implies that a distillation column will have a high reflux ratio and a large diameter. The combined effect is to increase both the operating and capital costs for distillation.

However, acetic acid can be recovered from dilute aqueous solutions using extraction. Several solvents are selective for acetic acid over water. A simplified extraction operation is shown in [Fig. 58.1](#). The denser aqueous *feed* flows down the extractor and countercurrently contacts the less dense *solvent* flowing up. The solvent-rich *extract* leaves the top. It contains more acid than water. The water-rich *raffinate* stream leaves the bottom with reduced acid content. The extractor is followed by a distillation column that simultaneously purifies the solvent and recovers the solute from it. [Figure 58.1\(a\)](#) shows a process with an ideal separation from a more volatile solvent such as ethyl or isopropyl acetate, whereas [Fig. 58.1\(b\)](#) shows a configuration with a less volatile solvent such as diisobutyl ketone. Finally, [Fig. 58.1\(c\)](#) shows a process configuration for acetic acid recovery with the additional column and separators needed for using a

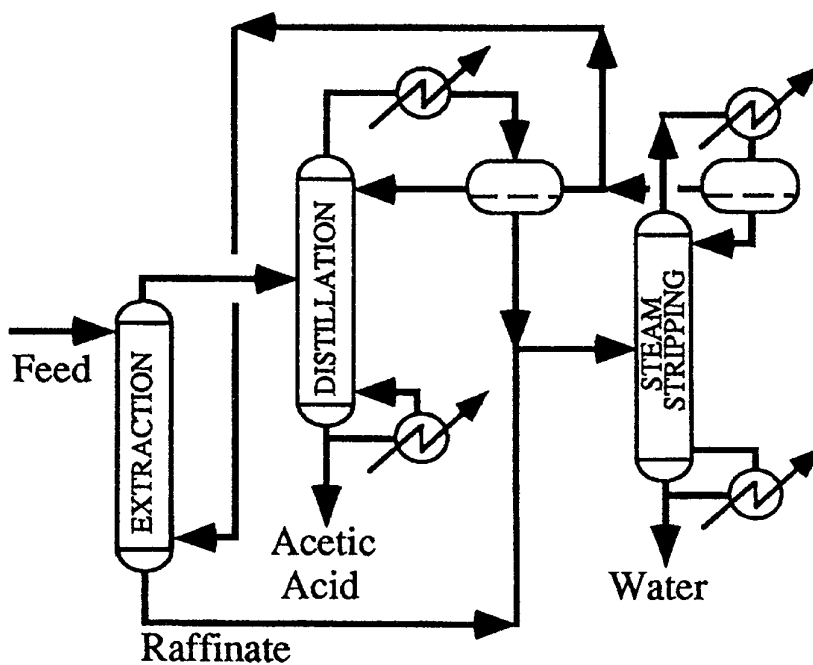
more volatile solvent. The stripping column is needed to reduce solvent loss in the raffinate.

**Figure 58.1** Acetic acid extraction processes.



a) More Volatile Solvent Process

b) Less Volatile Solvent Process



c) Process with Solvent Recovery from Raffinate

Both the capital and the operating costs for recovering acetic acid by the extraction route are less than those for distillation. Choosing between a solvent that is more or less volatile than acetic acid depends on several factors, including relative **selectivity** for acetic acid versus water, relative capacity for acetic acid called **loading**, and relative volatility versus that of acetic acid. The purity of acid required in the product and the initial composition of acid fed to the process greatly influence solvent selection. In general, as the composition of the acid in the aqueous feed diminishes, the quantity of solvent needed for extraction will increase. This, in turn, will necessitate more solvent vaporization in the recovery step. A way around this increase in recovery cost for a more volatile solvent is to switch to a less volatile one. The acid will then become the distillate product in the recovery step, reducing the energy and operating costs for its recovery. Comparative costs are related to the entire process and not dominated by the extraction step. In fact, it usually represents only a small fraction of the overall expenses.

The example shows that extraction *needs to be* considered in the context of an entire separation process. It is justified only when three factors are met. First, the solute must be easier to separate from the new solvent than the original. Second, the solvent must be able to be regenerated and recycled. And third, the solvent losses must be insignificant to the overall process. Thus, extraction is an indirect separation process that requires its evaluation and consideration with a second process step to regenerate the solvent and a third process step to recover the solute from it. In the example already given even though the solute recovery and solvent regeneration steps were combined, still another process step is needed to reduce solvent loss in the raffinate.

## 58.1 Representative Extraction Processes

---

Extraction is used for purifying both inorganic and organic chemicals, see [Rydberg *et al.*, 1992] and [Blumberg, 1988]. Examples of industrial uses are shown in Tables 58.1 through 58.4. The inorganic applications are represented by purification of phosphoric acid and recovery of metals from ore leach solutions. In the latter case the metal salts are nonvolatile and usually in the presence of impurities that prevent direct recovery via, for example, electrowinning. Extraction permits the desired metal to be isolated and then recovered. The solvent for hydrometallurgical applications often relies on a specific chemistry. The examples chosen represent a sampling of the chemistries available [Sekine and Hasegawa, 1977].



**Table 58.1** Representative Inorganic Extraction Systems<sup>1</sup>

Separation	Solvent/Extractant Type	Extractant
Cobalt/nickel	Phosphoric acid	Di-2-ethyl hexyl phosphoric acid (D2EHPA)
Copper	Acid chelating—hydroxyime	2-hydroxyl 5-nonyl benzo phenone oxime (LIX65N)
Uranium	Anion exchanger—tertiary amine	n-trioctyl amine (Alamine336)
Vanadium	Anion exchanger—quaternary ammonium salt	n-trioctyl methylammonium chloride (Aliquat 336)
Zirconium/Hafnium	Solvating—phosphoric acid ester	n-tributyl phosphate (TBP)
Actinides—fuel reprocessing	Solvating—phosphoric acid ester	TBP in dodecane
Phosphoric acid	Solvating—alcohols	Mixtures of butanol and pentanol
Actinides	Aqueous biphasic, polyethylene glycol (PEG) rich	Crown ethers

<sup>1</sup>All feeds are aqueous acid solutions; all diluents are kerosene, except as noted.

Source: Lo, T. C., Baird, M. H. I., and Hanson, C. 1983. *Handbook of Solvent Extraction*. Wiley-Interscience, New York.

**Table 58.2** Representative Organic Extraction Systems

Separation	Solvent
Aromatic/lube oil	Liquid sulfur dioxide
	Furfural
Aromatic/aliphatic	Tetrahydrothiophene-1,1-dioxide (Sulfolane)
	Triethylene glycol–water, tetraethylene glycol–water
	N-methyl 2-pyrrolidone(MPYROL, NMP)–water
Caprolactam	Toluene, benzene
Penicillin, antibiotics	Esters such as isoamyl acetate or butyl acetate
Lipids	Methanol-water
Acetic acid	Low–molecular-weight esters such as ethyl or isopropyl acetate, n-butyl acetate
Caffeine	Supercritical carbon dioxide
Flavors and aromas	Supercritical carbon dioxide

Enzymes	Aqueous biphasic, polyethylene glycol rich
---------	--

*Source:* Lo, T. C., Baird, M. H. I., and Hanson, C. 1983. *Handbook of Solvent Extraction*. Wiley-Interscience, New York.

**Table 58.3** Representative Dual-Solvent Extraction Systems

	Solvent 1	Solvent 2
	Inorganic	
Zinc/iron	Triisooctyl amine (TIOA) Water strip	D2EHPA Sulfuric acid strip
	Organic	
Lube oil	Mix of phenol + cresol	Propane
Aromatic/aliphatic	Dimethyl sulfoxide (DMSO)–water	Paraffin

*Source:* Lo, T. C., Baird, M. H. I., and Hanson, C. 1983. *Handbook of Solvent Extraction*. Wiley-Interscience, New York.

**Table 58.4** Representative Extraction-Reaction Systems

	Reaction
	Inorganic
Uranium / plutonium Potassium nitrate	Plutonium reduction/partitioning with hydroxylamine nitrate (HAN) Potassium chloride + nitric acid → potassium nitrate + hydrochloric acid
	Organic
p-cresol/m-cresol m-xylene/xylenes	m-cresol partitioning using NaOH to form an aqueous soluble salt Preferential extraction and isomerization of m-xylene with HF-BF <sub>3</sub>
	Irreversible reactions
Aromatic nitration	Use of aqueous nitric and sulfuric acids to nitrate toluene

*Source:* Lo, T. C., Baird, M. H. I., and Hanson, C. 1983. *Handbook of Solvent Extraction*. Wiley-Interscience, New York.

The organic applications include recovering heat sensitive chemicals such as penicillin, antibiotics, and vitamins, as well as separating close boiling mixtures. Of particular note is the separation of aromatics from aliphatics, for which there are several competing processes.

Supercritical carbon dioxide is used for the recovery of essential oils, flavors, and aromas and for decaffeination [McHugh and Krukonis, 1986].

Aqueous biphasic systems have a second, stable, less dense, aqueous phase formed by adding polyethylene glycol (PEG) and an inorganic salt such as ammonium sulfate. Both inorganic and organic processes are being developed that use water-soluble complexants to provide selectivity.

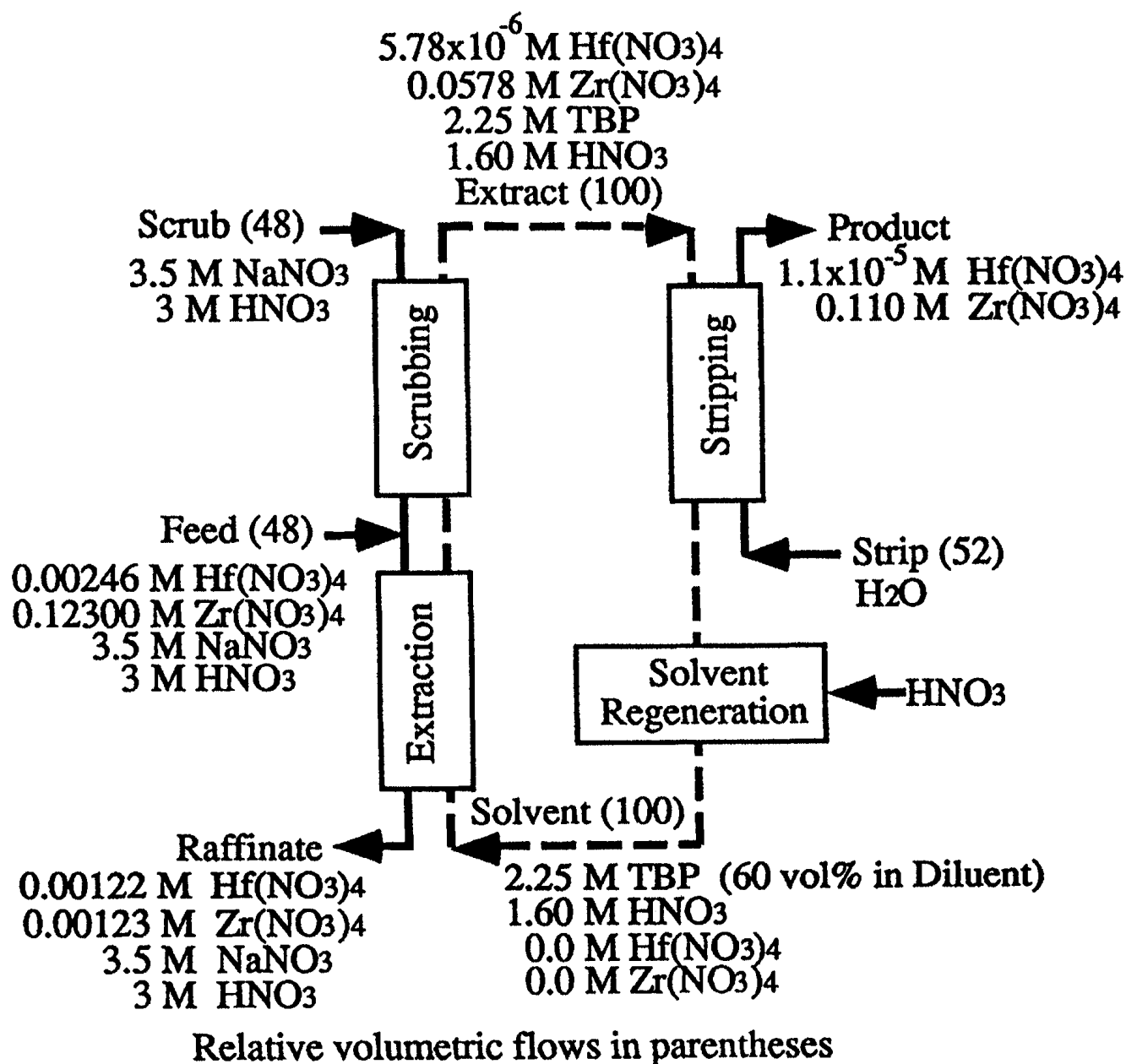
Figure 58.2 shows the key steps in a hydrometallurgical fractional **extraction cycle**. In this example zirconium is recovered and separated from hafnium. As in a distillation column, the feed enters between a rectifying and stripping cascade. The **extraction cascade** below the feed recovers zirconium from it. Each additional extraction stage improves the zirconium fractional recovery from the feed. Leaving the extraction cascade, the loaded solvent also contains coextracted impurities, including hafnium. As in distillation, the upper **scrubbing cascade** enhances the purity of the zirconium product by back-extracting hafnium (and other contaminants) from the solvent. Each additional scrubbing stage further decontaminates the zirconium by reducing the solvent's hafnium content.

In distillation only the reflux flow rate and its temperature can be adjusted to modify the overhead purity. In the extraction analog the *scrub* flow rate, pH, chemistry, and even temperature can all be independently adjusted to purify the zirconium. Note the presence of the nonextracted sodium nitrate in both the scrub and the feed. This **salting agent** permits an independent adjustment of the common anion, nitrate, in addition to a pH adjustment by changing the nitric acid content. The zirconium is recovered from the solvent by contacting it with an aqueous *strip* solution. Not shown are zirconium recovery from the strip solution by precipitation and the fate of the acid that is recovered and recycled.

Multiple extraction steps are used in several processes to enhance loading and to partition species extracted into the first solvent. Representative inorganic and organic mixtures are shown in Table 58.3. The zinc purification from spent electrolyte acid solution first separates zinc chloride using triisooctyl amine (TIOA). The first solvent is stripped with water before contact with the di-2-ethylhexylphosphoric acid (D2EHPA), which further purifies the zinc. Extraction of rare earths from a leach liquor is a sequential use of extraction cycles. In organic systems the second solvent may be used for solvent regeneration to lower process energy costs or to aid in solute and solvent recovery.

Both reversible and irreversible reactions are performed in the presence of two liquid phases. Extraction and reaction is used for fine separations such as isomer purification and for actinide partitioning. Isomer purification processes exploit a reversible reaction in an aqueous phase to enhance the selectivity between the isomers. Likewise, the plutonium reduction uses an irreversible reaction to effect a similar result by significantly reducing its organic phase solubility. Other representative systems are shown in Table 58.4.

**Figure 58.2** Zirconium-hafnium extraction cycle. (Adapted from Benedict, M., Pigford, T. H., and Levi, H. 1981. *Nuclear Chemical Engineering*, 2nd ed. McGraw-Hill, New York.)



## 58.2 Solvent Characteristics and Solvent Screening

Since the solvent is present in the entire process, each one of its characteristics and properties must be weighed to consider the net effect. The most significant considerations are shown in Table 58.5. The first grouping examines the solvent's relative process compatibility. A solvent or class of solvents that, even in minute quantities, reduces the value or marketability of the solute can be removed from consideration. Solute selectivity and loading ability must be balanced against ease of reversing and unloading it, recovering the solvent, and reducing the solvent losses.

**Table 58.5** Desirable Solvent Characteristics

Process compatibility	High solute selectivity
	Regeneration capability
	High solute distribution coefficient
	High solute loading
	Low raffinate solubility
	Solute compatibility
Processing and equipment	Low viscosity
	Density different from the feed (>2%, preferably >5%)
	Moderate interfacial tension
	Reduced tendency to form a third phase
	Low corrosivity
Safety and environmental	Low flammability, flash point
	Low toxicity
	Low environmental impact, both low volatility and low effect on raffinate disposal costs
Overall	Low cost and availability

Adapted from Robbins, L. 1983. Liquid-liquid extraction. In *Perry's Chemical Engineers' Handbook*, 6th ed., ed. R. H. Perry and D. Green. McGraw-Hill, New York; Cusack, R. W., Fremeaux, P., and Glatz, D. 1991. A fresh look at liquid-liquid extraction. *Chem. Eng.* February.

Once a solvent has met these tests it must pass those in the second grouping that affect processing equipment design and specification. Since extraction relies on intermittent dispersion and coalescence, desirable processing characteristics are those that enhance settling and coalescence, such as density difference with the feed and moderate interfacial tension. In general, a solvent having a lower viscosity is preferred. Likewise, the preferred solvent is less corrosive than the feed solution. Solvent chemistry can be altered to improve processing behavior.

The last grouping covers safety and environmental concerns. Safety issues such as toxicity and flammability may dominate over all others. If the raffinate or any other process stream is sent to waste treatment facilities, the solvent must be compatible with them or have low disposal costs. Finally, the preferred solvent is a readily available, and low-cost commercial product.

## 58.3 Extraction Equilibria

---

Extraction equilibria affect the initial solvent screening. However, since many solvents contain more than one component, conclusions drawn from three component systems may not readily translate into the selection of a practical one. Almost always, experimental testing with the actual feed solution is needed. It is common practice to consider a solvent as having three parts, each piece contributing to desirable solvent properties. The **diluent** provides the bulk properties, such as viscosity and density. The **extractant** provides the active reversible solvent-solute interactions. Its efficiency affects both selectivity and loading. The **modifier** improves the interfacial tension and affects third-phase formation. An effective modifier will either increase the interfacial tension, thereby preventing or reducing emulsion stability or, conversely, it may be chosen to reduce the interfacial tension to enhance dispersion and interfacial area.

Organic diluents are usually petroleum fractions of varying naphthenic, aromatic, and aliphatic composition. Water may be added to aqueous soluble solvents as a diluent. Common modifiers are 2-ethylhexanol, isodecanol, tributyl phosphate and n-nonyl phenol. [Ritcey and Ashbrook, 1984] discusses their behavior. Extractant chemistry is discussed in Sekine and Hasegawa [1977] and in Marcus and Kertes [1969].

For organic systems the thermodynamics of extraction equilibria can be used to screen potential solvents. The **distribution coefficient** is the ratio of a solute's mole fraction in the two phases. Here  $y$  refers to the organic, or less dense phase, mole fraction and  $x$  refers to the aqueous, or more dense phase, mole fraction.

$$K_i = y_i/x_i \quad (58.1)$$

For two chemical constituents the selectivity,  $\beta$ , can be expressed as

$$\beta_{ij} = K_i/K_j = (y_i/x_i)/(y_j/x_j) \quad (58.2)$$

Here the selectivity is written to evaluate the relative concentrations of  $i$  and  $j$  in the solvent phase, designated  $S$ , versus that in the feed phase, designated  $F$ . At equilibrium the chemical potential for each species is the same in each phase. In terms of activities this becomes

$$\gamma_i^F x_i = \gamma_i^S y_i \quad (58.3)$$

where  $\gamma$  is the activity coefficient. Using Eq. (58.3), the distribution coefficient can be expressed as

$$K_i = \gamma_i^F / \gamma_i^S \approx \gamma_i^{\infty F} / \gamma_i^{\infty S} \quad (58.4)$$

where the  $\infty$  indicates infinite dilution. Likewise, the selectivity becomes

$$\beta_{ij} = \frac{\gamma_i^F / \gamma_i^S}{\gamma_j^F / \gamma_j^S} = \left( \frac{\gamma_i}{\gamma_j} \right)^F \left( \frac{\gamma_j}{\gamma_i} \right)^S \quad (58.5)$$

Since the composition of the species in the feed is known, the ratio of activity coefficients at the feed composition can be determined using one of the nonideal models, such as Van Laar, NRTL, UNIQUAC, or UNIFAC. The reader is referred to Sorensen and Arlt [1980] for more details. Examining Eq. (58.5), one sees that the solvent selectivity for species  $i$  over  $j$  increases if the solvent phase activity coefficient for  $i$  is less than that for  $j$ . A preferred solvent would create solvent-solute interactions for species  $i$ , which reduces its activity coefficient below 1.0—that is, to have negative deviations from ideality. Furthermore, the ideal solvent would preferentially reject species  $j$  and have its activity coefficient greater than 1.0—that is, have positive deviations from ideality.

These effects can be used to identify potential solvent classes. Table 58.6 shows the expected deviations from ideality for solute-solvent pairs. For the acetic acid–water mixture the solute-solvent pairs that have negative deviations from ideality for the acid and no deviation or positive deviation for water are groups 2 through 5. Suitable solvents will probably come from those classes. In fact, solvents from those classes have all been used in practice or research. See C. J. King's chapter in [Lo et al., 1983] for more details.

**Table 58.6** Qualitative Solvent Screening

Group	Solute	Solvent								
		1	2	3	4	5	6	7	8	9
1	Acid, aromatic OH, for example, phenol	0	–	–	–	–	0	+	+	+
2	Paraffinic OH (alcohol), water, amide or imide with active H	–	0	+	+	+	+	+	+	+
3	Ketone, aromatic nitrate, tertiary amine, pyridine, sulfone, phosphine oxide, or trialkyl phosphate	–	+	0	+	+	–	0	+	+
4	Ester, aldehyde, carbonate, nitrite or nitrate, phosphate, amide without active H; intermolecular bonding, for example, o-nitrophenol	–	+	+	0	+	–	+	+	+
5	Ether, oxide, sulfide, sulfoxide, primary, or secondary amine or imine	–	+	+	+	0	–	0	+	+
6	Multihalo-paraffin with active H	0	+	–	–	–	0	0	+	0
7	Aromatic, halogenated aromatic, olefin	+	+	0	+	0	0	0	0	0
8	Paraffin	+	+	+	+	+	+	0	0	0
9	Mono-halogenated paraffin or olefin	+	+	+	+	+	0	0	+	0

Notes: The + sign means that the solvent tends to raise the activity coefficient of the solute in a row group; the – sign means that the solvent tends to lower the activity coefficient of the solute in a row group. 0 indicates no appreciable effect. Solvation is expected for negative group interactions. Potential solvents are those that lower the activity, that is, have minus signs.

Source: Cusack, R. W., Fremeaux, P., and Glatz, D. 1991. A fresh look at liquid-liquid extraction. *Chem. Eng.* February.

For the separation of benzene from n-hexane, [Table 58.6](#) still provides some guidance. Although no negative deviations from ideality are shown, groups 3, 5, and 6 show positive deviations for the aliphatics (n-hexane) and 0 deviations for aromatics (benzene). This implies that solvents from those classes will probably be selective for benzene. The table also indicates that a solvent that is preferential for the aliphatics (n-hexane) over the aromatics (benzene) cannot be identified readily.

Another refinement for solvent screening is to use infinite dilution activity coefficients. Equation 58.5 can be approximated as



$$\beta_{ij}^{\infty} \cong \left( \frac{\gamma_i}{\gamma_j} \right)^F \left( \frac{\gamma_j^{\infty}}{\gamma_i^{\infty}} \right)^S \approx \left( \frac{\gamma_j^{\infty}}{\gamma_i^{\infty}} \right)^S \quad (58.6)$$

Here one neglects the activity coefficient ratio in the feed solvent as a given quantity and focuses attention on the ratio for the solutes to be separated in potential solvents. Further, the ratio is approximated by the ratio at infinite dilution—that is, for each of the solutes at infinite dilution in the solvent. Again, one sees that the selectivity of the solvent for  $i$  over  $j$  is *inversely* proportional to the ratio of their activity coefficients. If one uses infinite dilution activity coefficient data for common solvents from those groups suggested from Table 58.6, one obtains Table 58.7. The high values obtained for the industrial solvents are striking. Also, the difference between the two nitriles indicates an effect of carbon number. There are generally lower values for solvents having groups not identified using the qualitative screening table. Prediction of infinite dilution activity coefficients for solvent screening may also be performed using the MOSCED model and refinements to it; see Park and Carr [1987] and Hait *et al.* [1993].

**Table 58.7** Infinite Dilution Selectivity of Benzene–n-Hexane

Solvent	Table 58.6 Group	$\beta_{\text{BH}}^{\infty} \frac{\gamma_{\text{n-hexane}}^{\infty}}{\gamma_{\text{benzene}}^{\infty}}$
Sulfolane	5, 5	30.5
Dimethylsulfoxide	5	22.7
Diethylene glycol	1, 5	15.4
Triethylene glycol	1, 5	18.3
Propylene carbonate	4	13.7
Dimethylformamide	2	12.5
n-methylpyrrolidone	3	12.5
Acetonitrile	3	9.4
Succinonitrile	3	46.8
g-butyrolactone	3	19.5
Aniline	5	11.2
Dichloroacetic acid	1	6.1

Modified from Tiegs, D. *et al.*, 1986. *Activity Coefficients at Infinite Dilution. C1-C9*. DECHEMA Chemistry Data Series, vol. IX, part 1. DECHEMA, Frankfurt.

## 58.4 Extraction Staging and Equipment

---

Procedures for extraction equilibrium-stage calculations are given in standard references, for example, Treybal [1963] and Lo *et al.* [1983]. For dilute systems, staging can be calculated by assuming a constant selectivity and constant solvent-to-feed ratio. An extraction factor—equivalent to a stripping factor and equal to the ratio of equilibrium to operating lines—can be used to calculate the number of equilibrium stages using the Kremser equation. For concentrated systems neither the operating nor the equilibrium lines are of constant slope, and calculation procedures must account for the change in the solvent-to-feed ratio.

Although equilibrium-stage calculations are suitable for costing and comparing processes, detailed design *requires* an experimental evaluation of extractor performance. Because extraction uses phases that are the same order of magnitude in density, column performance is reduced by dispersion (also called *axial mixing* or *backmixing*). Dispersion will reduce extraction efficiency and can be approximated; see Lo *et al.* [1983] and Godfrey and Slater [1994]. However, it is extremely equipment- and chemistry-dependent.

Actual equipment sizing is tied to the physical properties of the specific extraction system being evaluated. Unlike distillation, many different types of equipment are used. Logic diagrams for specifying extractors are given in standard texts such as Ladda and Degaleesan [1978], Lo *et al.* [1983], and Godfrey and Slater [1994]. Figure 58.3 presents a simplified version of the equipment decision tree [Ladda and Degaleesan, 1973]. It is strongly recommended that pilot testing be performed to determine both sizing needs and operating limits for new processes. Pilot campaigns with actual process chemicals will enable understanding of real processing behavior—for example, observation of interfacial solids buildup or changes in emulsion stability with degree of agitation.

### Defining Terms

**Diluent:** The bulk constituent of a solvent, primarily used to improve solvent properties that affect processing and equipment choice.

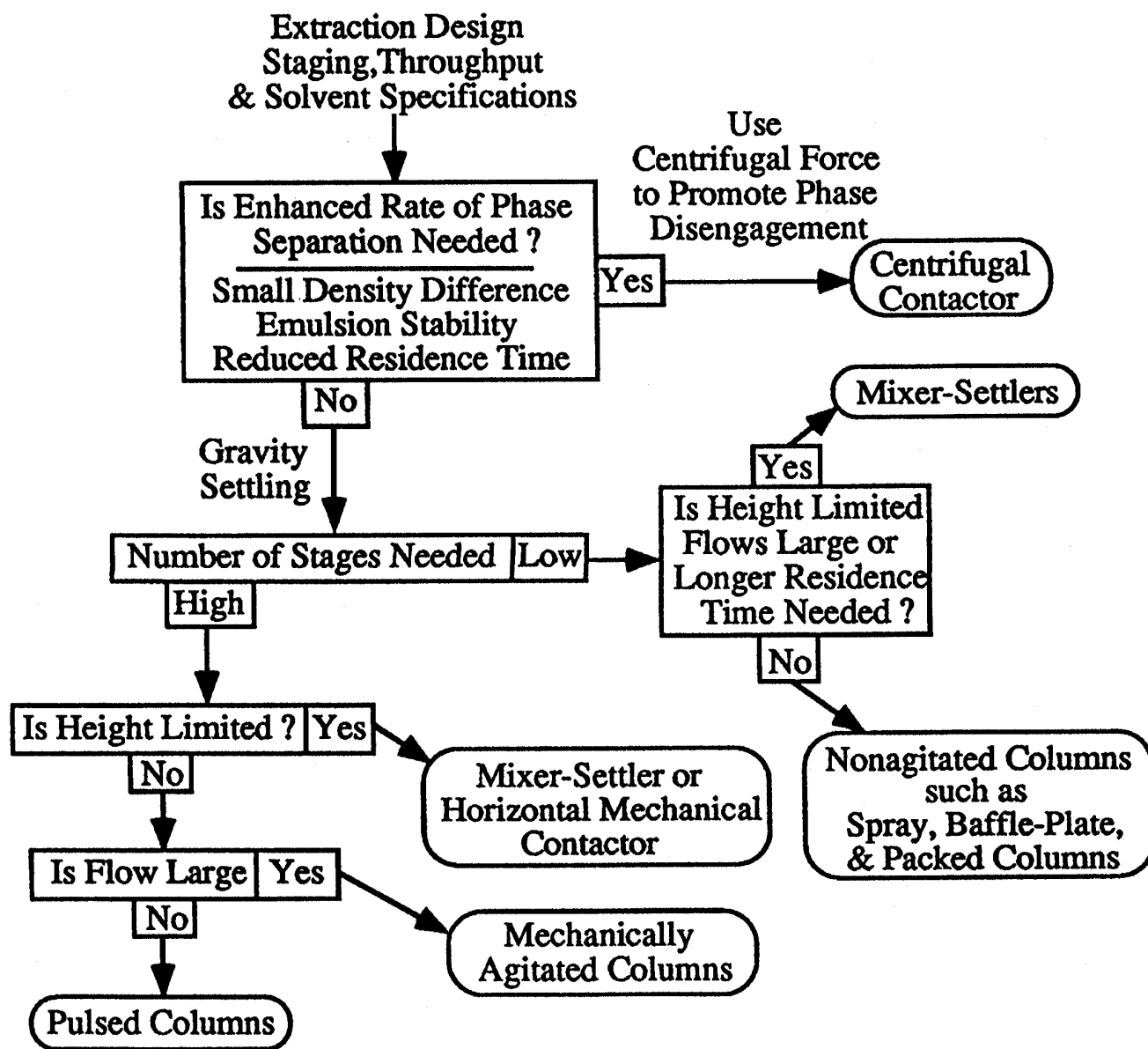
**Distribution coefficient ( $K$ ):** The ratio of the mole fraction of a chemical species in the solvent phase divided by the mole fraction of the same chemical in the feed phase.

**Extractant:** The active part of a solvent that has solvating characteristics with the solute or specific solute-solvent chemistry such as chelation or ion-pair formation.

**Extraction cascade:** A series of stages to recover the solute from a feed solution by contacting it with a relatively immiscible solvent.

**Extraction cycle:** A sequence of extraction, scrubbing, and stripping cascades to recover and purify a particular solute. Cycles may themselves be staged to enhance solute purity.

**Figure 58.3** Equipment type selection decision tree. (Adapted from Ladda, G. S. and Degaleesan, T. E. 1978. *Transport Phenomena in Liquid Extraction*. McGraw-Hill, New York.)



**Loading:** The capacity of the solvent for the solute. Usually, this is expressed on a per mole solvent basis. Higher values indicate reduced solvent needs.

**Modifier:** A chemical added to a solvent to retard third-phase formation.

**Salting agent:** A nonextracted chemical, usually a salt, used to change the process chemistry. For inorganic systems the salting agent may have an anion in common with the solute but have the wrong cation valence or size to be extracted.

**Scrubbing:** The operation of decontaminating a loaded solvent of chemicals coextracted with the solute. Scrubbing purifies the solute-loaded solvent. A *scrubbing cascade* is a series of stages to perform scrubbing.

**Selectivity ( $\beta$ ):** The separation factor for extraction is the ratio of distribution coefficients for two solutes. The higher the value is, the greater the ability of the solvent to separate the two species.

**Stripping:** The operation of unloading the solute from the solvent by extracting it back into a feed-like phase, from which it can be easily removed. Stripping also recovers the solvent for recycle back to the extraction cascade. A *stripping cascade* is a series of stages to perform stripping.

## References

- Blumberg, R. 1988. *Liquid-Liquid Extraction*. Academic Press, New York.
- Cusack, R. W., Fremeaux, P., and Glatz, D. 1991. A fresh look at liquid-liquid extraction. *Chem Eng*. February.
- Godfrey, J. C. and Slater, M. J. 1994. *Liquid-Liquid Extraction Equipment*. Wiley, New York.
- Hait, M. J., *et al.* 1993. Space predictor for infinite dilution activity coefficients. *Ind. Eng. Chem. Res.* 32: 2905–2914.
- Ladda, G. S. and Degaleesan, T. E. 1978. *Transport Phenomena in Liquid Extraction*. McGraw-Hill, New York.
- Lo, T. C., Baird, M. H. I., and Hanson, C. 1983. *Handbook of Solvent Extraction*. Wiley-Interscience, New York.
- Marcus, Y. and Kertes, A. S. 1969. *Ion Exchange and Solvent Extraction of Metal Complexes*. Wiley-Interscience, New York.
- McHugh, M. A. and Krukonis, V. J. 1986. *Supercritical Fluid Extraction, Principles and Practices*. Butterworth, Boston, MA.
- Park, J. H. and Carr, P. W. 1987. Predictive ability of the MOSCED and UNIFAC activity coefficient methods. *Anal. Chem.* 59: 2596–2602.
- Ritcey, G. M. and Ashbrook, A.W. 1984. *Solvent Extraction, Principles and Applications to Process Metallurgy, Part I*. Elsevier, New York.
- Robbins, L. Liquid-liquid extraction. In *Perry's Chemical Engineer's Handbook*, 6th ed. McGraw-Hill, New York.
- Rydberg, J., Musikas, C., and Choppin, G. R. 1992. *Principles and Practices of Solvent Extraction*. M. Dekker, New York.

- Schultz, W. W., and Navratil, J. D. *Science and Technology of Tributyl Phosphate* (vol. 1, 1984; vols. IIA and IIB, 1987; vol. III, 1990; vol. IV, 1991). CRC Press, Boca Raton, FL.
- Sekine, T. and Hasegawa, Y. 1977. *Solvent Extraction Chemistry, Fundamentals and Applications*. M. Dekker, New York.
- Sorensen, J. M. and Arlt, W. *Liquid-Liquid Equilibrium: Chemistry Data Series, Volume V* (part 1, Binary Systems, 1979; part 2, Ternary Systems, 1980; part 3, Ternary and Quaternary Systems, 1980; supplement 1, 1987, by Macedo, E. A. and Rasmussen, P.). DECHEMA, Frankfurt.
- Tiegs, D., 1986. *Activity Coefficients at Infinite Dilution, C1-C9*. DECHEMA Chemistry Data Series, vol. IX, part 1. DECHEMA, Frankfurt.
- Treybal, R. E. 1963. *Liquid Extraction*, 2nd ed. McGraw-Hill, New York.
- Wisniak, J. and Tamir, A. *Liquid-Liquid Equilibrium and Extraction: A Literature Source Book* (vol. 1, 1980; vol. 2, 1980; supplement 1, 1987). Elsevier, New York.

## Further Information

The first general reference for extraction chemistry, equipment, and its operation is the *Handbook of Solvent Extraction*. Equipment is further described in *Transport Phenomena in Liquid Extraction* and in *Liquid-Liquid Extraction Equipment*.

The European Federation of Chemical Engineering has established three systems for evaluation of extraction operations. These guidelines are available in Misak, T. 1984. *Recommended Systems for Liquid-Liquid Extraction*, Institute of Chemical Engineers. Rugby, UK.

Sources of reliable extraction data include all the volumes of Sorenson and Arlt [1979, 1980, 1980, 1987] as well as all the volumes of Wisniak and Tamir [1980, 1980, 1987]. The text by Francis, A. W., 1963, *Liquid-Liquid Equilibriums*, Wiley-Interscience, New York also contains references to reliable data.

The International Solvent Extraction Conference (ISEC) is held every three years with bound proceedings available. The conference presents the latest information about extraction chemistry and equipment. See listings under *Proceedings of ISEC*. Each proceedings has been published by a different professional organization.

Many extraction processes and refinements of them are discussed in the journals *Solvent Extraction and Ion Exchange* and *Separation Science and Technology*. The series *Ion Exchange and Solvent Extraction* contains tutorial articles about specific topics.

The science of extraction has been greatly influenced by nuclear applications, including nuclear reprocessing. Many references are available, including chapters in Long, J. T. 1978. *Engineering for Nuclear Fuel Reprocessing*. Amer. Nuclear. Soc., La Grange, IL; Benedict, M., Pigford, T. H., and Levi, H. L. 1981. *Nuclear Chemical Engineering*, 2nd ed. McGraw-Hill, New York; and all four volumes of Schultz and Navratil [1984, 1987, 1990, 1991].

Sircar, S. "Adsorption"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

- 59.1 Adsorbent Materials
- 59.2 Adsorption Equilibria
- 59.3 Heat of Adsorption
- 59.4 Thermodynamic Selectivity of Adsorption
- 59.5 Adsorption Kinetics
- 59.6 Adsorption Column Dynamics
- 59.7 Adsorptive Separation Processes and Design

**Shivaji Sircar**

*Air Products and Chemicals, Inc.*

Adsorption is a surface phenomenon. When a pure fluid (gas or liquid) is contacted with a solid surface (adsorbent), fluid-solid intermolecular forces of attraction cause some of the fluid molecules (adsorbates) to be concentrated at the surface. This creates a denser region of fluid molecules, which extends several molecular diameters near the surface (adsorbed phase). For a multicomponent fluid mixture, certain components of the mixture are preferentially concentrated (selectively adsorbed) at the surface due to differences in the fluid-solid forces of attraction between the components. This creation of an adsorbed phase having a composition different from that of the bulk fluid phase forms the basis of separation by adsorption technology.

Adsorption is a thermodynamically spontaneous process. Energy is released (exothermic) during the process. The reverse process by which the adsorbed molecules are removed from the surface to the bulk fluid phase is called **desorption**. Energy must be supplied to the adsorbed phase (endothermic) for the desorption process. Both adsorption and desorption form vital steps in a practical separation process in which the adsorbent is repeatedly used. This concept of regenerative ad(de)sorption is key to the practical use of this technology. It has found numerous commercial applications in chemical, petrochemical, biochemical, and environmental industries for separation and purification of fluid mixtures, as listed in [Table 59.1](#).

**Table 59.1** Key Commercial Applications of Adsorption Technology

Gas Separation	Liquid Separation	Environmental Separation	Bioseparation
Gas drying	Liquid drying	Municipal and	Recovery of
Trace impurity	Trace impurity	industrial waste	antibiotics
removal	removal	treatment	Purification and
Air separation	Olefin-paraffin	Ground and surface	recovery of
Carbon	separation	water treatment	enzymes
dioxide–methane	Xylene, cresol,	VOC removal	Purification of
separation	cymene isomer		proteins
Carbon	separation		Removal of
monoxide–hydroge	Fructose and glucose		microorganisms
n separation	separation		Recovery of
Hydrogen and	Breaking		vitamins
carbon dioxide	azeotropes		
recovery from SMR			
off gas			
Production of			
ammonia-synthesis			
gas			
Normal-iso paraffin			
separation			
Ozone enrichment			
Solvent vapor			
recovery			

## 59.1 Adsorbent Materials

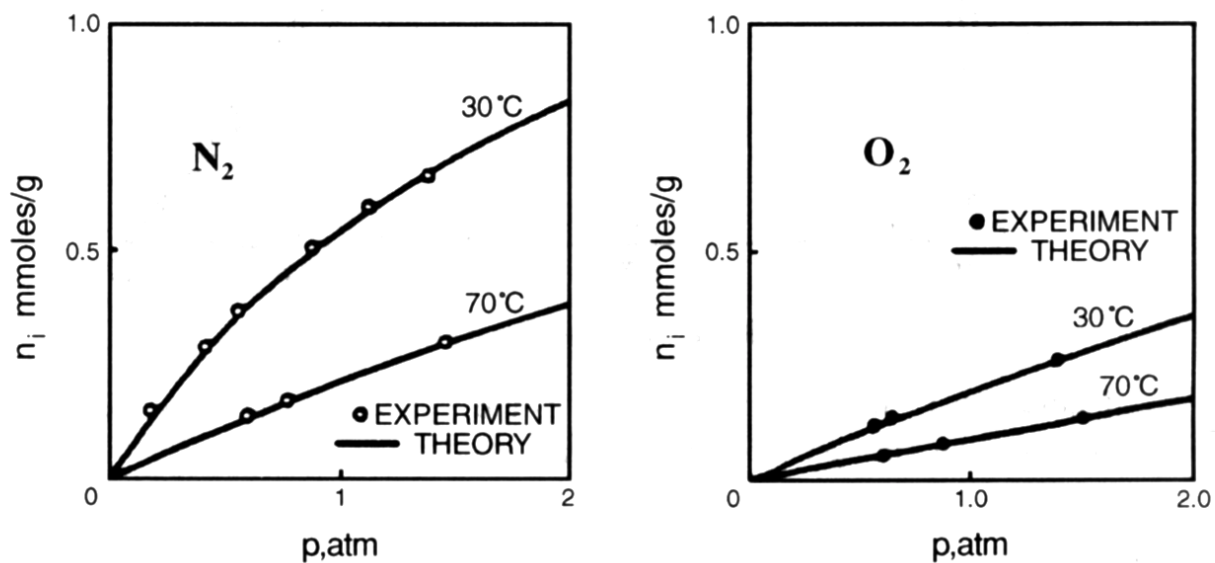
A key element in the development of adsorption technology has been the availability of a large spectrum of micro- and mesoporous adsorbents. They have large specific surface areas (500–1500 m<sup>2</sup>/g), varying pore structures, and surface properties (polar and nonpolar) that are responsible for selective adsorption of specific components of a fluid mixture. These include activated carbons, zeolites, aluminas, silica gels, polymeric adsorbents, and ion-exchange resins. Adsorbents may be energetically homogeneous, containing adsorption sites of identical adsorption energy, or heterogeneous, containing a distribution of sites of varying adsorption energies.



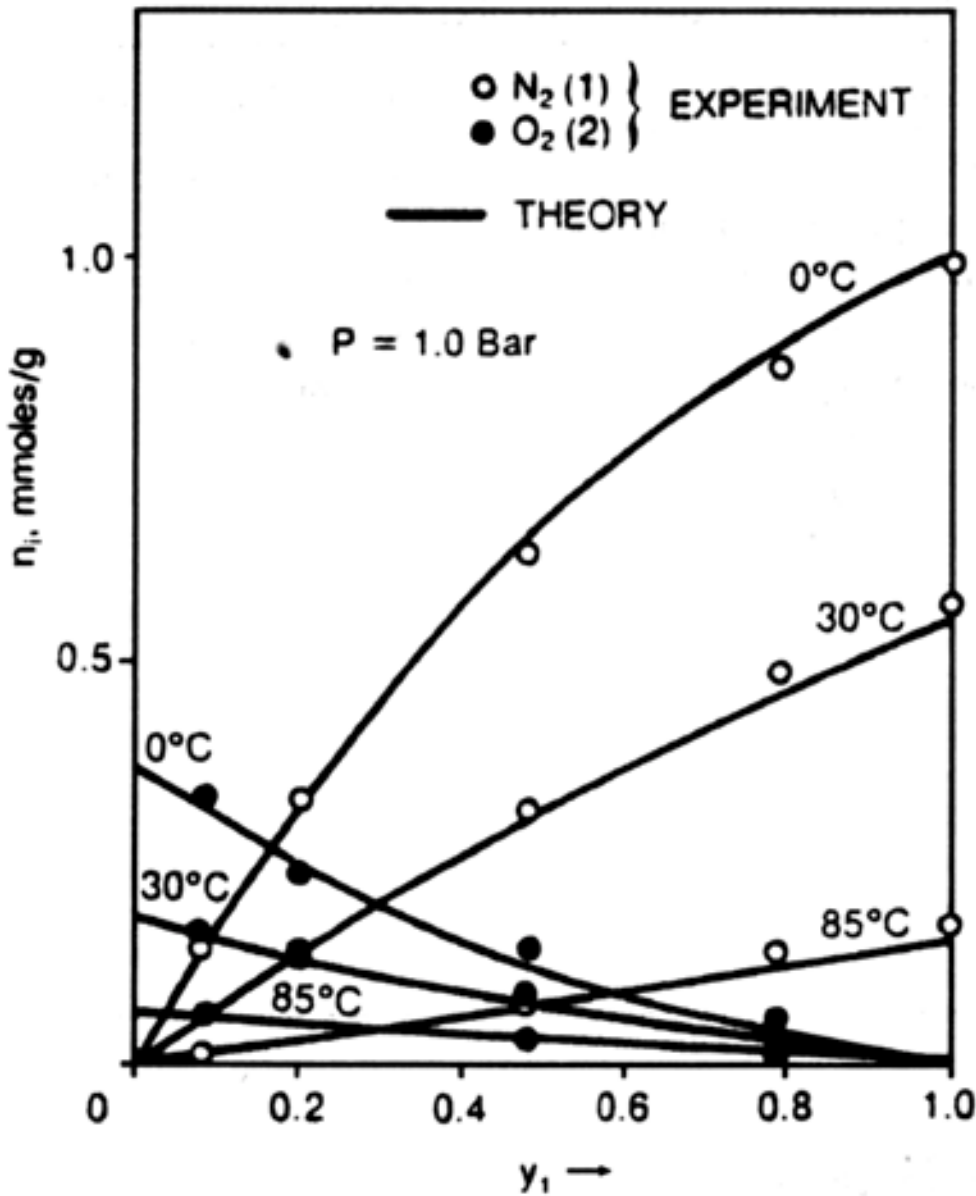
## 59.2 Adsorption Equilibria

Adsorption equilibria determine the thermodynamic limits of specific amounts of adsorption (moles/g) of a pure fluid or the components of a mixture under a given set of conditions [pressure ( $P$ ), temperature ( $T$ ), and mole fraction ( $y$ )] of the bulk fluid phase. A convenient way to represent adsorption equilibria is in terms of adsorption isotherms in which the specific amount adsorbed of a pure gas ( $n_i^o$ ) or that of component  $i$  ( $n_i$ ) from a multicomponent gas mixture is expressed as a function of  $P$  (pure gas) or as functions of  $P$  and  $y_i$  (fluid mixtures) at constant  $T$ . The values  $n_i^o$  and  $n_i$  decrease with increasing  $T$  for a given  $P$  and  $y_i$ . Adsorption isotherms can have many different shapes, but most microporous adsorbents exhibit the shape (type I) shown in Fig. 59.1 (pure gas) and Fig. 59.2 (binary gas).

**Figure 59.1** Langmuirian pure gas adsorption isotherms on Na-mordenite. (Source: Kumar and Sircar, 1986.)



**Figure 59.2** Langmuirian binary gas adsorption isotherms on Na-mordenite. (Source: Kumar and Sircar, 1986.)



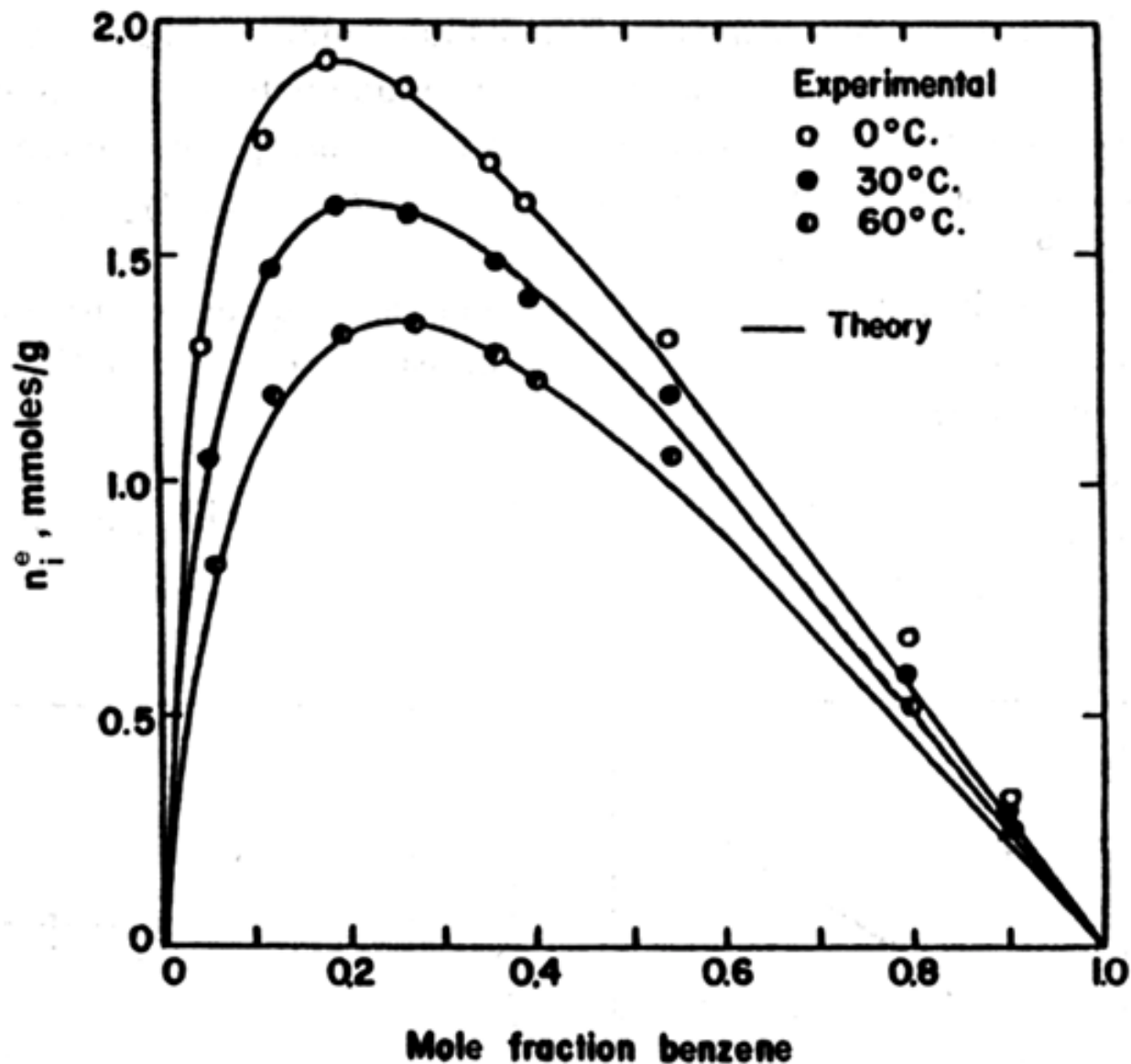
The simplest thermodynamic system for adsorption from liquids consists of a binary liquid mixture. The adsorption isotherm in this case is measured in terms of Gibbs surface excess (moles/g) of component  $i$  ( $n_i^e$ ):

$$n_i^e = n_i - y_i \sum_i n_i; \quad \sum_i n_i^e = 0, \quad i = 1, 2; \quad n_i^e = 0 \text{ for pure liquid} \quad (59.1)$$

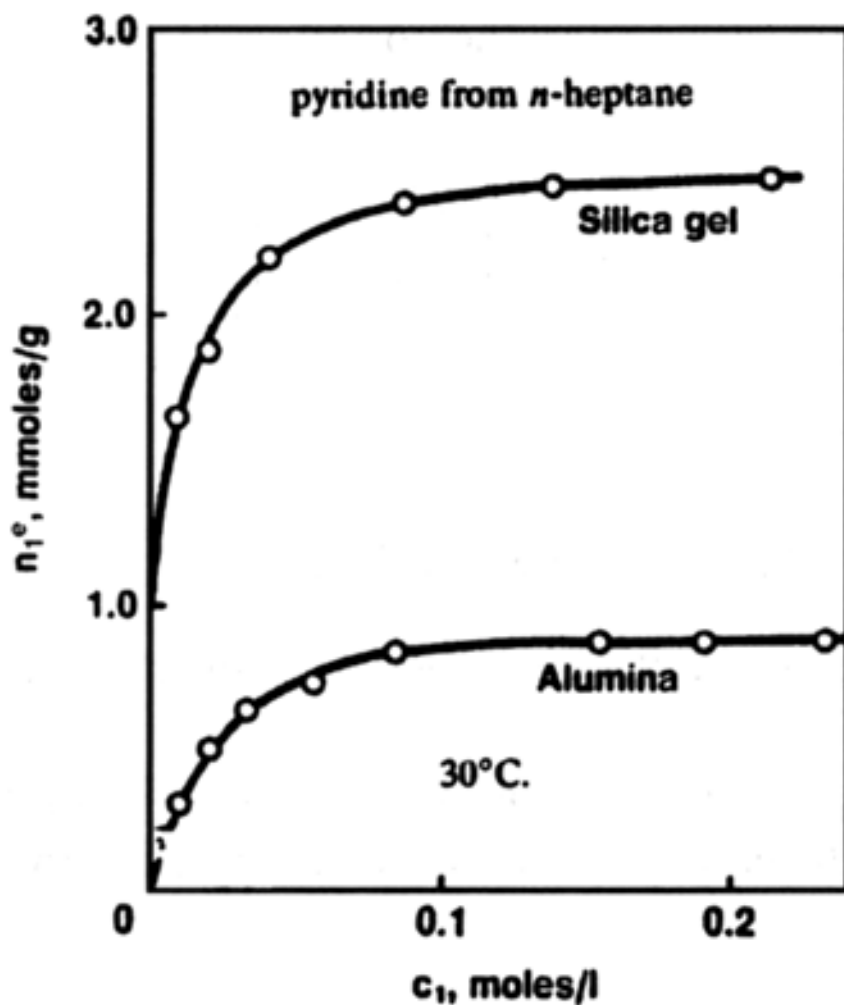
The liquid phase surface excess is a function of  $y_i$  and  $T$ . [Figure 59.3](#) shows an

example of bulk binary liquid phase isotherm. For adsorption of a very selective trace component ( $y_i \ll 1$ ),  $n_i^e$  is approximately equal to  $n_i$ . The isotherm in that case (Fig. 59.4) is type I in shape.

**Figure 59.3** Benzene-cyclohexane liquid phase surface excess isotherms on silica gel.  
(Source: Sircar *et al.*, 1972.)



**Figure 59.4** Surface excess isotherms for a trace component from liquids. (Source: Rahman and Ghosh, 1980.)



The simplest model to describe pure and multicomponent gas adsorption isotherms on a homogeneous adsorbent is the Langmuir equation,

$$\text{Pure gas: } n_i^o = \frac{mb_iP}{1 + b_iP}; \quad \text{Mixed gas: } n_i = \frac{mb_iPy_i}{1 + \sum_i b_iPy_i} \quad (59.2)$$

where  $b_i$  is the gas-solid interaction parameter for component  $i$ . The term  $b_i$  is a function

of temperature [Eq. (59.7)] only. The term  $m$  is the saturation-specific adsorption capacity (moles/g) for the components.

The Toth equation is often satisfactory for describing gas adsorption isotherms on heterogeneous adsorbents:

$$\text{Pure gas: } n_i^o = \frac{mb_iP}{[1 + (b_iP)^k]^{1/k}}; \quad \text{Mixed gas: } n_i = \frac{mb_iPy_i}{[1 + (\sum_i b_iPy_i)^k]^{1/k}} \quad (59.3)$$

The term  $k$  ( $\leq 1$ ) is the heterogeneity parameter. The lower the value of  $k$  is, the larger is the degree of adsorbent heterogeneity. The adsorbent is homogeneous when  $k$  is unity (Langmuir model).

Statistical mechanics of adsorption dictates that the gas adsorption isotherms become linear functions of pressure (pure gas) and component partial pressures (mixed gas) when the total gas pressure approaches zero ( $P \rightarrow 0$ ):

$$\text{Pure gas: } n_i^o = K_iP; \quad \text{Mixed gas: } n_i = K_iPy_i \quad (59.4)$$

$K_i$  is called the Henry's law constant for pure gas  $i$ . It is a function of temperature only [Eq. (59.7)].

The simplest model isotherm for adsorption of an ideal binary liquid mixture of equal adsorbate sizes on a homogeneous adsorbent is given by

$$n_1^e = \frac{my_1y_2(S-1)}{Sy_1+y_2} = -n_2^e \quad (59.5)$$

where  $S$  is the selectivity of adsorption of component 1 over component 2.

## 59.3 Heat of Adsorption

---

The thermodynamic variable of practical use that describes the differential heat evolved (consumed) during the ad(de)sorption process due to a differential change in the adsorbate loading ( $n_i^o$  or  $n_i$ ) of a pure gas ( $q_i^o$ ) or the components of a gas mixture ( $q_i$ ) is called the isosteric heat of adsorption (kcal/mole). It is given by

$$\text{Pure gas: } q_i^o = RT^2 \left[ \frac{\partial \ln P}{\partial T} \right]_{n_i^o}; \quad \text{Mixed gas: } q_i = RT^2 \left[ \frac{\partial \ln Py_i}{\partial T} \right]_{n_i}, \quad i = 1, 2, \dots \quad (59.6)$$

It follows from Eqs. (59.2), (59.3), (59.4), and (59.6) that

$$K_i = K_i^* \exp[q_i^*/RT]; \quad b_i = b_i^* \exp[q_i^*/RT] \quad (59.7)$$

where  $q_i^*$  is the isosteric heat of adsorption of pure gas  $i$  at the limit  $P \rightarrow 0$  (Henry's law region), and  $K_i^*$  and  $b_i^*$  are constants.

The terms  $q_i^o$  and  $q_i$  are constants and equal to  $q_i^*$  for any adsorbate loading for a homogeneous adsorbent. They are functions of loadings for a heterogeneous adsorbent. Thus, for the Toth model,

$$q_i^o = q_i^* + \left( \frac{RT^2}{k} \right) \left( \frac{d \ln k}{dT} \right) F_i(\theta_i^o); \quad q_i = q_i^* + \left( \frac{RT^2}{k} \right) \left( \frac{d \ln k}{dT} \right) F(\theta) \quad (59.8)$$

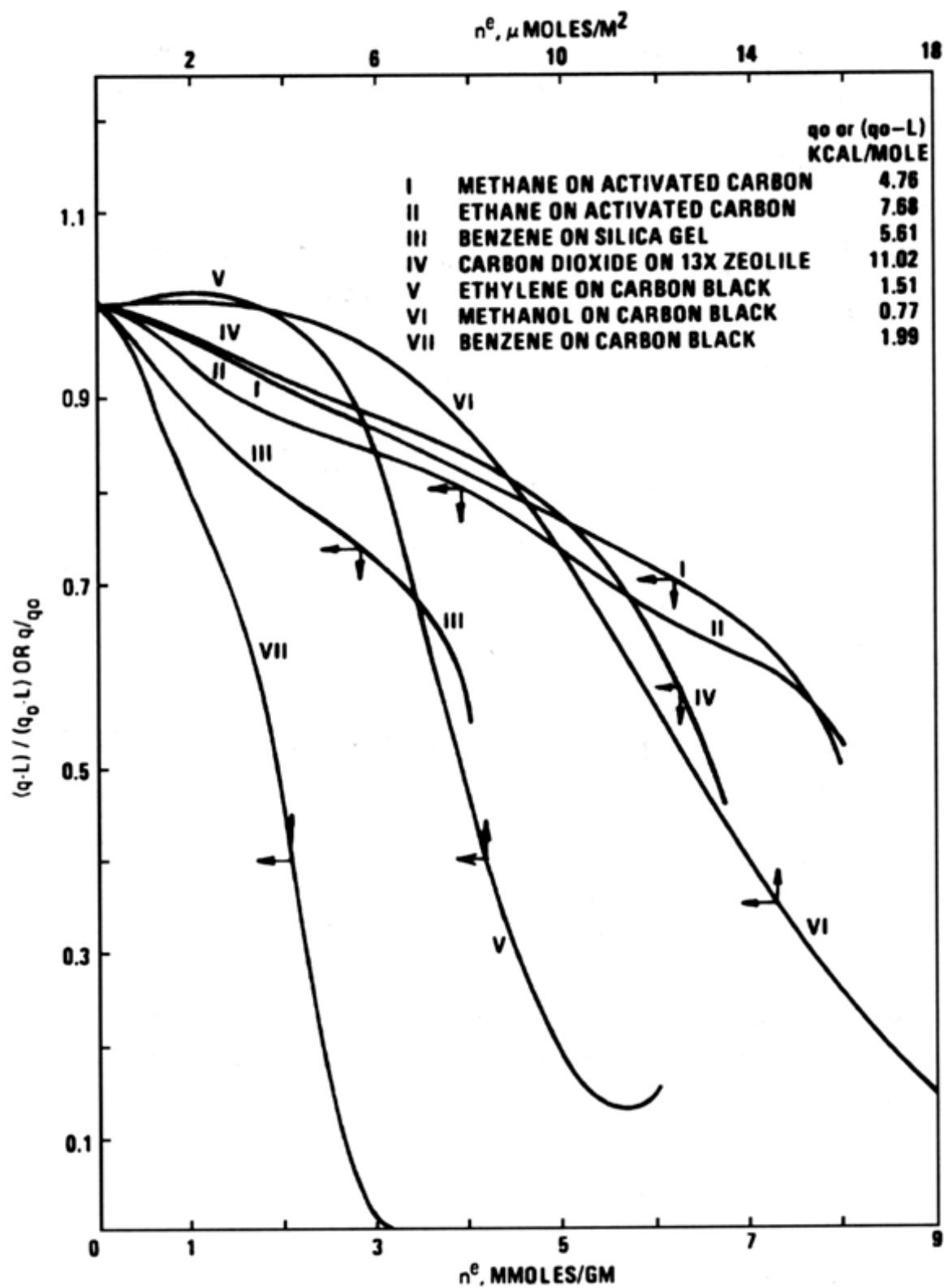
$$F(\theta_i^o) = \frac{[1 - (\theta_i^o)^k] \ln [1 - (\theta_i^o)^k] + (\theta_i^o)^k \ln (\theta_i^o)^k}{[1 - (\theta_i^o)^k]} \quad (59.9)$$

where  $\theta_i^o$  is fractional coverage ( $= n_i^o/m$ ) by pure gas  $i$ . The term  $\theta (= \sum \theta_i)$  is the total fractional coverage ( $= \sum n_i/m$ ) by all adsorbates for the mixture.  $F(\theta)$  has the same mathematical form as Eq. (59.9) except that  $\theta_i^o$  is replaced by  $\theta$ .

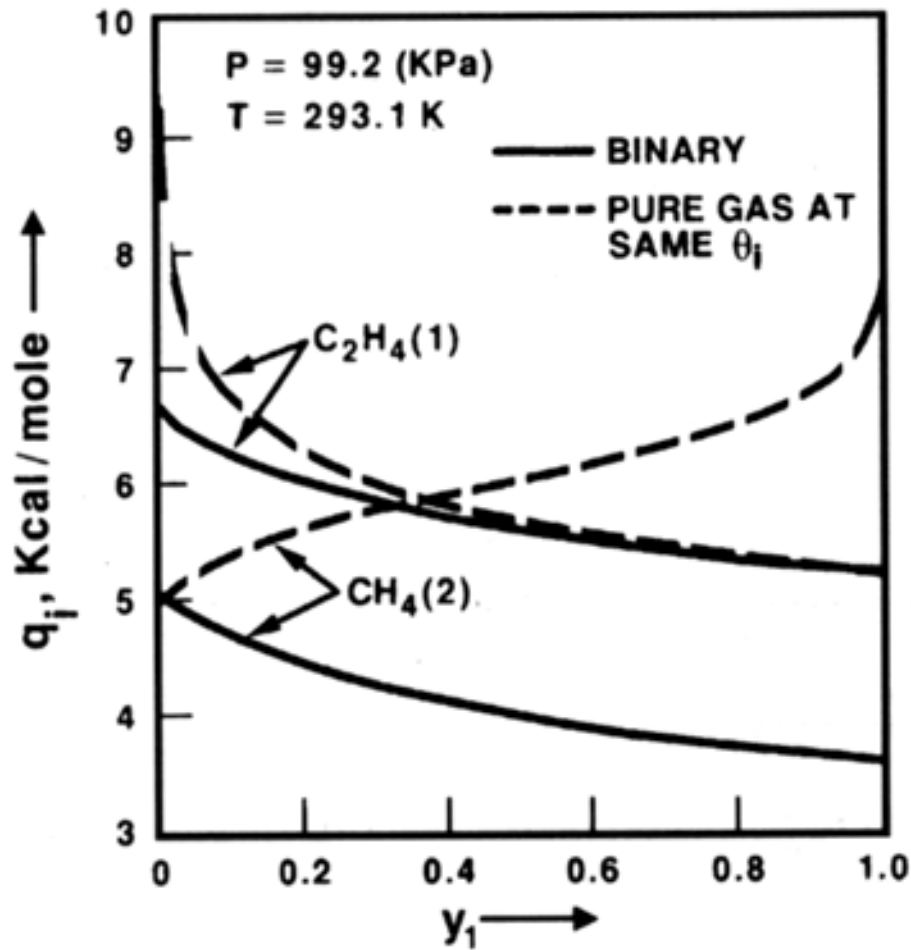
Equation (59.8) shows that  $q_i^o$  or  $q_i$  decreases with increasing  $\theta_i^o$  or  $\theta_i$  for a heterogeneous adsorbent. The higher energy sites of the adsorbent are predominantly filled at lower adsorbate loadings, and the lower energy sites are progressively filled at higher coverages.

Figure 59.5 shows examples of isosteric heats of adsorption of pure gases on heterogeneous adsorbents. Figure 59.6 shows the variation of isosteric heats of the components of a binary gas mixture with coverage (or gas composition) at a constant total gas pressure according to the Toth model.

**Figure 59.5** Isothermic heats of adsorption of pure gases on heterogeneous adsorbents. (Source: Sircar and Gupta, 1981.)



**Figure 59.6** Isosteric heats of adsorption of binary gas on heterogeneous adsorbents. (Source: Sircar, 1991a.)

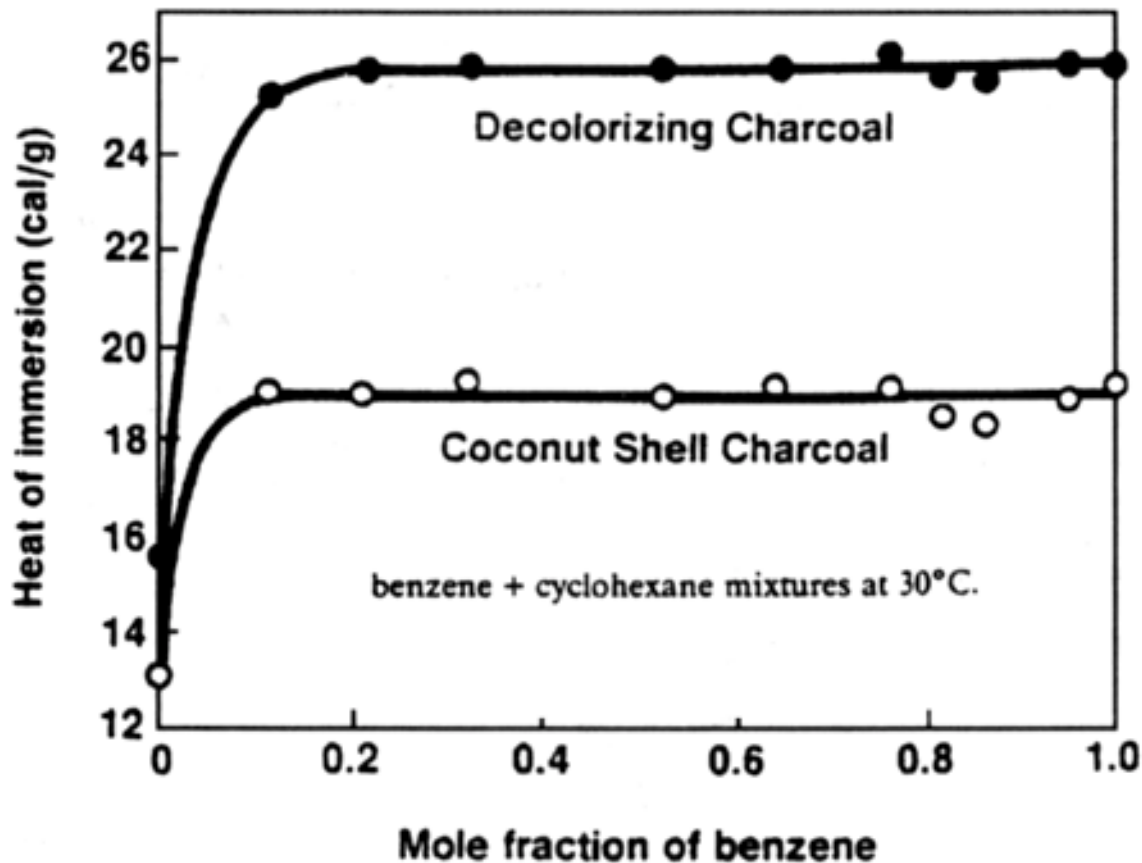


An integral heat of adsorption called *heat of immersion* (kcal/mole) can be measured by contacting a clean adsorbent with a pure liquid ( $\Delta H_i^o$ ) or a liquid mixture ( $\Delta H$ ). The typical variation of ( $\Delta H$ ) as a function of bulk liquid phase concentration is shown in Fig. 59.7. For an ideal binary liquid system [Eq. (59.5)],  $\Delta H$  is given by

$$\Delta H = \frac{S y_i \Delta H_1^0 + y_2 \Delta H_2^0}{S y_1 + y_2} \quad (59.10)$$



**Figure 59.7** Heats of immersion of binary liquid mixtures. (Source: Wright, 1967.)



## 59.4 Thermodynamic Selectivity of Adsorption

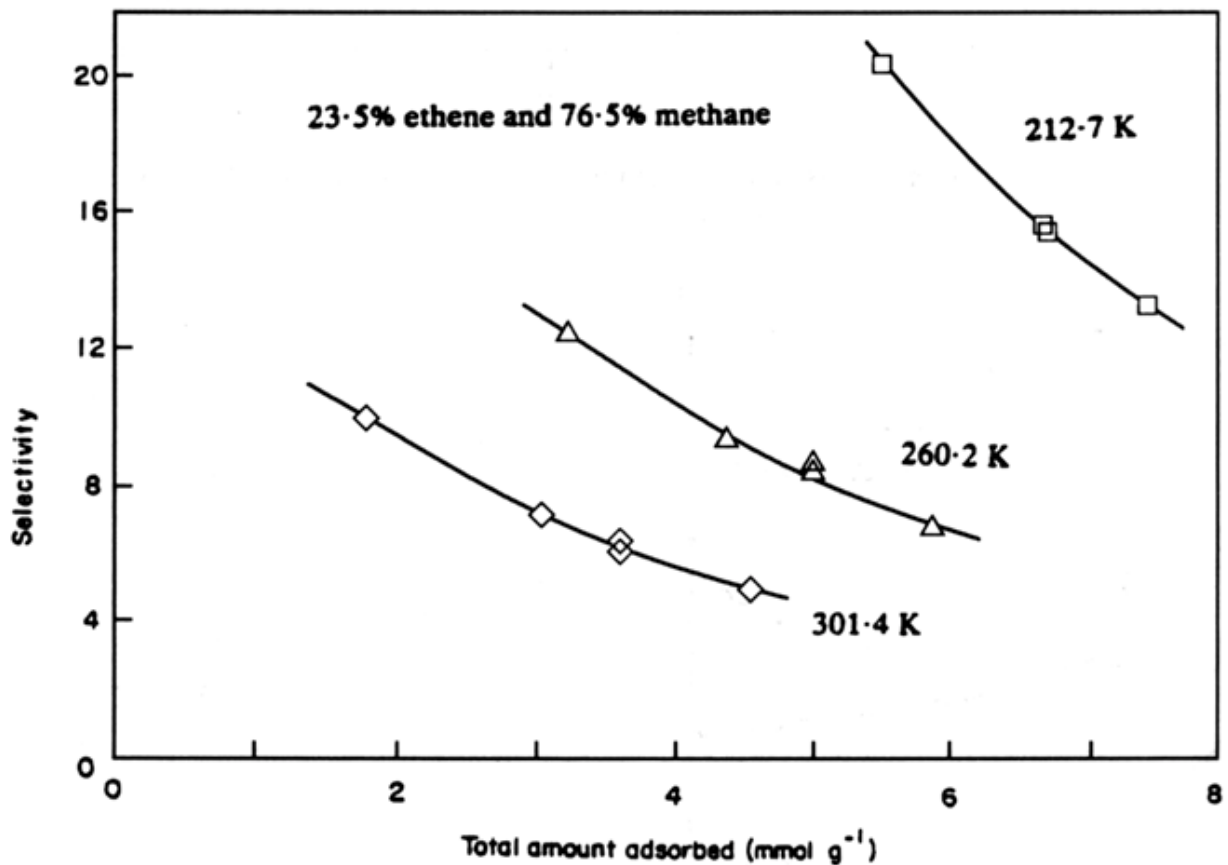
Most practical adsorptive separations are based on thermodynamic selectivity. The selectivity  $S_{ij}$  ( $= n_i y_j / n_j y_i$ ) of adsorption of component  $i$  over component  $j$  from a mixture determines the maximum achievable separation between the components under equilibrium conditions. Component  $i$  is selectively adsorbed over component  $j$  when  $S_{ij} > 1$ .  $S_{ij}$  can approach infinity if component  $j$  is excluded from entering the pores of the adsorbent (molecular sieving).

The selectivity of adsorption ( $S_{ij}^*$ ) in the Henry's law region ( $P \rightarrow 0$ ) is given by

$$S_{ij}^* = \frac{K_i}{K_j} = \left( \frac{K_i^*}{K_j^*} \right) \exp \left[ \frac{q_i^* - q_j^*}{RT} \right] \quad (59.11)$$

$S_{ij}$  values can be strong functions of adsorbate loadings of the components when the adsorbates have different molecular sizes and when the adsorbent is heterogeneous, as shown in Fig. 59.8.

**Figure 59.8** Binary selectivity of adsorption on activated carbon. (Source: Sircar and Myers, 1985.)



For adsorption of an ideal binary liquid system [Eq. (59.5)],

$$S = \exp[\phi_2^o - \phi_1^o]/mRT \quad (59.12)$$

where  $\phi_i^o$  is the surface potential for adsorption of pure liquid  $i$ .  $S$  for liquid mixtures can be strong functions of bulk phase concentrations when the adsorbates have different

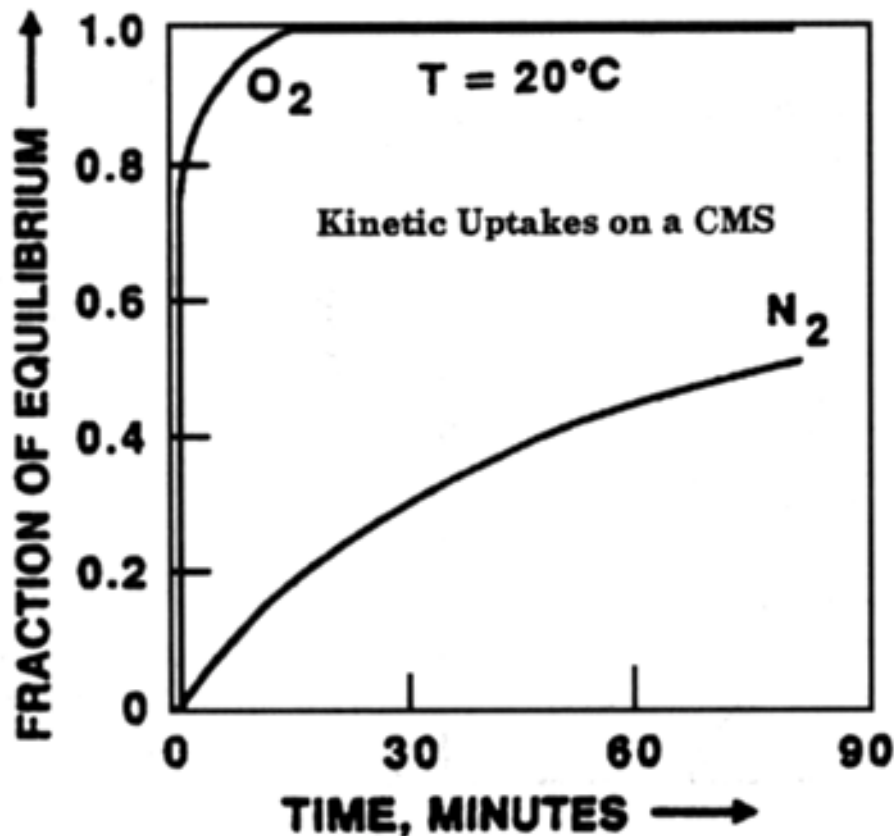
sizes or when the adsorbent is heterogeneous.

## 59.5 Adsorption Kinetics

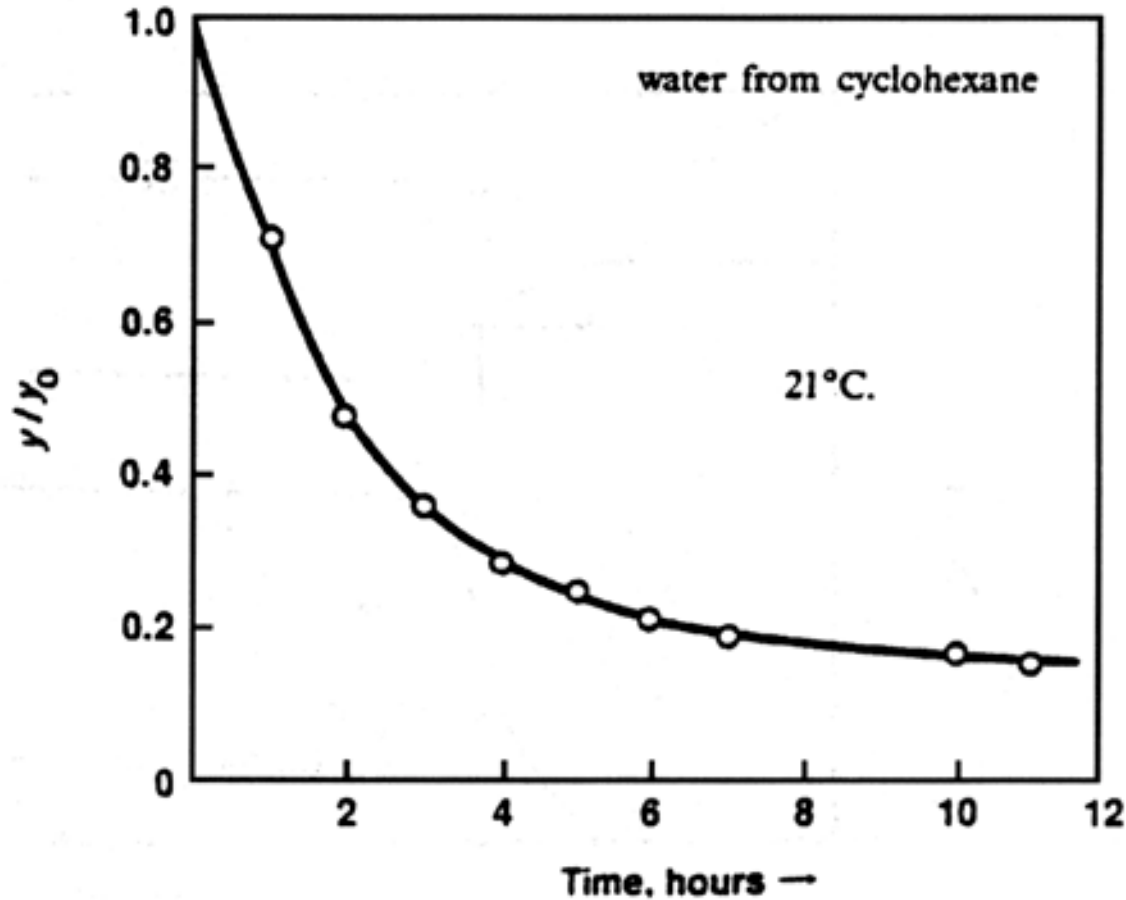
The actual physisorption process is very fast (milliseconds to reach equilibrium). However, a finite amount of time may be required for an adsorbate molecule to travel from the bulk fluid phase to the adsorption site in a microporous adsorbent. This rate process is generally referred to as *adsorption kinetics*. Adsorbate mass transfer resistance may be caused by (a) fluid film outside the adsorbent particle (for mixture adsorption), (b) anisotropic skin at the particle surface, and (c) internal macro- and microporous diffusional resistances (pore and surface diffusion). The transport of an adsorbate can be severely affected by (a) the presence of other adsorbates in the pores and (b) local temperature changes caused by heat of ad(de)sorption. Gas phase adsorption is generally faster than liquid phase adsorption. The overall mass transfer coefficient for adsorption can be increased by reducing the length of diffusional path (adsorbent particle size).

Figure 59.9 shows examples of uptakes of pure gases by a carbon molecular sieve. Figure 59.10 shows an example of uptake of a trace adsorbate from a bulk liquid mixture by a zeolite. The terms  $y$  and  $y_0$  are, respectively, bulk phase mole fractions at time  $t$  and at the start of the test.

**Figure 59.9** Pure gas adsorption kinetics on carbon molecular sieve. (Source: Sircar, 1994.)



**Figure 59.10** Adsorption kinetics of a trace liquid adsorbate on 3A zeolite. (Source: Sircar and Myers, 1986.)



The simplest model for describing gas phase adsorption kinetics is known as the linear driving force (LDF) model:

$$\begin{aligned}
 \text{Pure gas: } \frac{dn_i^o(t)}{dt} &= k_i^o [n_i^{o*}(t) - n_i^o(t)]; \\
 \text{Mixed gas: } \frac{dn_i(t)}{dt} &= k_{ii}[n_i^*(t) - n_i(t)] + \sum_j k_{ij}[n_j^*(t) - n_j(t)]
 \end{aligned}
 \tag{59.13}$$

The terms  $n_i^o(t)$  and  $n_i$  are, respectively, the transient adsorbate loading of pure component  $i$ , and that for component  $i$  from a mixture at time  $t$ . The terms  $n_i^{o*}(t)$  and  $n_i^*(t)$  are, respectively, the corresponding equilibrium adsorbate loadings of component  $i$  at the instantaneous bulk phase conditions. The term  $k_i^o$  (seconds<sup>-1</sup>) is the overall mass transfer coefficient for pure component  $i$ . The terms  $k_{ii}$  and  $k_{ij}$  are the overall straight

and cross (between component  $i$  and  $j$ ) mass transfer coefficients for component  $i$  in the mixture. The temperature and adsorbate loading dependence of the mass transfer coefficients is determined by the governing transport mechanism. The LDF model can also describe the kinetics of adsorption of liquid mixtures by rewriting Eq. (59.13) in terms of transient surface excess  $n_i^e(t)$  and equilibrium surface excess  $n_i^{e*}$  of component  $i$ .

An experimental gas adsorption kinetics process is generally nonisothermal, whereas liquid phase adsorption kinetics can be measured isothermally due to high heat capacity of the liquid.

Separation of the components of a fluid mixture can also be achieved by a kinetic selectivity when certain components of the mixture are adsorbed at a much faster rate than the others, even though there is no thermodynamic selectivity of adsorption between the components. [Figure 59.9](#) is an example of such a case.

## 59.6 Adsorption Column Dynamics

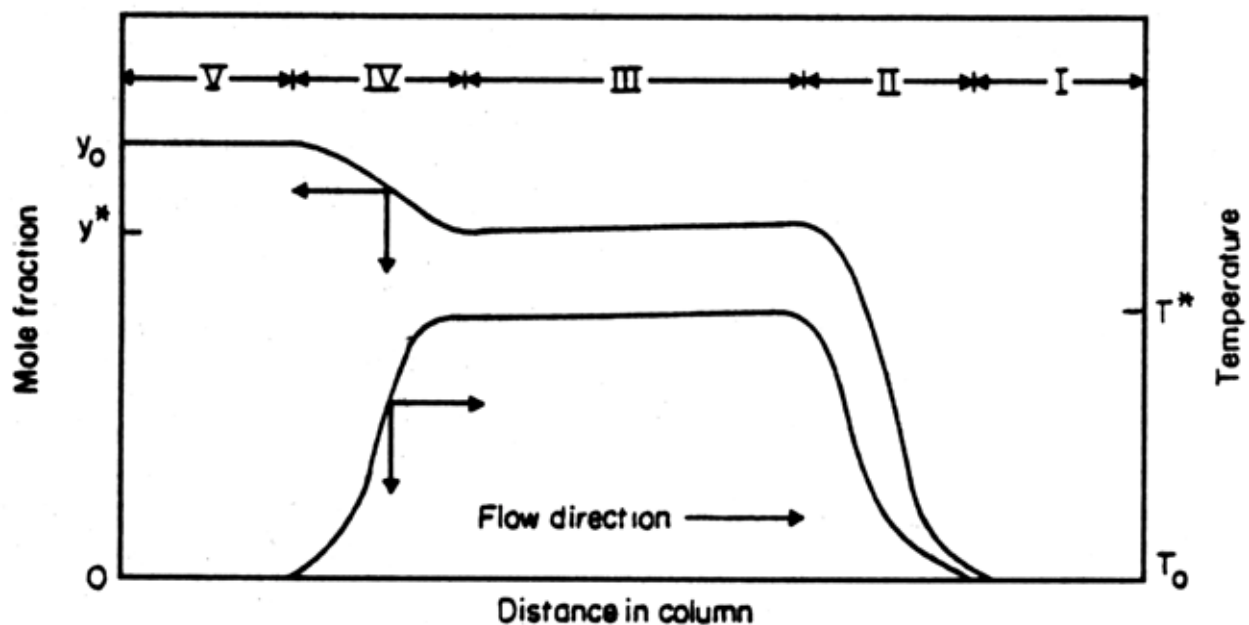
---

Practical separation and purification of fluid mixtures are carried out in packed columns of adsorbent materials. The dynamics of the ad(de)sorption process in columns is determined by the adsorption equilibria, heats, and kinetics and by the modes of operation of the process. The simplest case of the adsorption dynamics is to flow a binary gas mixture consisting of a single adsorbate (mole fraction  $y^o$ ) and an inert carrier gas through a packed column that has previously been saturated with the pure inert gas at the pressure ( $P^o$ ) and temperature ( $T^o$ ) of the feed gas. Two types of behavior may be observed:

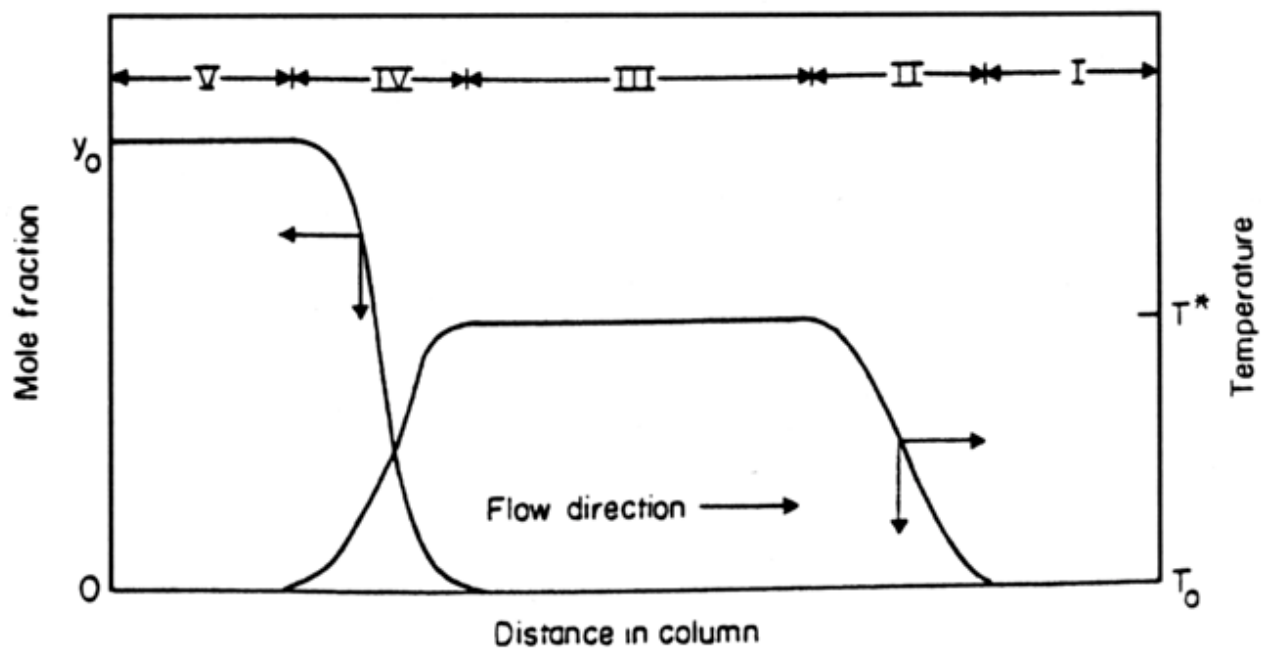
*Type I.* Two pairs of mass and heat transfer zones are formed in the column, as shown in [Fig. 59.11](#). The column ahead (section I) of the front zones (section II) remains saturated with the carrier gas at initial conditions. The column (section III) between the front and rear zones (section IV) is equilibrated with a gas mixture of mole fraction  $y^*$  ( $< y^o$ ) at temperature  $T^*$  ( $> T^o$ ). The column (section V) behind the rear zones is equilibrated with feed gas mixture at feed conditions.

*Type II.* A pure heat transfer zone (section II) is formed, followed by a pair of mass and heat transfer zones (section IV), as shown in [Fig. 59.12](#). The adsorbate is absent in sections I to III in this case. The column behind (section V) the rear zones remains equilibrated with feed gas at feed conditions.

**Figure 59.11** Type I column dynamics. (Source: Sircar and Myers, 1985.)



**Figure 59.12** Type II column dynamics. (Source: Sircar and Myers, 1985.)



Approximate criteria for formation of these two types of systems are as

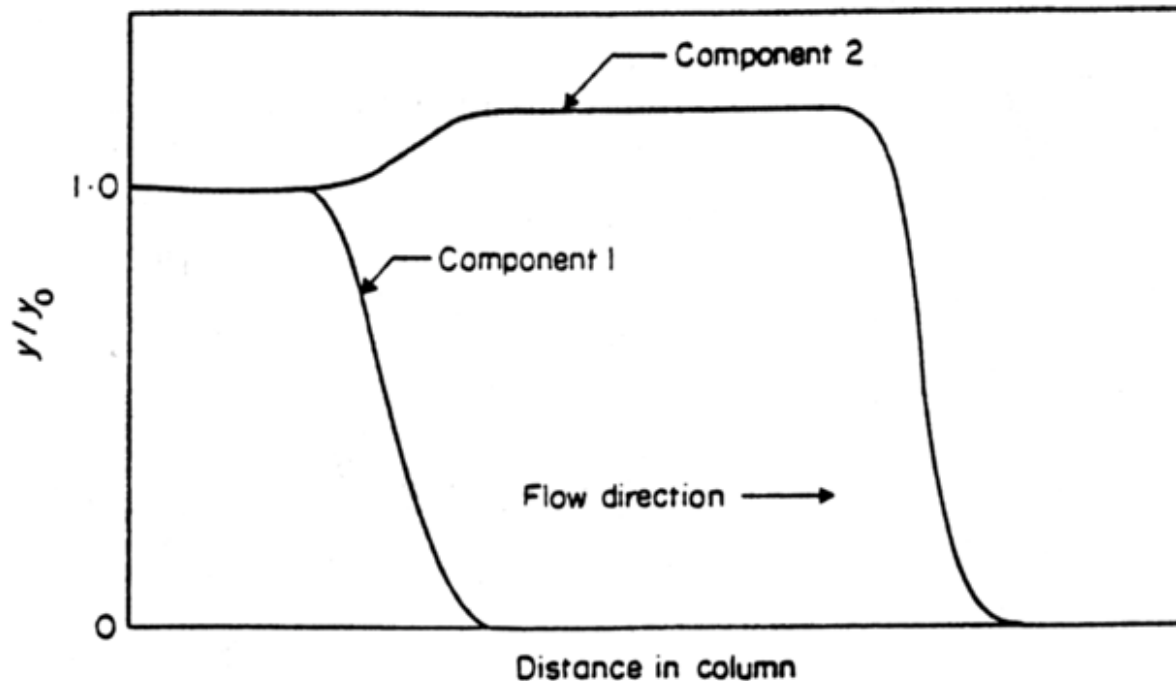
follows:

$$\text{Type I: } n^o/y^o < C_s/C_g; \quad \text{Type II: } n^o/y^o > C_s/C_g \quad (59.14)$$

The term  $n^o$  is adsorbate loading at feed conditions.  $C_s$  and  $C_g$  are heat capacities of the adsorbent and the feed gas. The zones move through the column as more feed gas is introduced.

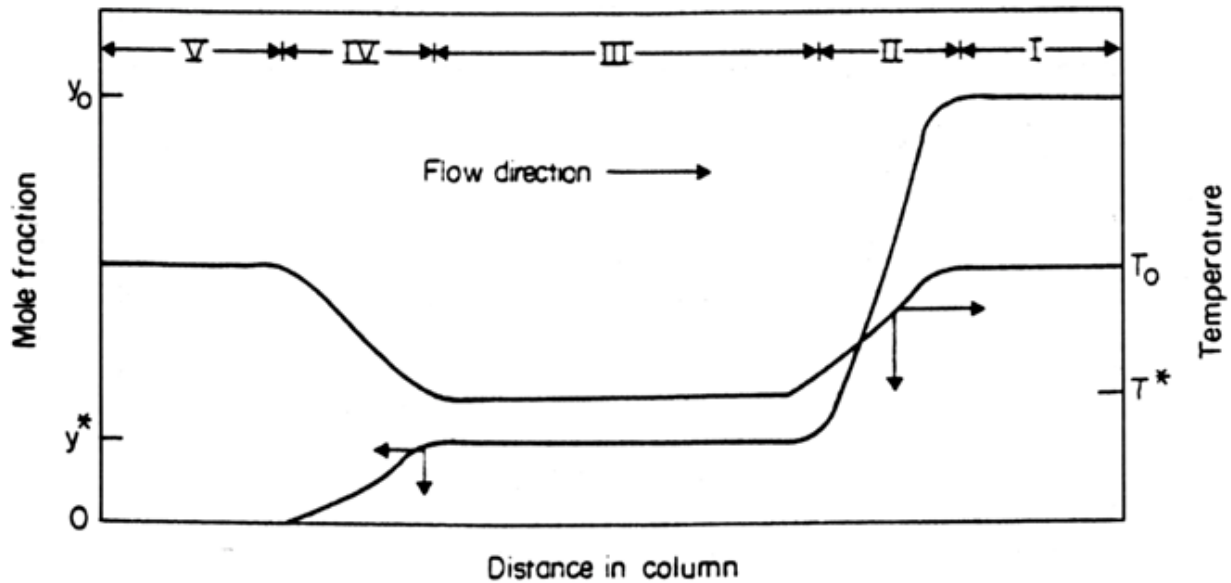
Adsorbates with type I adsorption equilibria generally yield constant pattern (unchanging shape and size) front zones (type I dynamics) and rear zones (type II dynamics) in a long column. The zones can be proportionate pattern (expanding in size with time) when the adsorption isotherms are linear. The zones eventually leave the column, and the measured adsorbate concentration-time and temperature-time profiles (breakthrough curves) are mirror images of these profiles within the column. Multiple transfer zones and equilibrium sections are formed in systems with multicomponent adsorbates. They are often characterized by rollover effects in which the more strongly adsorbed species (component 1) displaces the weaker species (component 2) as the zones propagate. This effect is illustrated in Fig. 59.13 for isothermal adsorption of a binary adsorbate from an inert carrier gas.

**Figure 59.13** Rollover effect for binary adsorbate system. (Source: Sircar and Myers, 1985.)



Well-defined mass and heat transfer zones and equilibrium sections can also be formed in the column during the desorption process. [Figure 59.14](#) is an example for isobaric desorption of a single adsorbate from a column saturated with the adsorbate at mole fraction  $y^o$  and temperature  $T^o$  that is being purged with an inert gas at  $T^o$ . Again, two pairs of mass and heat transfer zones (sections II and IV) and three equilibrium sections (I, III, V) are formed.

**Figure 59.14** Column dynamics for desorption by purge. (Source: Sircar and Myers, 1985.)



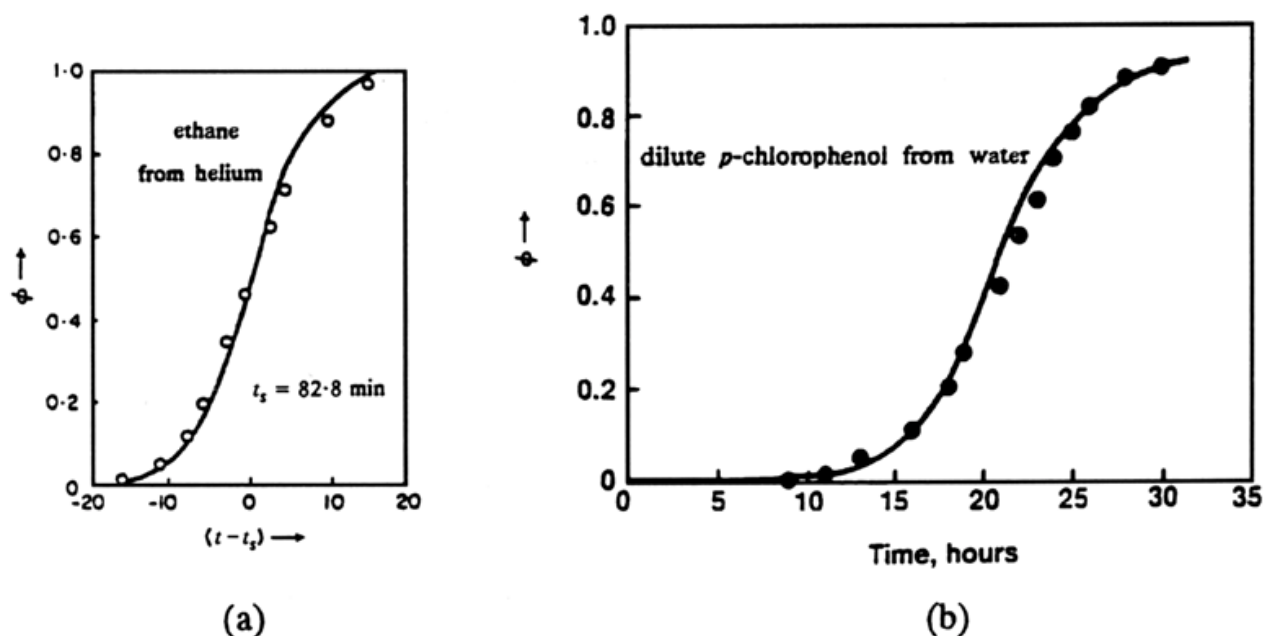
Nonisothermal column dynamics is generally a rule. Near-isothermal dynamics (mass transfer zones only) can be obtained for ad(de)sorption from (a) liquid mixtures and (b) trace gaseous adsorbates. [Figure 59.15](#) shows two examples of isothermal breakthrough curves for adsorption of trace single adsorbate from (a) an inert gas and (b) an inert liquid. For isothermal-isobaric adsorption of a trace Langmuirian adsorbate  $i$  from an inert carrier gas in a column, which is initially free of the adsorbate, and where a constant pattern mass transfer zone is formed, the LDF mechanism of mass transfer yields

$$(t_2 - t_1) \cong \frac{b_1 m}{(1 + b_1) k_i^o n_i^o} \ln \left[ \frac{\phi_2 (1 - \phi_1)}{\phi_1 (1 - \phi_2)} \right] + \ln \frac{\phi_1}{\phi_2} \quad (59.15)$$

The terms  $t_1$  and  $t_2$  correspond to time difference in the breakthrough curve corresponding to two arbitrary composition levels  $\phi_1$  and  $\phi_2$  [ $\phi = y_i(t)/y_i^o$ ].



**Figure 59.15** Isothermal breakthrough profiles for a single-trace adsorbate. [Source: (a) Garg and Ruthven, 1974. (b) Mathews, 1984.]



## 59.7 Adsorptive Separation Processes and Design

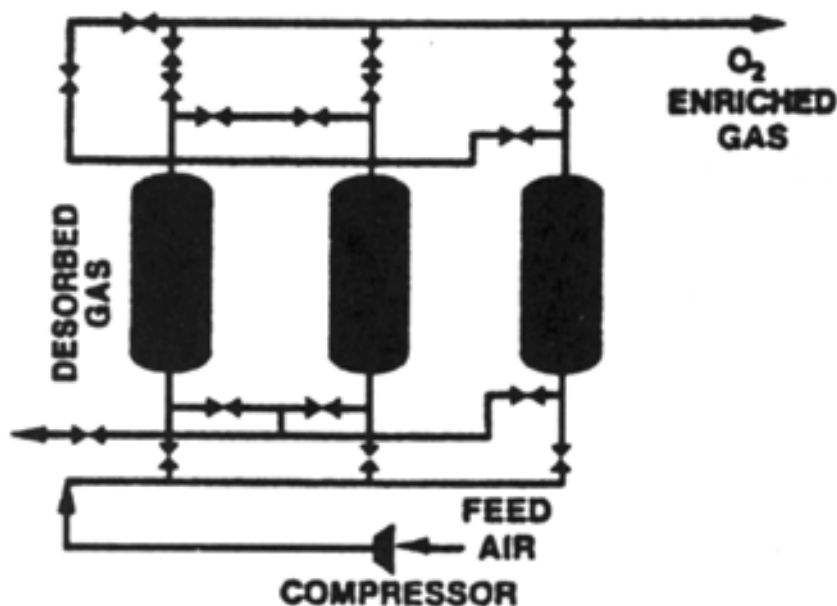
Three generic adsorptive process schemes have been developed to serve most of the applications described by Table 59.1. These are (a) **thermal swing adsorption (TSA)**, (b) **pressure swing adsorption (PSA)**, and (c) **concentration swing adsorption (CSA)**.

TSA is by far the most frequently used industrial application of this technology. It is used for removal of trace impurities as well as for drying gases and liquids. The adsorption is carried out at near ambient temperature and the desorption is achieved by directly heating the adsorbent using a part of the cleaned fluid or steam. The adsorbent is then cooled and reused.

PSA is primarily used for bulk gas separations and for gas drying. These processes consist of a series of sequential cyclic steps. The adsorption step is carried out at higher partial pressures of the adsorbates and the desorption is achieved by lowering their partial pressures in the column by (a) decreasing the total pressure of the column, and (b) flowing an adsorbate free gas through the column. Many complementary process steps are also used in modern PSA processes in order to (a) increase the recovery and purity of desired products from a multicomponent feed gas mixture, (b) produce more

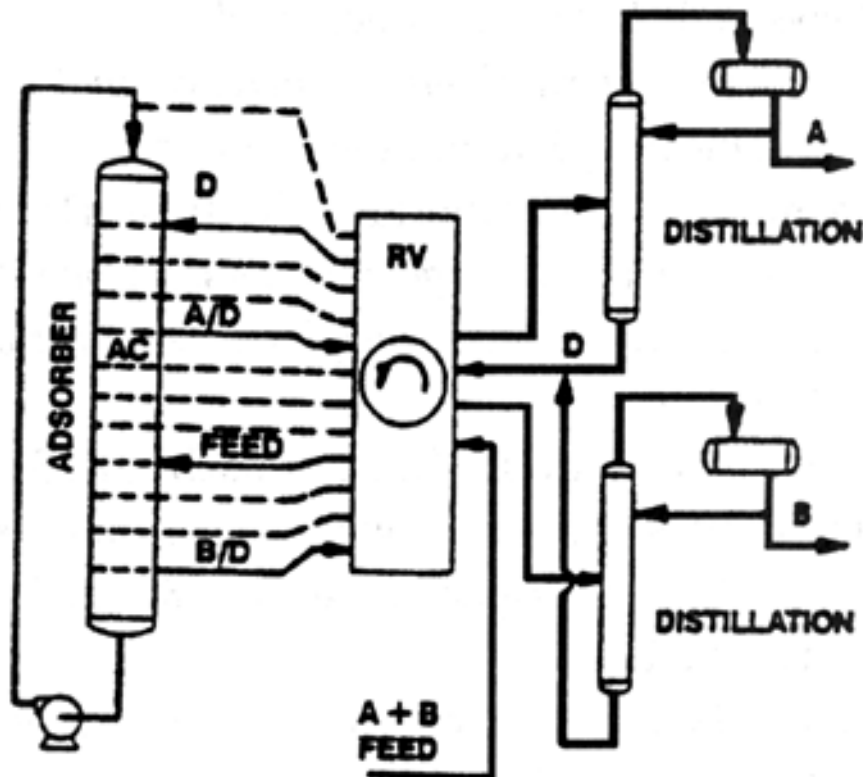
than one pure product, (c) decrease adsorbent inventory, and (d) reduce energy of separation. Figure 59.16 shows a schematic flowsheet for a PSA process.

**Figure 59.16** Schematic flowsheet of a PSA process. (Source: Sircar, 1989.)



The CSA processes are generally designed for separation of bulk liquid mixtures. Certain components of a liquid mixture are adsorbed at ambient conditions and the desorption is effected by flowing a less strongly adsorbed liquid (eluent) over the adsorbent. Simple distillation may be necessary to separate the eluent from the components of the feed mixture. Simulated moving bed (SMB) adsorbers have been designed for this purpose, as shown schematically in Fig. 59.17. The feed and eluent injection points as well as the liquid withdrawal points are changed periodically in a fixed column to simulate continuous countercurrent operation.

**Figure 59.17** Schematic flowsheet of an SMB process. (Source: Keller et al., 1987.)



The design of adsorptive processes requires simultaneous solutions of differential mass, heat, and momentum balance equations describing the operations of the cyclic process steps in the adsorbent column using the appropriate initial and boundary conditions for each step. The final column conditions at the end of a step become the initial conditions for the next step. Numerical integration of the equations is often necessary in order to reach a cyclic steady state solution. Multicomponent adsorption equilibria, heats, and kinetics form the key input variables for the solution. Bench- and pilot-scale process performance data are generally needed to confirm design calculations.

## Defining Terms

**Adsorbent:** A material used for carrying out the adsorption process.

**Adsorption:** The surface phenomenon by which the molecules of a bulk fluid phase are attracted by a solid surface in contact with the fluid.

**Adsorption equilibria:** The thermodynamic property describing the extent of adsorption of a fluid species by a solid surface.

**Adsorption kinetics:** The measure of travel time of an adsorbate molecule from bulk

fluid phase to the adsorption site.

**Column dynamics:** Defines behavior of mass and heat transfer zone movements within an adsorption column during the ad(de)sorption process.

**Concentration swing adsorption:** An adsorptive process in which desorption is effected by changing the fluid phase concentration of the adsorbate.

**Desorption:** The process of removing the adsorbed molecules from the solid surface to the bulk fluid phase.

**Heat of adsorption:** The measure of thermal energy released during the exothermic adsorption process.

**Pressure swing adsorption:** An adsorptive process in which desorption is effected by lowering the partial pressure of the adsorbate.

**Selectivity of adsorption:** The measure of extent of separation of a component of a fluid mixture by adsorption process.

**Thermal swing adsorption:** An adsorptive process in which desorption is effected by heating the adsorbent.

## References

- Broughton, D. B. and Gembicki, S. A. 1984. Adsorptive separations by simulated moving bed technology. In *Fundamentals of Adsorption*, ed. A. L. Myers and G. Belfort, p. 115. Engineering Foundation, New York.
- Garg, D. R. and Ruthven, D. M. 1974. The performance of molecular sieve adsorption columns: System with micropore diffusion control. *Chem. Eng. Sci.* 29:571–581.
- Keller, G. E, Anderson, R. A., and Yon, C. M. 1987. Adsorption. In *Handbook of Separation Process Technology*, ed. R. W. Rousseau, p. 644, John Wiley & Sons, New York.
- Kumar, R. and Sircar, S. 1986. Skin resistance for adsorbate mass transfer into extruded adsorbent pellets. *Chem. Eng. Sci.* 41:2215–2223.
- Mathews, A. P. 1984. Dynamics of adsorption in a fixed bed of polydisperse particles. In *Proceedings of First International Conference on Fundamentals of Adsorption*, ed. A. L. Myers, p. 345. Engineering Foundation, New York.
- Rahman, M. A. and Ghosh, A. K. 1980. Determination of specific surface area of ferric oxide, alumina and silica gel powders. *J. Colloid Interface Sci.* 77:50–52.
- Ruthven, D. M. 1984. *Principles of Adsorption and Adsorption Processes*. John Wiley & Sons, New York.
- Ruthven, D. M., Farouq, S., and Knaebel, K. S. 1994. *Pressure Swing Adsorption*.

VCH, New York.

- Sircar, S. 1989. Pressure swing adsorption technology. In *Adsorption Science and Technology*, ed. A. I. Rodrigues *et al.*, p. 285. NATO ASI Series E158. Kluwer Academic, Dordrecht, The Netherlands.
- Sircar, S. 1991a. Isosteric heats of multicomponent gas adsorption on heterogeneous absorbents. *Langmuir*. 7:3065–3069.
- Sircar, S. 1991b. Pressure swing adsorption—Research needs by industry. In *Proceedings of Third International Conference on Fundamentals of Adsorption*, ed. A. Mersmann, p. 815. Engineering Foundation, New York.
- Sircar, S. 1993. Novel applications of adsorption technology. In *Proceedings of Fourth International Conference on Fundamentals of Adsorption*, ed. M. Suzuki, p. 3. Kodansha, Tokyo.
- Sircar, S. 1994. Adsorption technology: A versatile separation tool. In *Separation Technology: The Next Ten Years*, ed. J. Garside, p. 49. Institute of Chemical Engineers, Rugby, Warwickshire, U. K.
- Sircar, S. and Gupta, R. 1981. A semi empirical adsorption equation for single component gas solid equilibrium. *Am. Insti. Chem. Eng. J.* 27:806–812.
- Sircar, S. and Myers, A. L. 1985. Gas adsorption operations: Equilibrium, kinetics, column dynamics and design. *Adsorp. Sci. Technol.* 2:69–87.
- Sircar, S. and Myers, A. L. 1986. Liquid adsorption operations: Equilibrium, kinetics, column dynamics, and applications. *Sep. Sci. Technol.* 21:535–562.
- Sircar, S., Novosad, J., and Myers, A. L. 1972. Adsorption from liquid mixtures on solids: Thermodynamics of excess properties and their temperature coefficients. *Ind. Eng. Chem. Fundam.* 11:249–254.
- Wright, E. H. M. 1967. Thermodynamic correlation for adsorption from non-ideal solutions at the solid-solution interface. *Trans. Faraday Soc.* 63:3026–3038.
- Yang, R. T. 1987. *Gas Separation by Adsorption Processes*. Butterworths, London.
- Young, D. M. and Crowell, A. D. 1962. *Physical Adsorption of Gases*. Butterworths, London.

## Further Information

- Barrer, R. M. 1978. *Zeolites and Clay Minerals as Sorbents and Molecular Sieves*. Academic Press, New York.
- Breck, D. W. 1974. *Zeolite Molecular Sieves*. Wiley-Interscience, New York.
- Capelle, A. and deVooy, F. (Eds.) 1983. *Activated Carbon—A Fascinating Material*. Norit N. V., Amersfoort, The Netherlands.

- Gregg, S. J. and Sing, K.S.W. 1982. *Adsorption Surface Area and Porosity*. Academic Press, London.
- Karger, J. and Ruthven, D. M. 1992. *Diffusion in Zeolites*. Wiley-Interscience, New York.
- Rousseau, R. W. (Ed.) 1981. *Handbook of Separation Technology*. John Wiley & Sons, New York.
- Suzuki, M. 1990. *Adsorption Engineering*. Kodansha, Tokyo.
- Wankat, P. C. 1986. *Large Scale Adsorption and Chromatography*. CRC, Boca Raton, FL.

Bennett, R. C. "Crystallization and Evaporation"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Crystallization and Evaporation

---

- 60.1 Methods of Creating Supersaturation
- 60.2 Reasons for the Use of Crystallization
- 60.3 Solubility Relations
- 60.4 Product Characteristics
- 60.5 Impurities Influencing the Product
- 60.6 Kinds of Crystallization Processes
- 60.7 Calculation of Yield in a Crystallization Process
- 60.8 Mathematical Models of Continuous Crystallization
- 60.9 Equipment Designs
- 60.10 Evaporation

**Richard C. Bennett**

*Swenson Process Equipment, Inc.*

A **crystal** is a solid bounded by plane surfaces. Crystallization is important as an industrial process because there are a large number of commodity chemicals, pharmaceuticals, and specialty chemicals that are marketed in the form of crystals. Its wide use is due to the highly purified and attractive form in which the compounds can be obtained from relatively impure solutions by means of a single processing step. Crystallization can be performed at high or low temperatures, and generally requires much less energy for separation of pure materials than other commonly used methods of purification. While crystallization may be carried on from vapor or a melt, the most common industrial method is from a solution.

A solution is made up of a liquid (solvent)—most commonly water—and one or more dissolved species which are solid in their pure form (solute). The amount of solute present in solution may be expressed in several different units of concentration. For engineering calculations, expressing the solubility in mass units is the most useful. The solubility of a material is the maximum amount of solute which can be dissolved in a solvent at a particular temperature. Solubility varies with temperature and, with most substances, the amount of solute dissolved increases with increasing temperature.

For crystallization to occur, a solution must be supersaturated. **Supersaturation** means that, at a given temperature, the actual solute concentration exceeds the concentration under equilibrium or saturated conditions. A supersaturated solution is metastable and all crystallization occurs in the metastable region. A crystal suspended in saturated solution will not grow. Supersaturation may be expressed as the ratio between the actual concentration and the concentration at saturation [Eq. (60.1)] or as the difference in concentration between the solution and the saturated solution at the



same temperature [Eq. (60.2)].

$$S = C/C_s \quad (60.1)$$

$$\Delta C = C - C_s \quad (60.2)$$

where  $C$  is the concentration (g/100 g of solution), and  $C_s$  is the concentration (g/100 g of solution) at saturation. This difference in concentration may also be referenced to the solubility diagram and expressed as degrees ( $^{\circ}\text{C}$ ) of supersaturation.

Nucleation is the birth of a new crystal within a supersaturated solution. Crystal growth is the layer-by-layer addition of solute to an existing crystal. Both of these phenomena are caused by supersaturation. Nucleation is a relatively rapid phenomenon that can occur in a matter of seconds. Growth is a layer-by-layer process on the surface of an existing crystal and takes considerably more time. The ratio of nucleation to growth controls the size distribution of the crystal product obtained. Generating a high level of supersaturation spontaneously leads to both nucleation and growth. The competition between these two processes determines the character of the product produced.

## 60.1 Methods of Creating Supersaturation

---

Supersaturation may be created by cooling a solution of normal solubility into the metastable zone. Typically, the amount of supersaturation which can be created in this way without causing spontaneous nucleation is in the range of  $1\text{--}2^{\circ}\text{C}$ . Evaporation of solvent at a constant temperature also produces supersaturation by reducing the amount of solvent available to hold the solute. The reaction of two or more chemical species, which causes the formation of a less soluble species in the solvent, can also produce supersaturation. Finally, the addition of a miscible nonsolvent in which the solute is not soluble to a solvent will cause a decrease in the solubility of the solute in the solution. This technique is most often used in pharmaceutical operations involving the addition of alcohol or similar solvents to the primary solvent (water).

## 60.2 Reasons for the Use of Crystallization

---

Crystallization is important as an industrial process because there are a large number of materials that can be marketed in the form of crystals which have good handling properties. Typically, crystalline materials can be separated from relatively impure solutions in a single processing step. In terms of energy requirements, the energy required for crystallization is typically much less than for separation by distillation or other means. In addition, crystallization can often be done at relatively low temperatures on a scale that involves quantities from a few pounds up to thousands of tons per day.

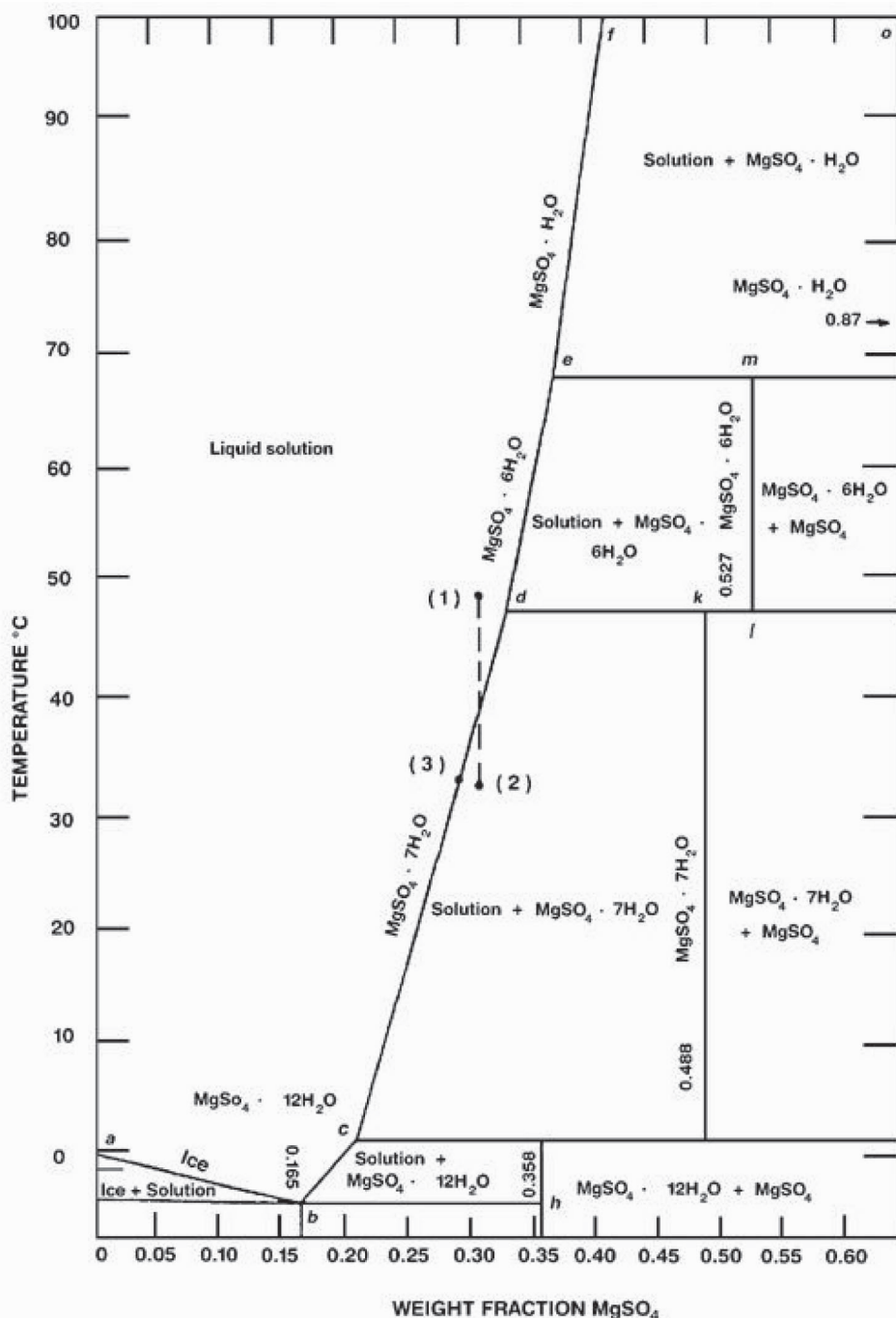
## 60.3 Solubility Relations

---

Equilibrium relations for crystallization systems are expressed in the form of solubility data which

are plotted as phase diagrams or solubility curves. The starting point in designing any crystallization process is knowledge of the solubility curve, which is ordinarily plotted in terms of mass units as a function of temperature. An example is given in Fig. 60.1 for the solubility of magnesium sulfate in water as a function of temperature. At any concentration and temperature, the information on the diagram allows one to predict the mixture of solids and solution that exist. Note that, in the case of magnesium sulfate, a number of different hydrates can exist in addition to the solution itself, or ice plus the solution. The line that forms a boundary between the solution area and the various crystal hydrate areas is a solubility curve.

**Figure 60.1** Weight fraction of  $\text{MgSO}_4$  versus temperature. (Source: Courtesy of Swenson Process Equipment Inc.)



Starting from point (1) at 50°C and cooling to 30°C at point (2) is a path that crosses the solubility line. During the cooling process, crossing the line in this manner indicates that the solution has become supersaturated for the concentration in question. If the supersaturation is within the metastable range—which is approximately 1°C—then growth can occur on existing crystals, but no substantial amount of nucleation will occur. If the cooling proceeds further, the system can become unstably supersaturated and spontaneous nucleation takes place. If spontaneous nucleation takes place, very small crystals or nuclei form, and they will grow as long as the solution remains supersaturated.

As growth takes place, the concentration drops in the direction of point (3), and, as it approaches the solubility line, growth ceases because the driving force approaches zero. Organic and inorganic materials have similar solubility curves and they vary in concentration and temperature for each compound. Some materials have no hydrates and others exhibit a wide range of hydrates similar to those shown in Fig. 60.1. Solubility information on most compounds is available from the literature in publications, such as the *International Critical Tables* [Campbell and Smith, 1951] and *Lang's Solubility of Inorganic and Organic Compounds* [Linke, 1958], and in various software packages which are becoming available.

## 60.4 Product Characteristics

---

The shape and size of a crystal are determined by its internal structure as well as external conditions that occur during its growth cycle. These external conditions include growth rate, the solvent system, the level of agitation, and the effect of impurities that may be present. Crystalline material is almost always separated from its mother liquor before the crystal can be dried or used. The filterability of the crystals, whether separation is done on a centrifuge or filter, is an important characteristic of the product. Generally, larger particles filter more readily, but the average particle size by itself is not an unfailing indication of filterability. The particle size distribution is important because, if it is very broad, small particles may be trapped between the larger particles, making the drainage rates much lower. This could lead to retention of mother liquor, which will degrade the purity of the final product. Broad distributions which increase the amount of mother liquor retained also make the cake less pervious to wash liquids. Products crystallized from continuous crystallizers typically have a coefficient of variation of 45–50%. Products made from batch crystallizers, which are fully seeded, often show narrower size distributions with a coefficient of variation of approximately 25–30%.

The bulk density of the dried material is affected not only by the crystal density itself, but also by the size distribution. A broader distribution leads to tighter packing and, therefore, a higher bulk density. The flow properties of a crystal product are affected by the crystal shape. Rounded crystals which are formed under conditions of relatively high attrition flow very well, particularly if the particles are in the size range of –8 to +30 U.S. Mesh.

## 60.5 Impurities Influencing the Product

---

Since crystallization is generally done to produce high-purity products, it is important that the

crystal be grown in such a way that the impurities which are part of the mother liquor are not carried out with the crystalline particles. Impurities can affect the growth rate and nucleation rate during crystallization and, as a consequence, affect both the mean particle diameter and the habit of the particles being crystallized. Most habit modifiers cause a change in the crystal shape because they are absorbed on one or more of the crystal faces, thereby altering the growth rate of that face and causing that face to either become predominant or largely disappear. Impurities which have this influence can be either ionic, surface-active compounds, or polymers.

Under some conditions, the impurities in a product can be increased by lattice incorporation, which occurs when an impurity in the mother liquor substitutes for molecules in the product crystal lattice. Mixed crystals—which are really two separate species crystallizing at the same time—can also be produced. Surface absorption of species that are in the mother liquor not only can add to the impurity level of the product, but can also alter the growth rate and, therefore, the habit of the crystals. Solvent inclusion can occur when rapidly growing crystals form around small volumes of mother liquor which become trapped inside the crystal lattice. The liquor in these inclusions may or may not find its way to the surface during the subsequent drying operations.

Solvent inclusion probably accounts for the largest increase in impurity levels in a crystal, with lattice incorporation generally less, and surface absorption accounting for only very minute amounts of contamination. Normally, recrystallization from a relatively pure solution will eliminate virtually all the impurities, except for a material whose presence is due to lattice incorporation.

## **60.6 Kinds of Crystallization Processes**

---

Crystallization can be carried on in either a batch or continuous manner, irrespective of whether evaporation, cooling, or solvent change is the method of creating supersaturation. Batch processes are almost always used for small capacities and have useful application for large capacities when a very narrow particle size distribution is required, such as with sugar, or when materials (e.g., pharmaceuticals) that require very accurate inventory control are being handled.

A continuous crystallization process normally must operate around the clock because the retention times typically used in crystallizers range from about one to six hours. As such, it takes at least four to six retention times for the crystallizer to come to equilibrium, which means there may be off-spec product when the system is started up. To minimize this, the unit should be kept running steadily as long as possible. The cost of at least three operators per day and the instrumentation required to continuously control the process represent a substantially greater investment than what is required for batch processing. This disadvantage can only be overcome by utilizing that labor and investment at relatively high production rates.

## 60.7 Calculation of Yield in a Crystallization Process

---

In order to calculate the yield in a crystallization process, it is necessary that the concentration of feed, mother liquor, and any change in solvent inventory (evaporation) be known. In most crystallization processes, the supersaturation in the residual mother liquor is relatively small and can be ignored when calculating the yield. With some materials, such as sugar, a substantial amount of supersaturation can exist, and under such circumstances the exact concentration of the solute in the final mother liquor must be known in order to make a yield calculation. The product crystal may be hydrated, depending on the compound and temperature at which the final crystal is separated from the mother liquor.

Shown below is a formula method for calculating the yield of a hydrated crystal from a feed solution [Myerson, 1993].

$$P = R \frac{100W_o - S(H_o - E)}{100 - S(R - 1)} \quad (60.3)$$

where

$P$  = weight of product

$R = \frac{\text{mole weight of hydrate crystal}}{\text{mole weight of anhydrous crystal}}$

$S$  = solubility at the mother liquor (final) temperature in units/100 units of solvent

$W_o$  = weight of anhydrous solute in feed

$H_o$  = weight of solvent in feed

$E$  = evaporation

## 60.8 Mathematical Models of Continuous Crystallization

---

Randolph and Larsen [1988] developed a method of modeling continuous crystallizers in which the growth rate is independent of size and the slurry is uniformly mixed. Such crystallizers are often referred to as the mixed-suspension mixed-product removal (MSMPR) type. For operation under steady conditions, the population density of an MSMPR crystallizer (FC and DTB types shown in Fig. 60.2 and Fig. 60.3, respectively) is

$$n = n^o e^{-L/GT} \quad (60.4)$$

where

$n$  = population density, number/mm

$G$  = growth rate, mm/h

$T$  = retention time, h

$L$  = characteristic length, mm

$n^o$  = nuclei population density (i.e., intercept at size  $L = 0$ )

A plot of the  $\ln n$  versus  $L$  will be a straight line if the system is operating under the conditions assumed above. The nucleation rate and the mean particle size (by weight) are

$$B^o = Gn^o \quad (60.5)$$

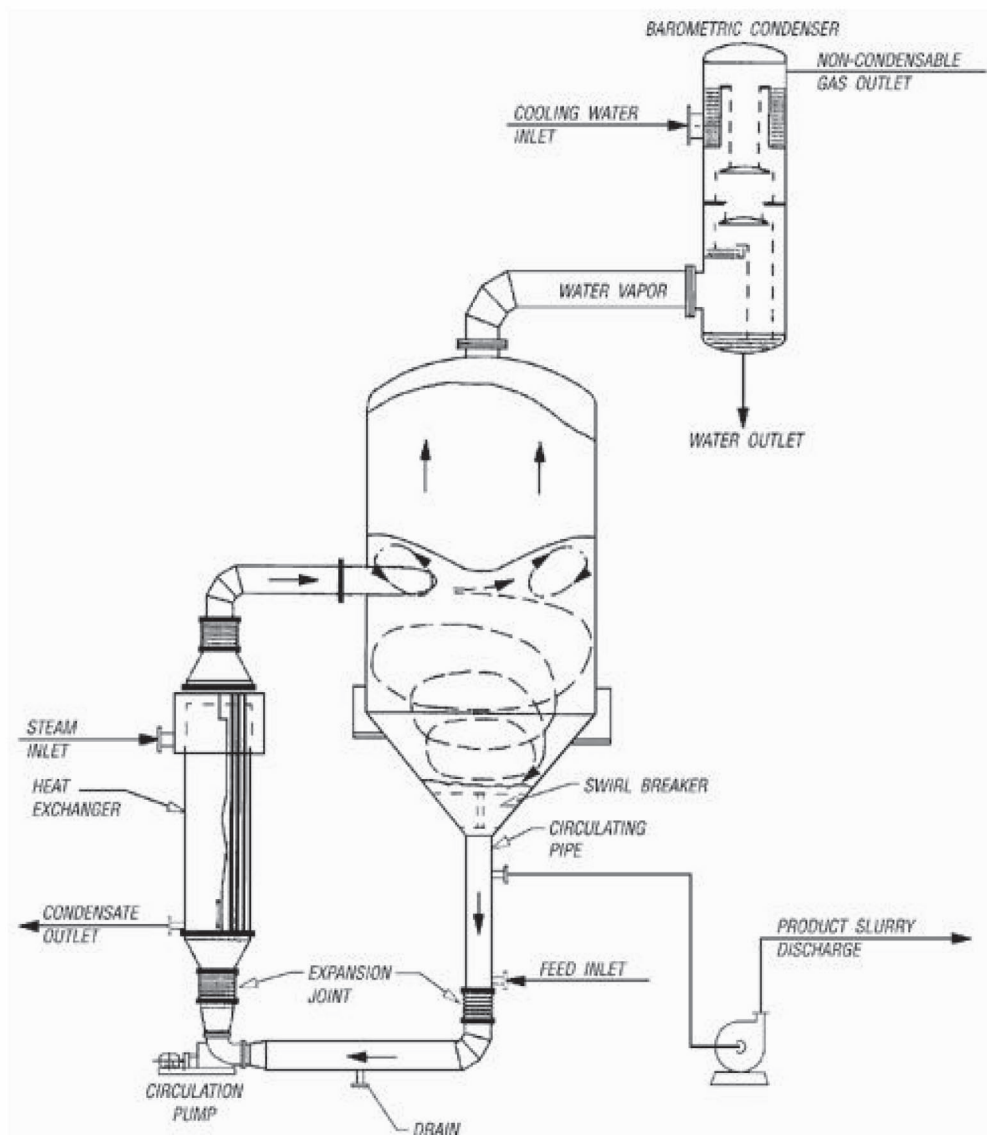
$$L_a = 3.67GT \quad (60.6)$$

where

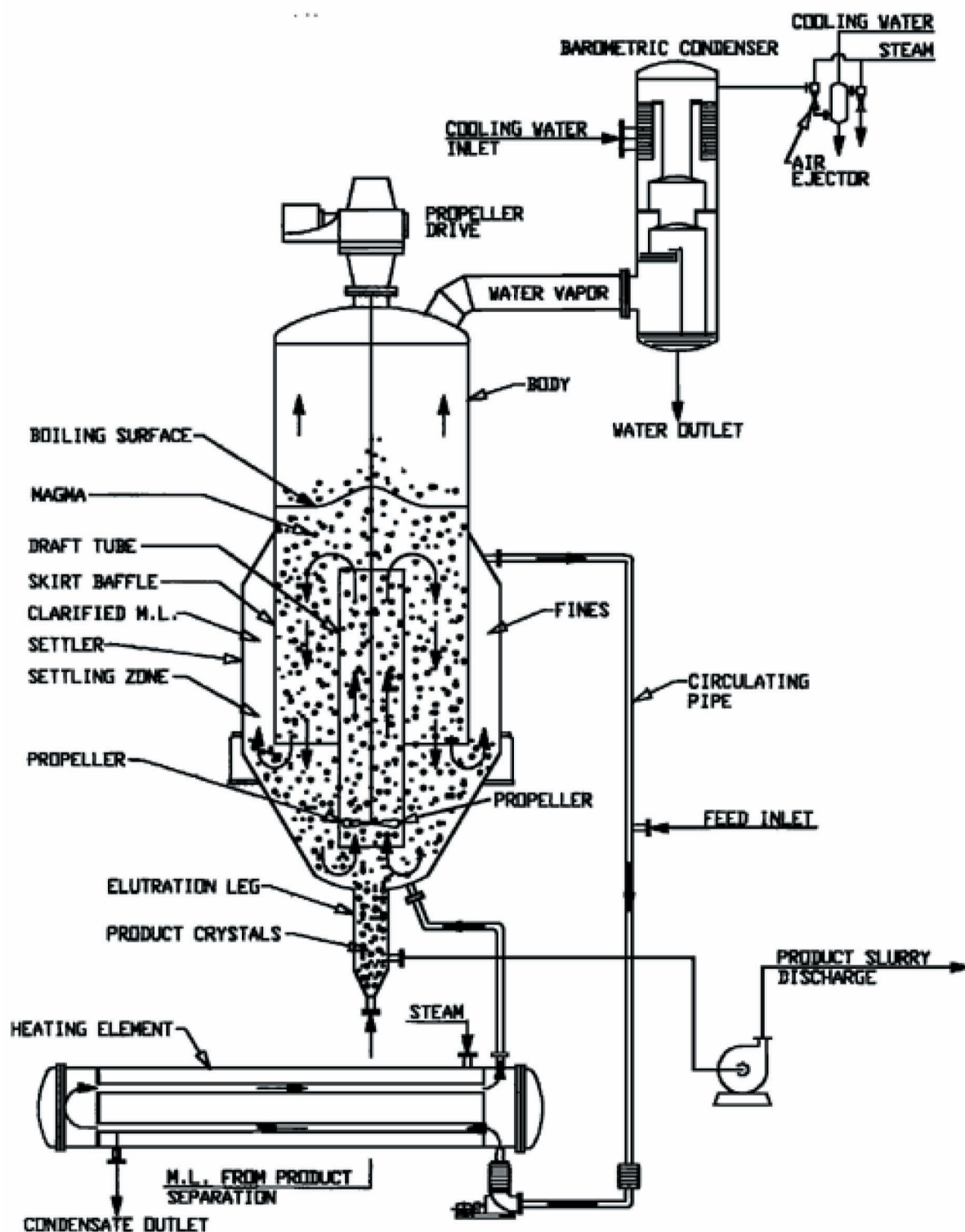
$L_a$  = average particle by weight

$B^o$  = nucleation rate, number/cc-s

**Figure 60.2** Swenson forced-circulation crystallizer. (Source: Courtesy of Swenson Process Equipment Inc.)



**Figure 60.3** Swenson draft-tube baffle crystallizer. (Source: Courtesy of Swenson Process Equipment Inc.)



It is possible to calculate the particle size distribution by weight if the assumptions above are valid and if the plot of  $\ln n$  versus  $L$  is a straight line. The weight fraction up to any size  $L$  is

$$W_x = 1 - e^{-x} \left( 1 + x + \frac{x^2}{2} + \frac{x^3}{6} \right) \quad (60.7)$$



where

$$x = L/GT$$

$W_x$  cumulative weight fraction up to size  $L$

=

In solving Eqs. (60.4) and (60.5), it must be remembered that the growth rate and the nucleation rate must be measured under the same conditions. In evaluating the performance of crystallization equipment, it is necessary to know the heat balance, material balance, and the population balance of the particles being used as seed (when used), as well as the product population balance.

## 60.9 Equipment Designs

---

While many solvent systems are possible, most large-scale industrial crystallizers crystallize solutes from water. Organic solvents are sometimes encountered in the petroleum industry, and alcohol solutions or mixtures of alcohol and water are found in pharmaceutical applications. Typically, water solutions have viscosities in the range of 1–25 cp and boiling point elevations from 1°C up to 12°C. The viscosity of a solution is very important because it determines the settling rates of particles within the solution and heat transfer rates in heat exchange equipment required for heating or cooling the solution. The boiling point elevation represents a temperature loss in an evaporative system where condensation of the vapor in a multiple-stage evaporative crystallizer or condenser is required.

The evaporation rate is determined from the basic process requirements and the heat balance around the system. The evaporation rate and the temperature at which evaporation occurs determine the minimum body diameter. The specific volume of water vapor is strongly influenced by pressure and temperature. Low temperatures, which represent relatively high vacuum for water at its boiling point, require larger bodies than do systems operating at atmospheric pressure. The other consideration in sizing the body is the minimum volume required to provide the retention time required for crystal growth.

Shown in Fig. 60.2 is a forced-circulation evaporator-crystallizer which is often used for the production of sodium chloride, citric acid, sodium sulfate, sodium carbonate, and many other inorganic compounds produced by evaporative crystallization. The body diameter and straight side are determined by the vapor release rate and retention time required to grow crystals of the desired size. The sizes of the circulating pipe, pump, heat exchanger, and recirculation pipe are based on the heat input required to cause the evaporation to take place. Crystals in the solution circulated throughout the body are kept in suspension by the action of the recirculating liquor. Tube velocities, heat transfer rates, and circulation rates are determined by the particular application and the physical properties of the solution. Slurry leaving the crystallizer is pumped by the product discharge pump onto a centrifuge, filter, or other separation equipment. This type of crystallizer is often referred to as an MSMR type, and the crystal size distribution can be described by the mathematical model described in Eqs. (60.4)–(60.7). The crystal size typically produced in equipment of this type is in the range of 30–100 Mesh, and slurry discharge densities



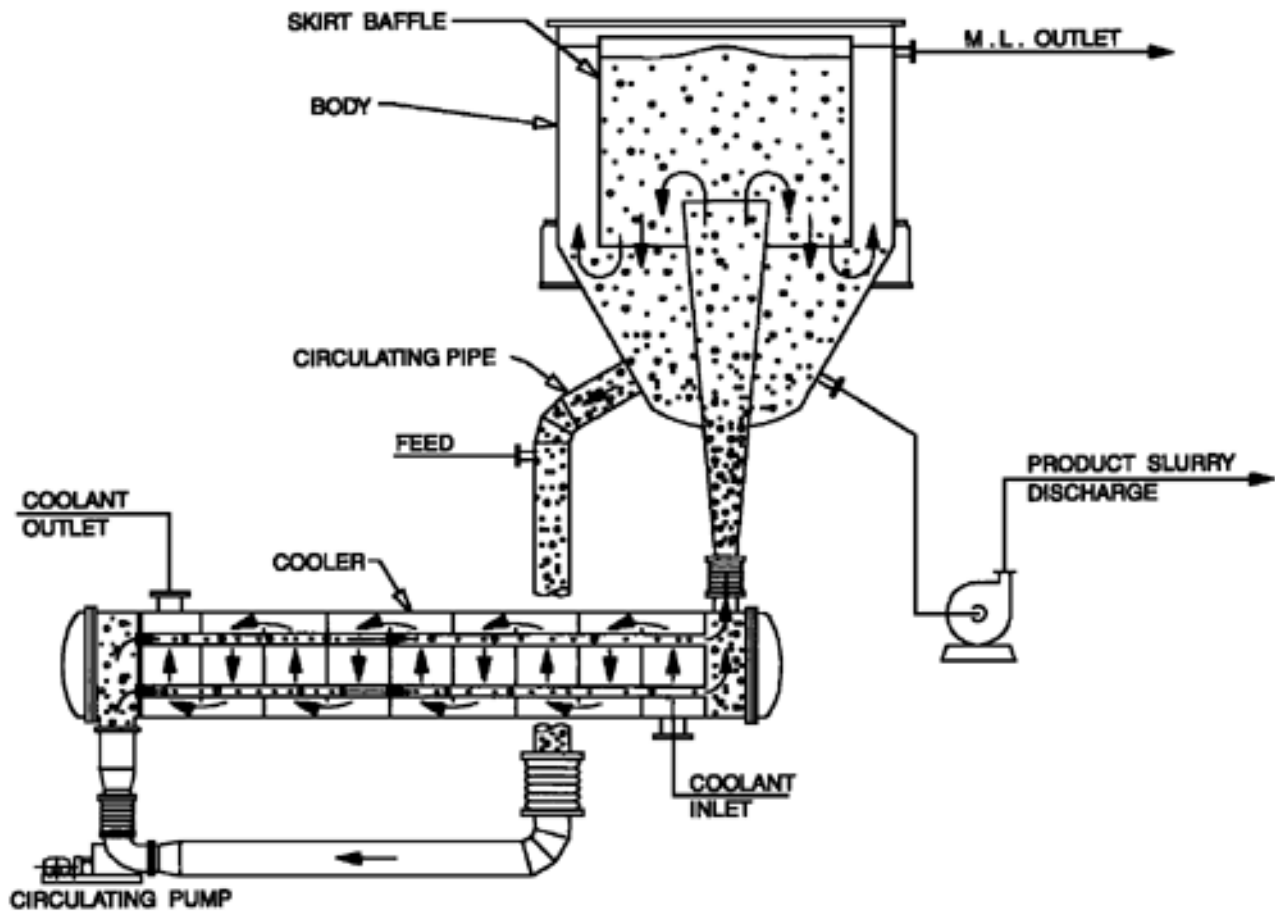
typically handled in such equipment range from about 20–40% by weight solids.

Shown in [Fig. 60.3](#) is a draft-tube baffle (DTB) crystallizer of the evaporative type, including an elutriation leg. Slurry within the crystallizer body is pumped to the surface by means of a slow propeller and recirculates to the suction of the propeller where it is mixed with heated solution exiting the heating element. Surrounding the body of slurry in the crystallizer is an annular space between the skirt baffle and the settler. Liquid is pumped from this annular space at a controlled rate so that small crystal particles from the body can be removed, but the bulk of the circulated liquor and crystals enters the propeller suction. The flow from the annular area is pumped through a circulating pipe by a circulating pump through the heat exchanger, where the temperature rise destroys small particles that are present. This continuous removal and dissolution of small particles by temperature increase serves two purposes: (1) the heat required for the evaporation is transferred into the liquid so that a constant vaporization rate can be maintained; (2) small particles are continuously removed so as to limit the seed crystals in the body to values low enough so that the production can be obtained in a coarse crystal size.

When the crystals become too large to be circulated by the propeller, they settle into the elutriation leg, where they are washed by a countercurrent stream of mother liquor pumped from behind the baffle. Crystals leaving the leg are therefore classified generally at a heavier slurry density than would be true if they were pumped from the body itself. This combination of removal of unwanted fines for destruction and classification of the particle size being discharged from the crystallizer encourages the growth of larger particles than would be obtained in a crystallizer like the forced circulation type in [Fig. 60.2](#). Typically, the DTB crystallizer is used for products in the range of 8–20 Mesh with materials such as ammonium sulfate and potassium chloride.

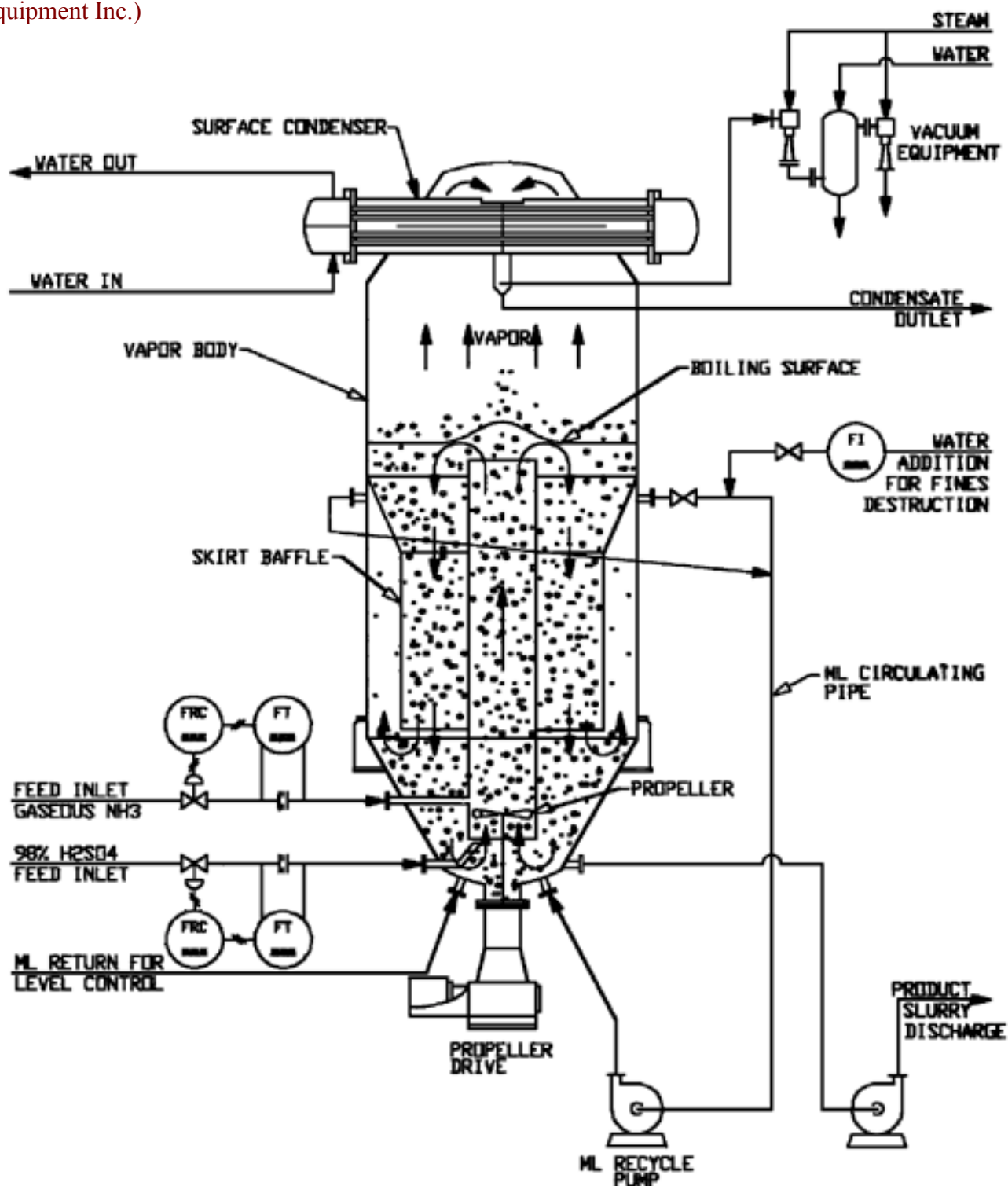
Shown in [Fig. 60.4](#) is a surface-cooled crystallizer which is frequently used at temperatures close to ambient or below. Slurry leaving the body is pumped through a heat exchanger and returns to the body through a vertical inlet. Surrounding the circulating slurry is a baffle that permits removal of unwanted fine crystals or provides for the removal of clarified mother liquor to increase the slurry density within the crystallizer body. Slurry pumped through the tubes of the cooler is chilled by a coolant that is circulated outside the tubes. The temperature difference between the coolant and the slurry flowing through the tubes must be limited to approximately 3–8°C. The temperature drop of the slurry passing through the tubes is normally about 0.5°C. These very low values are required in order to minimize the growth of solids on the tubes. Crystallizers of this type produce a product that ranges between 20 Mesh and 150 Mesh in size. Common applications are for the production of copper sulfate pentahydrate, sodium chlorate, sodium carbonate decahydrate, and sodium sulfate decahydrate.

**Figure 60.4** Swenson surface-cooled crystallizer. (Source: Courtesy of Swenson Process Equipment Inc.)



Shown in Fig. 60.5 is a reaction-type DTB crystallizer. This unit, while in many respects similar to the DTB crystallizer shown in Fig. 60.3, has the important difference that no heat exchanger is required to supply the heat required for evaporation. The heat of reaction of the reactants injected into the crystallizer body supplies this heat. Typically, this type of equipment is used for the production of ammonium sulfate, where sulfuric acid and gaseous ammonia are mixed in the draft tube of the crystallizer so as to produce supersaturation with respect to ammonium sulfate. The heat of reaction is removed by vaporizing water, which can be recirculated to the crystallizer and used for the destruction of fines. Whenever a chemical reaction causes a precipitation of crystalline product, this type of equipment is worth considering because the conditions used in crystallization are compatible with low temperature rises and good heat removal required in reactors. By combining the reactor and crystallizer, there is better control of the particle size with an obvious decrease in equipment costs.

**Figure 60.5** Swenson reaction-type DTB crystallizer. (Source: Courtesy of Swenson Process Equipment Inc.)



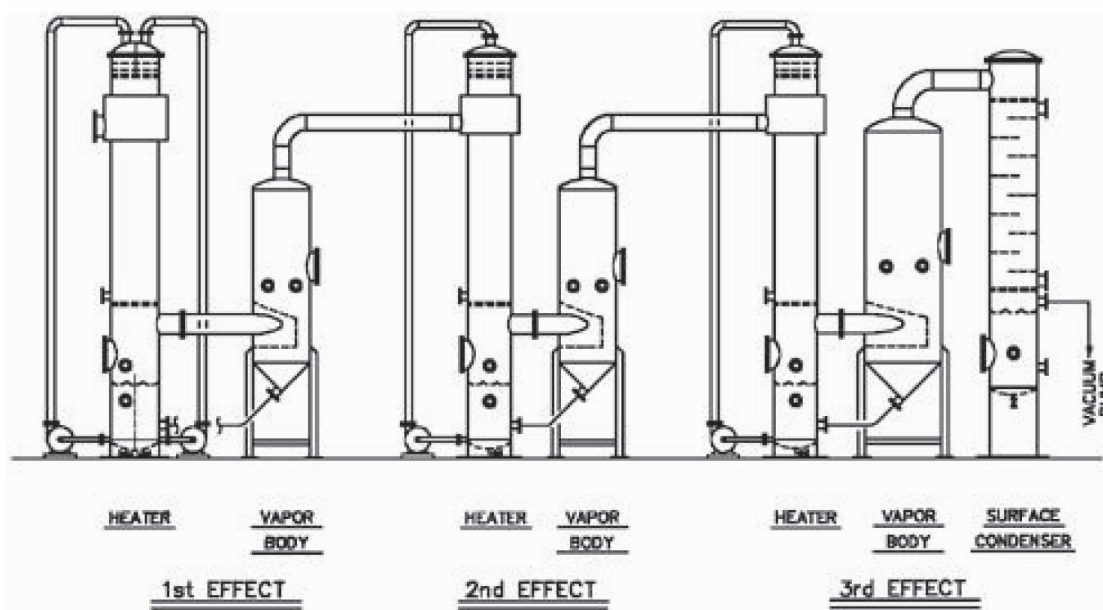
## 60.10 Evaporation

When a solution is boiled (evaporated) at constant pressure, the total pressure above the solution represents the sum of the partial pressures of the liquids that are boiling. If only water is present, then the pressure above the solution at any given temperature corresponds to water at its boiling point at that pressure. If there is more than one component present and that component has a vapor pressure at the temperature of the liquid, then the total pressure represents the vapor pressure of

water plus the vapor pressure of the other component. Vapor leaving such a system, therefore, represents a mixture of solvents in the ratio of their partial pressures. In a sense, an evaporator is a single plate distillation column. In most applications, the vapor pressure of the solute is negligible and only water is removed during boiling which can be condensed in the form of a pure solution. However, when volatile compounds are present (e.g.,  $\text{H}_3\text{BO}_3$ ,  $\text{HNO}_3$ ), some of the volatile material will appear in the overhead vapor.

Since the heat required to vaporize water is approximately 556 cal/kg (1000 Btu/lb), it is important to reduce the amount of energy required as much as possible so as to improve the economics of the process. For this reason, multiple-effect evaporators were developed in the middle of the 19th century and continue today as an important means for achieving good economy during evaporation or crystallization. A multiple-effect falling-film evaporator consisting of three vessels and a condenser is shown in Fig. 60.6. In this type of equipment, the vapor boiled from the first effect (the vessel where the steam enters) is conducted to the heat exchanger of the second effect, where it acts as the heating medium. Vapor boiled in the second effect is conducted to the third effect, where it again acts as the heating medium. Vapor leaving the third effect, in this case, is condensed in a condenser utilizing ambient temperature water. The flow of feed solution to the evaporator can be either forward, backward, or parallel. In a forward feed evaporator, the feed enters the first effect, then passes to the second effect, and is ultimately removed from the third effect as concentrated liquor. With this type of flowsheet, heat exchange means must be employed to minimize the sensible heat required for the liquid fed to the first effect. In a backward feed evaporator, this is not normally done.

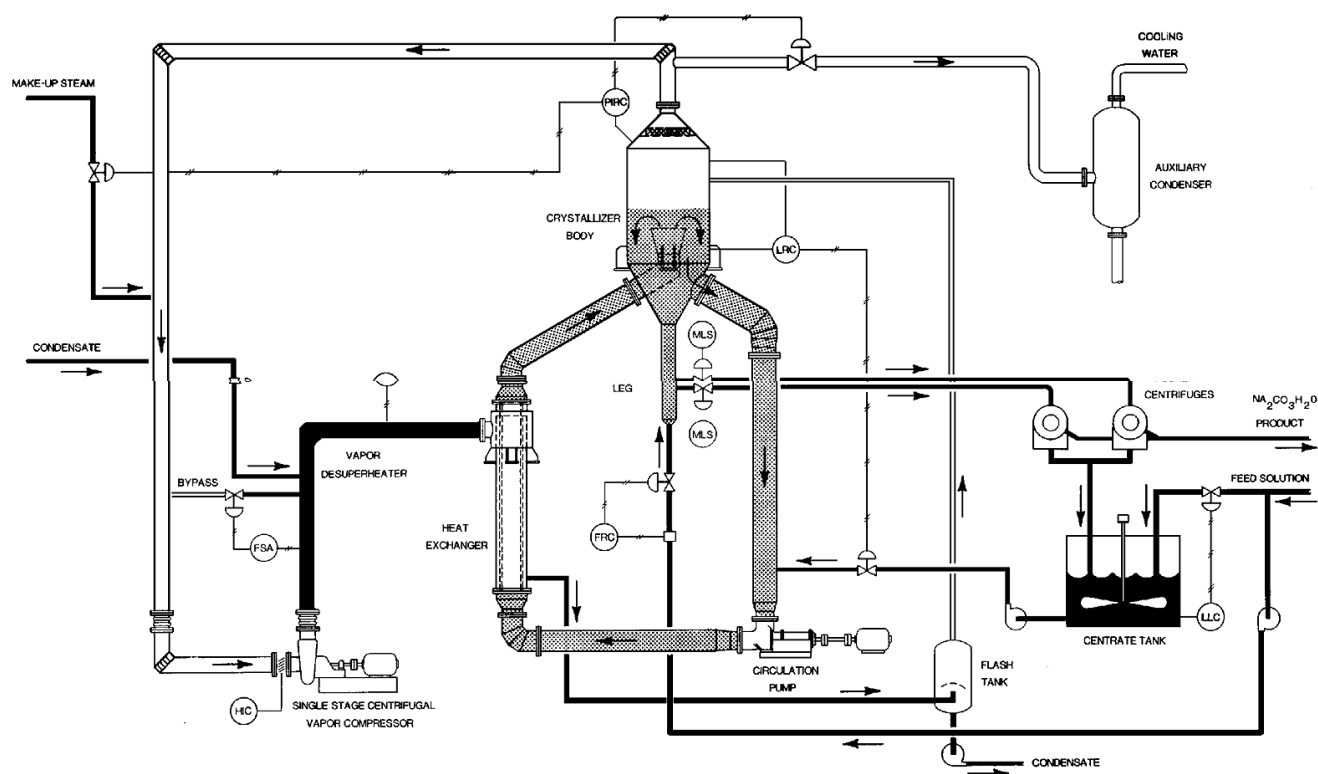
**Figure 60.6** Swenson triple-effect falling-film evaporator. (Source: Courtesy of Swenson Process Equipment Inc.)



An alternative means for reducing energy consumption during evaporation is shown in the recompression evaporative crystallizer in Fig. 60.7. The technique can be employed on both

evaporation and crystallization equipment. In this case, a single vessel is employed and the vapor boiled out of the solvent is compressed by a centrifugal compressor and used as the heating medium in the heat exchanger. The compressed vapor has a higher pressure, and a higher condensing temperature so that there is a change in temperature between the vapor being condensed in the heater and the liquid being heated in the heat exchanger. In utilizing this technique, it must be remembered that the boiling point elevation decreases the pressure of the vapor above the liquid at any given temperature and, thereby, represents a pressure barrier which must be overcome by the compressor. The efficiency of this process varies greatly with the boiling point elevation. As a practical matter, such techniques are limited to those liquids which have boiling point elevations of less than about 13°C. Typically, such compressors are driven at constant speed by an electric motor. The turndown ratio on a constant speed compressor is about 40%. A variable-speed drive would give a greater range of evaporative capacity.

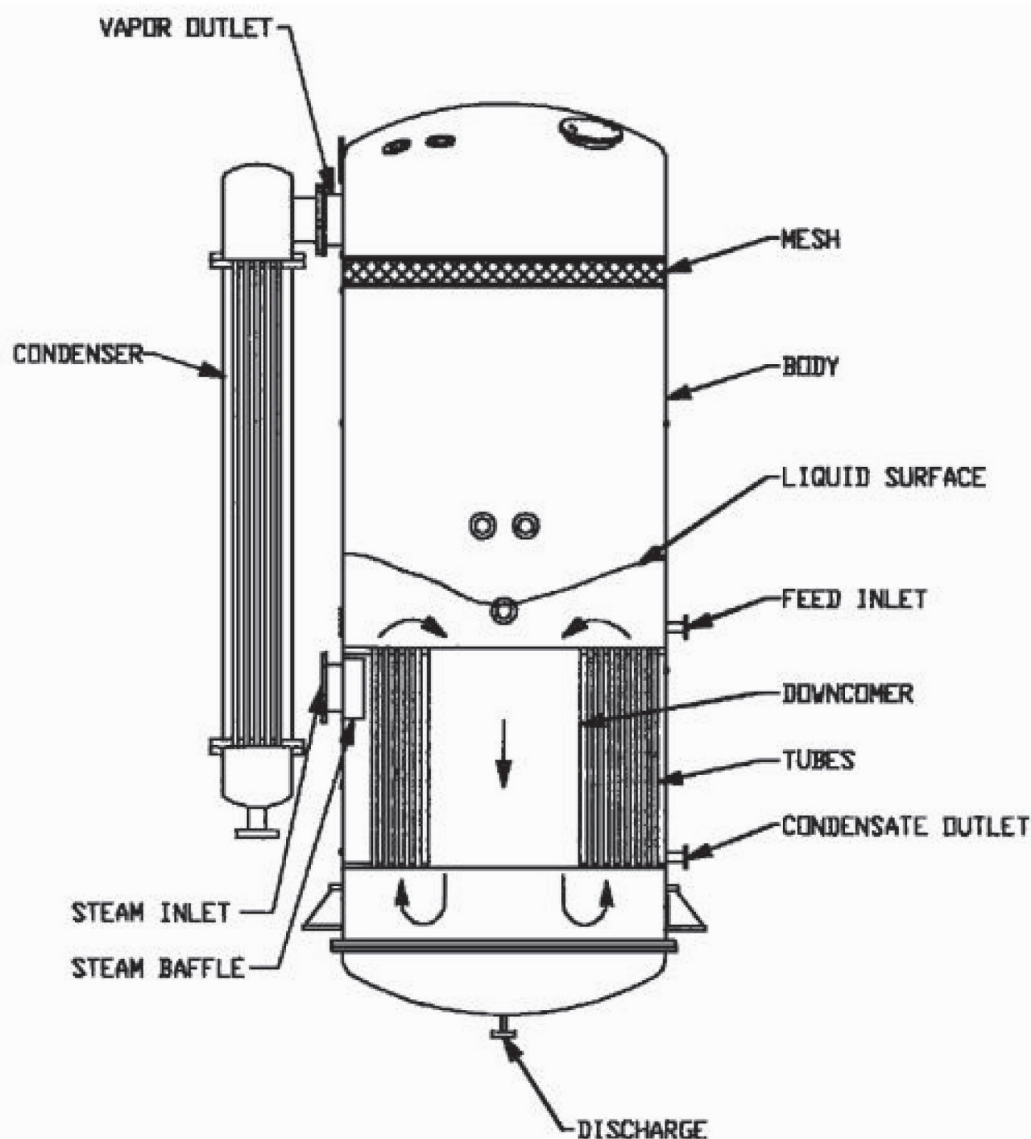
**Figure 60.7** Swenson recompression evaporator-crystallizer. (Source: Courtesy of Swenson Process Equipment Inc.)



During the last 100 years, a wide variety of evaporator types have evolved, each offering advantages for certain specific applications. The forced-circulation crystallizer shown in Fig. 60.2 is utilized for many applications where no crystallization occurs, but the liquids being handled are viscous and the use of the circulation system is needed to promote heat transfer. A

number of evaporator types have been developed that require no external circulating system. For the most part, these rely upon thermo-syphon effects to promote movement of liquid through the tubes as an aid to heat transfer. The calandria evaporator (or Roberts type) shown in Fig. 60.8 is a design that has been widely used since the 19th century for both crystallization and evaporation applications. It relies on natural circulation in relatively short tubes (1–2 m) to maintain heat transfer rates; a relatively large amount of recirculation occurs through the tubes. Since there is no recirculation pump or piping, this type of equipment is relatively simple to operate and requires a minimum of instrumentation. The volume of liquid retained in this vessel is much larger than in some of the rising or falling film designs and, therefore, in dealing with heat sensitive materials where concentration must proceed at relatively short retention times, the calandria would be a poor choice. In many situations, however, especially where some crystallization may occur, this evaporator may be operated successfully in a semi-batch or continuous manner.

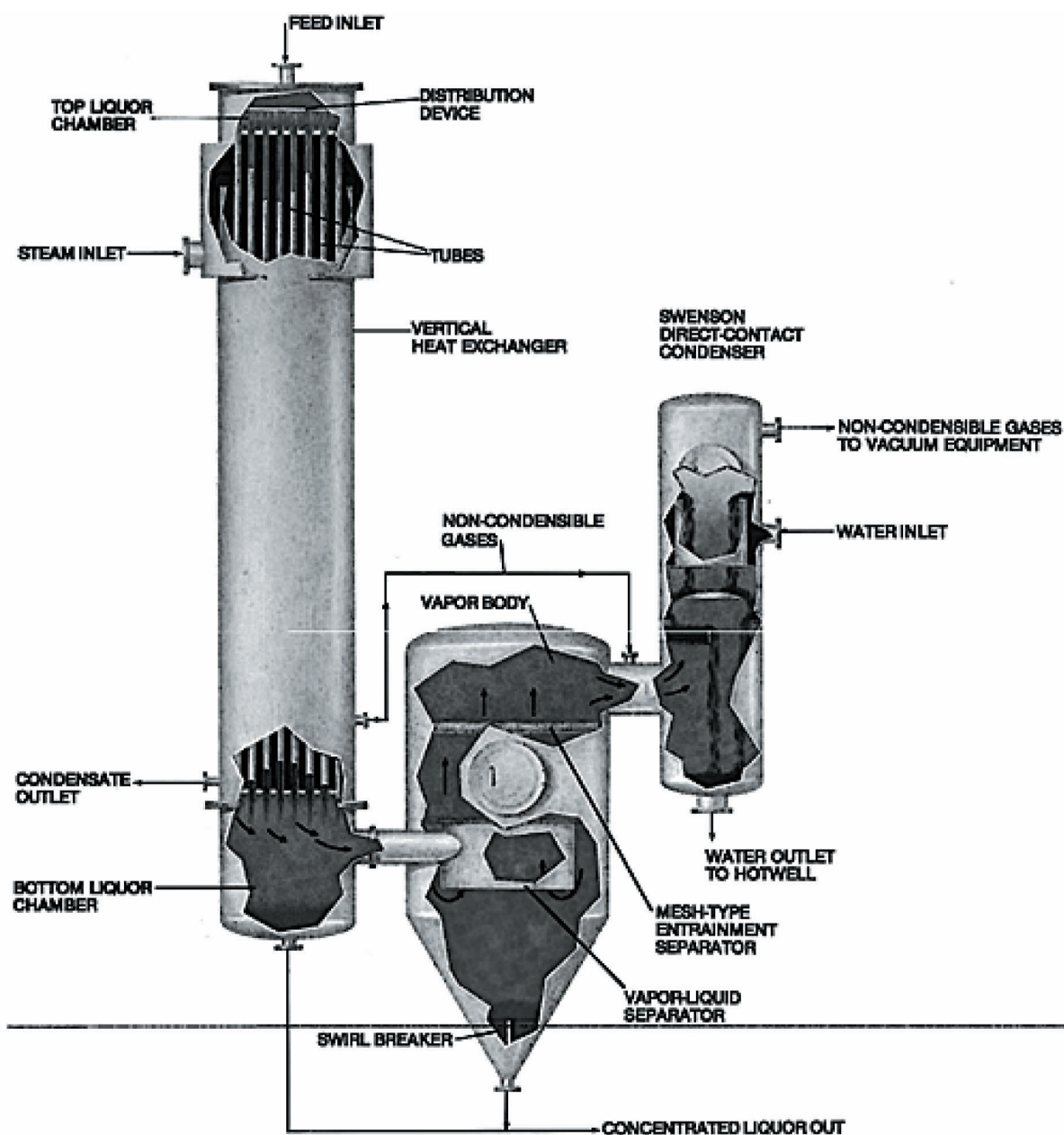
**Figure 60.8** Swenson calandria evaporator. (Source: Courtesy of Swenson Process Equipment Inc.)





The falling-film evaporator shown in Fig. 60.9 is similar to the rising-film evaporator, except that there must be sufficient liquid at all times entering the heater at the feed inlet to wet the inside surface of the tubes in the heat exchanger bundle. With insufficient circulation, solute material can dry on the tubes and cause a serious reduction in heat transfer. Many falling-film evaporators operate with a recirculating pump between the concentrated liquor outlet and the feed inlet to be certain that the recirculation rate is adequate to maintain a film on the tubes at all times. If this is done, the system can operate stably through a wide range of capacities and achieve very high rates of heat transfer, often 50–100% more than are obtained in a rising-film evaporator. The other advantage of the falling film evaporator is that it can operate with very low temperature differences between the steam and the liquid since there is no hydrostatic pressure drop of consequence within the tubes to prevent boiling at the inlet end of the heat exchanger. As a result, this type of design has found wide application as a recompression evaporator.

**Figure 60.9** Swenson falling-film evaporator. (Source: Courtesy of Swenson Process Equipment Inc.)



Even though evaporators are typically used where no precipitation of solids occurs, there is often a trace of precipitation in the form of scaling components which coat the inside of the tubes over a relatively long period of time. This scaling is analogous to that which occurs in boilers and many other types of heat transfer equipment. Typically it is due to either a small amount of precipitation or, because of the composition of the materials being concentrated, some inverted solubility components. Such scaling may often be reduced by a technique known as "sludge recirculation." This is commonly done in cooling tower blowdown evaporation and in the evaporation of salt brines where scaling components are present. In these cases, the evaporator flowsheet is designed in such a way that a thickened slurry of the scaling component can be recirculated from the discharge of the evaporator back to the feed side. By maintaining an artificial slurry density of the scaling component, which is higher than the natural slurry density, it is often possible to reduce the growth of scale which occurs on heat transfer surfaces.

## Defining Terms

**Crystal:** A solid bounded by plane surfaces which has an internal order with atoms or molecules in a fixed lattice arrangement.

**Crystallizer:** An apparatus for causing the crystallization of solutes from solvents by means of changes in heat or solvent inventory.

**Evaporator:** An apparatus for causing water or other solvents to be removed from a solution in order to increase the concentration of the solution.

**Nucleation:** The birth of a new crystal within a supersaturated solution.

**Recompression:** A process for collecting the vapor boiled from the solution in an evaporator or crystallizer and compressing it to a higher pressure where it can be used as the heating media for said evaporator or crystallizer.

**Supersaturation:** A metastable condition in a solution which permits nucleation and growth of crystals to occur.

## References

Campbell, A. N. and Smith, N. O. 1951. *Phase Rule*, 9th ed. Dover, Mineola, NY.

Linke, W. F. (Ed.) 1958. *Solubilities, Inorganic and Metal-Organic Compounds*. Van Nostrand, New York.

Myerson, A. (Ed.) 1993. *Handbook of Industrial Crystallization*, p. 104. Butterworth, Boston.

Randolph, A. D. and Larson, M. A. 1988. *Theory of Particulate Processes*, 2nd ed., p. 84. Academic Press, Boston.

Washburn, E. W. (Ed.) 1926. *International Critical Tables*, McGraw-Hill, New York.



Woods, D. G., Walker, D. R. B., Koros, W. J. "Membrane Separation"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Membrane Separation

---

- 61.1 Dialysis
- 61.2 Reverse Osmosis
- 61.3 Gas and Vapor Separations
- 61.4 Asymmetric Membranes
- 61.5 Membrane Stability and Fouling
- 61.6 Module Design Considerations

**David G. Woods**

*University of Texas, Austin*

**David R. B. Walker**

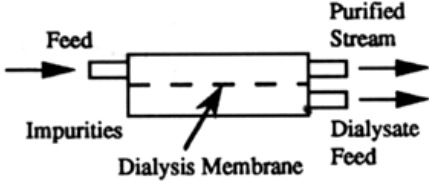
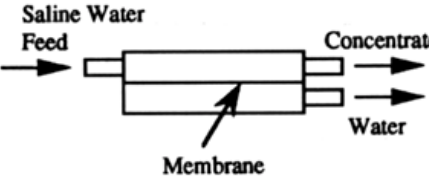
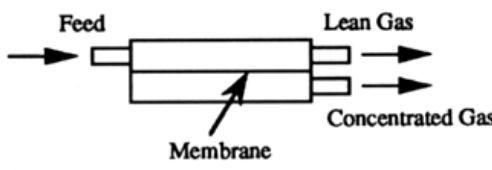
*University of Texas, Austin*

**William J. Koros**

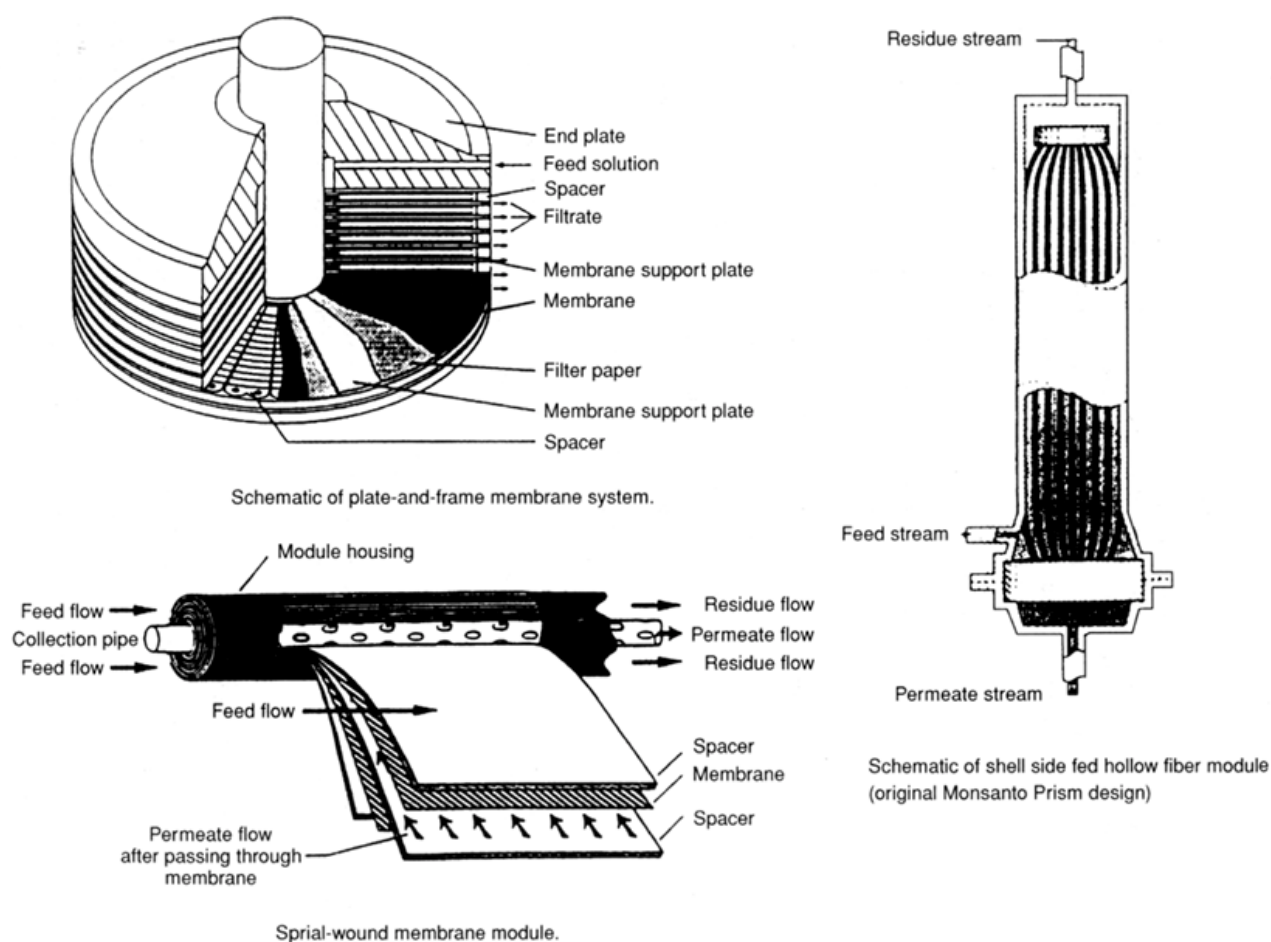
*University of Texas, Austin*

The separations referred to as *ultrafiltration* and *microfiltration* rely upon hydrodynamic sieving of a larger solute or particle from a fluid suspending medium. These types of separations are covered separately in other chapters under their respective headings. Contrary to these hydrodynamic sieving devices, dialysis, reverse osmosis, and gas separation membranes operate at a molecular scale to achieve selective passage of one or more components in a feed stream. [Figure 61.1](#) illustrates idealized flow schemes of these processes and the materials that they separate. Penetrant partitioning into the membrane and thermally activated diffusion within the membrane cooperate to yield a solution-diffusion process referred to as *permeation*. The two primary measures of the membrane performance are the membrane permeability to the desired penetrant and the membrane selectivity between penetrants. Membranes are typically "packaged" in three basic configurations: flat sheet, tubular, and hollow fiber ([Fig. 61.2](#)).

**Figure 61.1** Idealized flow schemes and materials separated.

PROCESS	CONCEPT	MATERIALS PASSED	DRIVING FORCE	MATERIAL RETAINED
Dialysis		Ions and Low-Molecular-Weight Organics (Urea, etc.)	Concentration Difference	Dissolved and Suspended Material with Molecular Weight >1000
Reverse Osmosis		Water	Pressure Difference, Typically 100-800 psi	Virtually All Suspended and Dissolved Material
Gas Separation		Gases and Vapors	Pressure Difference 1-100 atm	Membrane-Impermeable Gases and Vapors

**Figure 61.2** Configurations of membrane modules. (Source: Koros, W. J. and Fleming, G. K. 1993. Membrane-based separation. *J Membrane Sci.* 83(1):1–80. With permission.)



## 61.1 Dialysis

Dialysis processes are most effective for systems having large concentrations of solutes that differ significantly in molecular size from the solvent. However, the concentration-driven nature of dialysis allows its application in systems where the species to be separated are sensitive to high shear rates or high pressures. The most prominent use of dialysis membranes is in the treatment of end-stage renal disease. In this process, low and intermediate molecular weight waste products are removed from the blood of a patient. To a much smaller extent, dialysis is applied in microbiology for enzyme recovery from cultures and in the food industry for desalting cheese whey solids.

The principal membrane material for dialysis is regenerated cellulose. The water-swollen material is a hydrogel and will undergo irreversible, destructive collapse upon drying in the absence of **plasticizer**. Stability upon air drying can be achieved by using hydrophobic **glassy polymers** such as methacrylates, polysulfones, and polycarbonates. Unfortunately, the hydrophobic natures of these materials make them less blood compatible.

Although dialysis modules are available in the three basic configurations shown in Fig. 61.2, recent design emphasis has been on the development of hollow fiber modules because of their

high surface areas. A typical hollow fiber hemodialyzer utilizes laminar flow on the feed side with a pressure drop less than 0.1 MPa across the membrane [Klein *et al.*, 1987].

A **semipermeable** membrane is used to separate components based on their different mobilities through the membrane. The membrane separates two flowing streams, a solute-containing feed stream and a dialysate stream that may not contain solute. Solute diffuses from the feed side to the dialysate side due to the difference in chemical potential between the two membrane-liquid interfaces. When hydraulic pressure is added to provide additional convective transport, it is properly referred to as *pressure-assisted* dialysis. Under these conditions the process assumes aspects of ultrafiltration. Convective transport of the solvent can also occur due to osmotic effects if either of the streams is highly concentrated in the solute.

Since diffusion coefficients are weakly dependent on molecular size in highly swollen gels, dialysis is able to separate efficiently only species that differ significantly in molecular size. The chemical potential driving force arises from the concentration gradient across the membrane, so dialysis is not favorable for the removal of very low concentrations of solute.

As dialysis of a solute occurs, the solute is removed from the feed side solution. Solute is transported across the membrane from an interfacial layer between the membrane and the bulk solution. The concentration of solute must be restored by diffusion into this layer from the bulk solution. If the flow is laminar—as is common in small-diameter hollow fibers and spiral-wound and plate-and-frame modules—convective replenishment from the bulk far away from the membrane does not occur, and diffusion from the bulk must be relied upon to replenish the permeating solute at the membrane surface. A concentration gradient therefore exists between the bulk solution and the membrane-solution interface.

After the solute has crossed the membrane, it must be removed from the interfacial layer on the **dialysate** side to maintain the driving force of concentration difference across the membrane. In this interfacial layer a concentration gradient also exists between the membrane-dialysate interface and the bulk dialysate. Mass transfer from the membrane to the dialysate is dependent on the transmembrane flux and the bulk dialysate concentration.

The overall resistance to mass transfer from the bulk feed solution to the bulk dialysate solution can be associated with the sum of three terms: the resistance of the fluid boundary layer on the feed side, the resistance to diffusion through the membrane itself, and the resistance of the fluid boundary layer on the dialysate side. These terms are combined to give a total mass transfer coefficient,  $k_T$ , and, since resistances to flow are linearly additive, one finds:

$$\frac{1}{k_T} = \frac{1}{k_F} + \frac{1}{P_M} + \frac{1}{k_D} \quad (61.1)$$

where  $k_F$  and  $k_D$  are the mass transfer coefficients (cm/s) for the feed and dialysate concentration polarization layers respectively, and  $P_M$  (cm/s) is the **permeability** of the membrane. Both  $k_F$  and  $k_D$  are often considered the ratio of the solute diffusion coefficient in the boundary layer to the boundary layer thickness. As solute size increases, the **diffusivity** of the solute through the membrane decreases more rapidly than in the solution boundary layers; therefore, membrane resistance becomes the dominant term. With small solute molecules, however, membrane-limited dialysis will occur only when  $k_F$  and/or  $k_D$  are large relative to  $P_M$ . Under these conditions,

transport across the membrane can become limiting, and a thinner membrane will not increase the transport rate under comparable feed and dialysate flow conditions.

The simplest case of dialysis is one of countercurrent flow between feed and dialysate, with negligible transmembrane pressure or convection, and equal inlet and outlet flow rates for both the feed and dialysate streams. The overall mass balance is

$$\dot{m} = Q_F(C_{Fi} - C_{Fo}) = Q_D(C_{Do} - C_{Di}) \quad (61.2)$$

where  $\dot{m}$  is the mass flow rate of the solute across the membrane from feed to dialysate streams (grams/second),  $Q$  is the volumetric flow rate (cm<sup>3</sup>/s),  $F$  and  $D$  refer to feed and dialysate solutions, respectively, and  $C_i$  and  $C_o$  are the solute concentrations in and out of the dialyzer, respectively (grams/cm<sup>3</sup>). In terms of an overall mass transfer coefficient,  $k_T$ , the mass flow rate of solute is

$$\dot{m} = k_T A \frac{(C_{Fo} - C_{Di}) - (C_{Fi} - C_{Do})}{\ln[(C_{Fo} - C_{Di})/(C_{Fi} - C_{Do})]} \quad (61.3)$$

where  $A$  is the membrane area (cm<sup>2</sup>). Performance of the dialyzer can be expressed as a dialysance,  $D$ , which is defined as the ratio of the mass flow rate and the inlet concentration differences as follows (cm<sup>3</sup>/s):

$$D = \frac{\dot{m}}{C_{Fi} - C_{Di}} \quad (61.4)$$

Similar expressions for other flow configurations can be derived. With convective transport and ultrafiltration of the feed, more complex expressions can be derived to accurately describe the dialysis process.

## 61.2 Reverse Osmosis

---

Reverse osmosis (RO) membranes should display high water flux, high salt retention, hydrolytic stability, and chemical stability. The drive to improve membranes in these areas has led to the development of many varieties of materials in the past decade to replace the historically used asymmetric cellulose acetate membrane. Among the successes have been polyaramides and polyamides. The latter have been successfully manufactured as composites using an ultrafiltration membrane as the porous support. This design provides good stability and extremely high flux.

Reverse osmosis membranes are most frequently used in the desalination of brackish water and sea water. However, the membranes also find applications in the paper industry for the concentration and partial fractionation of liginosulfonates in wood pulping and in the food industry for the concentration of whey, milk, and maple syrup [Klein *et al.*, 1987]. More generally, reverse osmosis systems can be applied in a variety of processes to concentrate effluent streams; however, their application is limited by the need to overcome high osmotic pressures at high concentrations, as discussed in the following paragraphs.

Reverse osmosis modules are commonly found in all three types of configurations shown in Fig. 61.2. In an RO plant the feed is filtered and, if needed, biologically stabilized and then pumped through the membrane module(s). The pressure must be carefully balanced so that the end pressure is high enough to overcome the osmotic pressure, but not so high as to overstress the membrane, causing failure. This requirement sometimes dictates the use of booster pumps in the later stages of the process. The typical required driving pressure (operating pressure – osmotic pressure) of 1 to 5 MPa requires a high-pressure pump, which is the dominant energy consumer in the process. In large plants some energy recovery is practiced, but is not usually economically viable; a discussion of the economics of reverse osmosis is given by Koros [1988].

Reverse osmosis is a solution-diffusion membrane separation process that relies upon a pressure to make the chemical potential of the solvent overcome the osmotic pressure of the solution to be purified. At low solute concentrations, the osmotic pressure  $\pi$  (kPa) of the solution at temperature  $T$  (K) with a solute molar concentration  $C_s$  (moles/cm<sup>3</sup>) is given by

$$\pi = C_s RT \quad (61.5)$$

where  $R$  is the ideal gas constant. The concentration term must include dissociation effects of the solute since  $\pi$  is a colligative property of the solution. Some solute passes through the membrane in response to the concentration gradient that develops. Fixed charges within the membrane can help suppress solution of ionic solutes by *Donnan exclusion* and minimize this undesirable solute flux from the feed to the permeate side.

The molar flux of desired solvent (water) through a reverse osmosis membrane  $N$  (moles/cm<sup>2</sup>·s) is proportional to the difference between the applied pressure difference and the osmotic pressure difference across the membrane, as shown by Eq. (61.6),

$$N_w = \frac{C_w D_w v_w}{RT} \left( \frac{\Delta p - \Delta \pi}{l} \right) \quad (61.6)$$

where  $C_w$  is the concentration of the water in the membrane (moles/cm<sup>3</sup>),  $D_w$  is the diffusivity of the water in the membrane (cm<sup>2</sup>/s), and  $v_w$  is the partial molar volume of the water (cm<sup>3</sup>/mole). The solute flux is given as

$$N_s = \frac{D_s S_s \Delta C_s}{l} \quad (61.7)$$

where  $S_s$  is a dimensionless partition coefficient for the solute in the membrane material. From Eqs. (61.6) and (61.7) it is apparent that the water flux increases as the pressure applied to the feed increases, whereas the solute flux does not increase. The rejection of the solute may be 99% or higher, since water flux can be increased without causing much change in the concentration driving force for the dissolved salts.

## 61.3 Gas and Vapor Separations

---

Applications of membrane-based gas separation technology tend to fall into four major categories:

1. Hydrogen separation from a wide variety of slower-permeating **supercritical** components such as CO, CH<sub>4</sub>, and N<sub>2</sub>
2. Acid gas (CO<sub>2</sub> and H<sub>2</sub>S) separation from natural gas
3. Oxygen or nitrogen enrichment of air.
4. Vapor/gas separation, including gas drying and organic vapor recovery.

The order of the various types of gas-gas applications given in this list provides a qualitative ranking of the relative ease of performing these three types of separations. The fourth application, involving the removal of vapors from fixed gases, is generally carried out with the use of a rubbery polymer, whereas the first three applications rely upon glassy materials.

In an idealized sense, gas separation membranes act as molecular-scale filters separating a feed mixture of A and B into a **permeate** of pure A and **nonpermeate** or **retentate** of pure B. Figure 61.3 shows that gas separations by membranes can be performed using three types of transport mechanisms: Knudsen diffusion, solution diffusion, and molecular sieving. High-performance membranes based on solution-diffusion separations can achieve separation factors of greater than 100 in many cases. Transport through the membrane occurs due to the thermally activated motion of polymer chain segments that create penetrant-scale transient gaps in the matrix, allowing diffusion from the upstream to the downstream side of the membrane to occur.

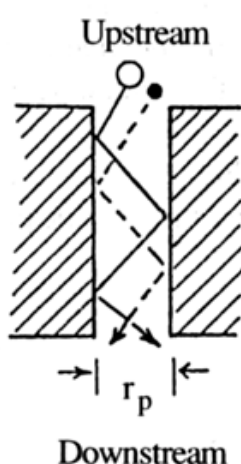
The driving force for gas separation is the transmembrane partial pressure typically achieved by compression of the feed gas. The permeability of the polymer for component  $i$ ,  $P_i$ , is defined as

$$P_i = \frac{N_i}{(\Delta p_i/l)} \quad (61.8)$$

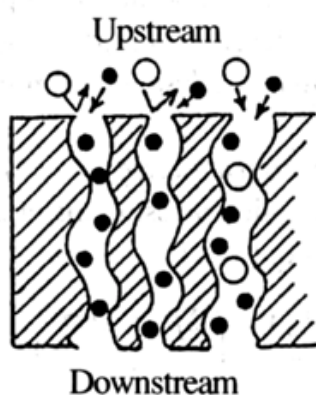
where  $N_i$  is the gas flux [cm<sup>3</sup> (STP)/cm<sup>2</sup> · s],  $\Delta p_i$  is the partial pressure difference between the upstream and downstream faces of the membrane respectively (cm Hg), and  $l$  is the thickness of the membrane (cm). Polymer permeabilities are often expressed using a unit defined as a barrer:



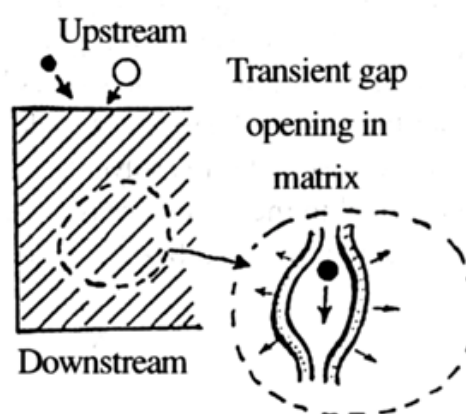
**Figure 61.3** Transport mechanisms for gas separations.



(a) Knudsen flow-separation is based on the inverse square root ratio of the molecular weights of A & B.



(b) Ultramicroporous molecular sieving separation is based **primarily** on the much higher diffusion rates of the smallest molecule, but sorption level differences may be important factors for similarly sized penetrants like  $O_2$  &  $N_2$ .



(c) Solution-Diffusion-separation is based on **both** solubility and mobility factors in essentially all cases. Diffusivity selectivity favors the **smallest** molecule. Solubility selectivity favors the most **condensible** molecule.

$$1 \text{ barrer} = \frac{10^{-10} \text{ cm}^3(\text{STP}) \cdot \text{cm}}{\text{cm}^2 \cdot \text{cm Hg} \cdot \text{s}} = \frac{10^{-10} \text{ cm}^3(\text{STP})}{\text{cm} \cdot \text{cm Hg} \cdot \text{s}} \quad (61.9)$$

The permeation of penetrants occurs due to a solution-diffusion mechanism, so the permeability coefficient is expressed as a product of two terms:

$$P_i = D_i S_i \quad (61.10)$$

where  $D_i$  is the average diffusion coefficient of the penetrant across the membrane ( $\text{cm}^2/\text{s}$ ) and  $S_i$  is the effective solubility coefficient of the penetrant in the membrane [ $\text{cm}^3(\text{STP})/\text{cm}^3 \cdot \text{cm Hg}$ ]. The solubility coefficient is defined as

$$S_i = C_i/p_i \quad (66.11)$$

where  $C_i$  is the concentration of the penetrant  $i$  in the polymer [ $\text{cm}^3(\text{STP})/\text{cm}^3$ ] at equilibrium when the external partial pressure is  $p_i$  (cm Hg).

The ideal separation factor,  $\alpha_{A/B}^*$ , is based on the individual permeabilities of two gases A and B and is given as

$$\alpha_{A/B}^* = \frac{P_A}{P_B} \quad (61.12)$$

The ideal separation factor provides a measure of the intrinsic **permselectivity** of a membrane material for mixtures of A and B. In the absence of strong polymer-penetrant interactions,  $\alpha_{A/B}^*$  in mixed feed situations can be approximated to within approximately 10 to 15% using the more easily measured ratio of permeabilities of pure components A and B. If Eq. (61.10) is substituted into Eq. (61.12), the ideal separation factor can be separated into two parts;

$$\alpha_{A/B}^* = \left[ \frac{D_A}{D_B} \right] \left[ \frac{S_A}{S_B} \right] \quad (61.13)$$

where  $D_A/D_B$  is the diffusivity selectivity and  $S_A/S_B$  is the solubility selectivity. The diffusivity selectivity is determined by the ability of the polymer to discriminate between the penetrants based on their sizes and shapes and is governed by intrasegmental motions and intersegmental packing. The solubility selectivity, like the solubility, is thermodynamic in nature.

The actual separation factor between two components A and B is defined in terms of the upstream ( $X_{A2}$  and  $X_{B2}$ ) and downstream ( $X_{A1}$  and  $X_{B1}$ ) mole fractions.

$$\text{SF} = (X_{A1}/X_{B1})/(X_{A2}/X_{B2}) \quad (61.14)$$

Equation (61.14) can be written in terms of  $D$  and  $S$  but also includes partial pressure terms.

$$SF = \left[ \frac{D_A}{D_B} \right] \left[ \frac{S_A}{S_B} \right] \left[ \frac{\Delta p_A/p_{A2}}{\Delta p_B/p_{B2}} \right] \quad (61.15)$$

From Eq. (61.15) it is apparent that the actual separation factor may be less than the ideal selectivity. Also apparent is that when the downstream pressure is nearly zero Eq. (61.15) simplifies to Eq. (61.12).

So far in the discussion of membrane productivity and selectivity, no attention has been given to the gas that is rejected by the membrane. During actual usage of a membrane module, at least three streams are involved: the feed, the permeate, and the retentate. A sweep gas may also be used on the permeate side in special cases to reduce stagnation at the downstream face of the membrane for vapor removal. In a real membrane application the residue and permeate pressures will be nonzero, and the driving force across the length of the membrane will therefore vary. Under realistic conditions even more complex analysis is needed to account for the variations of driving force terms used in Eq. (61.15) [Koros and Fleming, 1993].

The extraordinarily small molecular size of H<sub>2</sub> makes it highly permeable through membranes and easily collected as a permeate product compared to many larger-sized gases such as N<sub>2</sub>, CH<sub>4</sub>, and CO. Perhaps surprisingly, the separation of H<sub>2</sub> from CO<sub>2</sub> and H<sub>2</sub>S is difficult, although these latter gases are clearly much larger in molecular size than H<sub>2</sub>. This phenomenon can be understood from consideration of Eq. (61.13). The somewhat higher solubilities of gases such as N<sub>2</sub>, CH<sub>4</sub>, and CO compared to H<sub>2</sub> do not offset the much smaller diffusivities of these bulky gases, so H<sub>2</sub> permeates more rapidly. However, the highly soluble, slower-diffusing compound CO<sub>2</sub> has a permeability similar to that of H<sub>2</sub> due to a compromise between these two factors, which illustrates the importance of understanding the two-part nature of the permeability coefficient in Eq. (61.10).

For the second type of separations noted, the high permeabilities of CO<sub>2</sub>, H<sub>2</sub>S, and H<sub>2</sub>O can be used to an advantage. The relatively high solubilities of these compounds in membranes at low partial pressures, coupled with the low diffusivity and **solubility** of the bulky supercritical methane molecule, give high membrane productivity for these penetrants with good selectivity over CH<sub>4</sub>. For example, with an equimolar upstream feed, commercial materials such as polysulfone and cellulose acetate yield permeation rates of CO<sub>2</sub> that are roughly 20 to 30 times higher than for CH<sub>4</sub>. These results can be improved with careful materials selection.

As noted earlier, the most difficult of the three types of gas-gas separations given earlier involves O<sub>2</sub>/N<sub>2</sub> processing. Unfortunately, currently available polymer membranes have only moderate selectivities for separation of oxygen and nitrogen. This again can be understood by considering Eq. (61.13). The size and shape (and hence diffusivity) of O<sub>2</sub> and N<sub>2</sub> are quite similar; moreover, the solubilities of the pair in most membranes are very similar. This leads to similar permeabilities for the two components and thus modest selectivities, rendering high-purity separation difficult. Nevertheless, well-engineered commercial systems do exist for producing nitrogen- enriched air above 98% purity.

The separation of vapors from gases is, in principle, extremely easy since the solubility of highly condensable vapor components makes their solubility selectivity high (and overall

selectivity is also high, since mobility selectivity is close to unity often) in rubbery polymers, which have liquid-like characteristics. Silicone rubber is a common choice for such applications since it offers high transport rates. This type of separation requires great care to minimize partial pressure buildup of the vapor in the permeate.

## 61.4 Asymmetric Membranes

---

Dense polymer films, though providing good separation of many process streams, generally yield fluxes too low to be economical in most membrane separation applications. Most membranes have "asymmetric" structures as shown in Fig. 61.4. Integrally skinned or "simple" membrane structures are typified by those constructed from cellulose acetate and aromatic polyamides in which a single membrane polymer is used for the support and thin selective layer. Polysulfone is easily formed into porous structures but is difficult to prepare with a truly pore-free skin. This obstacle is generally eliminated using a "caulking" procedure developed by Henis and Tripodi of Monsanto [1980]. An interesting extension of these approaches involves the generation of a "composite" membrane comprising a more-or-less unskinned porous support with subsequent attachment of a second material with desirable solution-diffusion separation properties. Asymmetric and composite structures permit the preparation of nearly pore-free membranes with skin thicknesses of roughly 1000 angstroms, thus increasing flux values by more than a factor of 200 over most dense films. This represents the latest breakthrough in the drive to increase membrane productivity.

In all of these membrane types the primary separation occurs in the thin skin on these membranes, whereas the underlying open-celled structure acts simply as a support and cushion for the skin. If the resistance to transport across the porous support layer exceeds roughly 10% of the resistance of the dense selective layer, serious loss in efficacy results. This effect places a premium on the ability to generate low-resistance open-celled porous supports.

## 61.5 Membrane Stability and Fouling

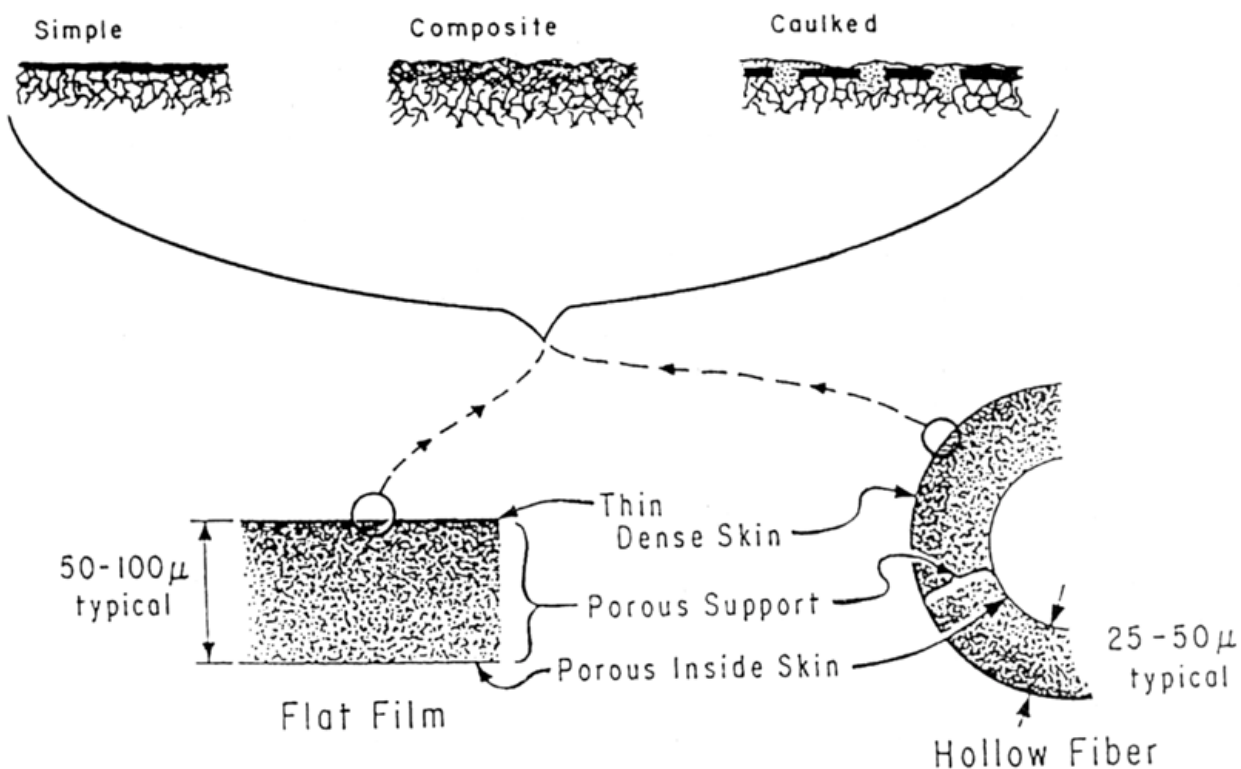
---

In design of a membrane process, the operating environment must be carefully considered. Although polymer materials today tend to be quite robust, they can be affected adversely by their environment in some cases.

In reverse osmosis applications the chemical stability of the membrane and the polymer itself in an aqueous environment is an obvious demand. Another consideration is the stability of the porous structure, which must be maintained in a wet and swollen state to prevent collapse. In dialysis applications where the selective skin itself is somewhat porous, wetting of the membrane and maintenance of the porosity are of primary concern. Fouling of membranes in both applications can be a problem due to the nature of the streams being separated, which often contain high molecular weight solutes and possibly particulate matter.

With gas separations the effects of high levels of sorbed penetrant can be very important in applications such as acid gas removal from natural gas. **Transport plasticization** typically refers to a situation in which the diffusivity of a penetrant increases significantly due to the presence of

**Figure 61.4** Asymmetric membrane structure. (Source: Koros, W. J. 1998. Membranes and membrane processes. In *Encyclopedia of Chemical Processing and Design*, Vol. 29, ed. J. J. McKenna and W. A. Cunningham p. 301-350. Marcel Dekker, New York. With permission)



other neighboring penetrants. This phenomenon is generally indicated by a reduction in the selectivity of the membrane in mixed gas situations at elevated partial pressures of a strongly sorbing component like CO<sub>2</sub> or H<sub>2</sub>S.

Additional complexities can influence the selectivity even in the absence of true plasticization. The simplest such effect is caused by gas phase nonideality induced by high pressures of an additional low-sorbing penetrant such as CH<sub>4</sub> in CO<sub>2</sub>/CH<sub>4</sub> mixtures. This effect can be understood with the Lewis-Randall rule that accounts for nonideal P-V-T behavior of gas mixtures [Koros and Fleming, 1993].

Operating temperature often influences the choice of polymer membrane material and can also affect the module design. In general, polymers exhibit increasing flux and decreasing selectivity with increasing temperature and suffer a catastrophic loss of selectivity near and above their **glass transition temperatures** ( $T_g$ ). This can be understood by modeling a rubbery polymer as a highly mobile liquid. Such a matrix presents little hindrance to particle diffusion and thus would yield low selectivity, except in the vapor-gas applications noted earlier.

## 61.6 Module Design Considerations

---

The efficiency of gas separation membrane modules can be optimized by the control of various parameters. Among the most important are membrane type, membrane surface area, flow pattern, feed pressure, and feed composition. A detailed discussion of how these parameters interact is given by Koros and Chern [1987] and more recently by Zolandz and Fleming [1992]. These authors and others [Bhide and Stern, 1991a, b; Prasad *et al.*, 1994] also consider the economics of various approaches to separation using competitive technologies.

### Defining Terms

**Dialysate:** The portion of the feed stream that penetrates a dialysis membrane (see **permeate**).

**Diffusivity:** A measure of a component's ability to diffuse through membrane materials.

**Glass transition temperature ( $T_g$ ):** The temperature below which large-scale segmental motions of a polymer are suppressed. Above the  $T_g$ , amorphous polymers act as mobile, viscoelastic liquids. Below the  $T_g$ , amorphous polymers act as rigid glasses.

**Glassy polymer:** An amorphous polymer below its glass transition temperature.

**Nonpermeate:** See **retentate**.

**Permeability:** A measure of the ease with which a penetrant passes through a membrane material.

**Permeate:** The portion of the feed stream that penetrates a membrane.

**Permselectivity:** A measure of the ability of a membrane to separate components based on a difference in permeabilities.

**Plasticizer:** A component added to a polymer to change its properties. Often added to increase processibility by lowering the  $T_g$ .

**Retentate:** The portion of the feed stream that does not penetrate a membrane.

**Semipermeable:** A membrane that allows some viscous flow of compounds of certain molecular

sizes or structures.

**Solubility:** A measure of a component's ability to absorb into a polymer matrix.

**Supercritical:** A gas at a temperature above its critical temperature.

**Transport plasticization:** Increase in the diffusivity of a penetrant caused by the high penetrant sorption levels. Leads to dramatic increases in permeability, but often at the expense of decreases in selectivity.

## References

- Bhide, B. D. and Stern, S. A. 1991a. A new evaluation of membrane processes for the oxygen-enrichment of air. I: Identification of optimum operating conditions and process configuration. *J. Membrane Sci.* 62(1):13–36.
- Bhide, B. D. and Stern, S. A. 1991b. A new evaluation of membrane processes for the oxygen-enrichment of air. II: Effects of economic parameters and membrane properties. *J. Membrane Sci.* 62(1):37–58.
- Henis, J. M. S. and Tripodi, M. K. 1980. *Multicomponent Membranes for Gas Separation*. U.S. Patent 4,230,463.
- Klein, E., Ward, R. A., and Lacey, R. E. 1987. Membrane processes—Dialysis and electrodialysis. In *Handbook of Separation Process Technology*, ed. R. W. Rousseau, p. 862–953. John Wiley & Sons, New York.
- Koros, W. J. 1988. Membranes and membrane processes. In *Encyclopedia of Chemical Processing and Design*, Vol. 29, ed. J. J. McKetta and W. A. Cunningham, p. 301–350. Marcel Dekker, New York.
- Koros, W. J. and Chern, R. T. 1987. Separation of gaseous mixtures using polymer membranes. In *Handbook of Separation Process Technology*, ed. R. W. Rousseau, p. 862–953. John Wiley & Sons, New York.
- Koros, W. J. and Fleming, G. K. 1993. Membrane-based gas separation. *J. Membrane Sci.* 83(1):1–80.
- Prasad, R., Shaner, R. L., and Doshi, K. J. 1994. Comparison of membranes with other gas separation technologies. In *Polymeric Gas Separation Membranes*, ed. D. R. Paul and Y. P. Yampol'ski, p. 513–614. CRC Press, Boca Raton, FL.
- Zolandz, R. R. and Fleming, G. K. 1992. Gas permeation. In *Membrane Handbook*, ed. W. S. W. Ho and K. K. Sirkar, p. 54–102. Van Nostrand Reinhold, New York.

## Further Information

- Kesting, R. E. and Fritzsche, A. K. 1993. *Polymeric Gas Separation Membranes*. John Wiley & Sons, New York.
- Koros, W. J., Coleman, M. R., and Walker, D. R. B. 1992. Controlled permeability membranes. In *Annual Review of Materials Science*, Vol. 22, ed. R. A. Huggins, J. A. Giordmaine, and J. B. Wachtman Jr., p. 47–89. Annual Reviews, Palo Alto, CA.

Shiao-Hung Chiang, Daxin He. “Fluid-Particle Separation”  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



## Fluid-Particle Separation

---

### 62.1 Equipment

### 62.2 Fundamental Concept

### 62.3 Design Principles

Cake Filtration • Ultrafiltration

### 62.4 Economics

#### **Shiao-Hung Chiang**

*University of Pittsburgh*

#### **Daxin He**

*University of Pittsburgh*

Fluid-particle separation plays a key role in nearly all major manufacturing industries (including chemical, mineral, paper, electronics, food, beverage, pharmaceutical, and biochemical industries), as well as in energy production, pollution abatement, and environmental control. It also serves to fulfill the vital needs of our daily life, since we must have cartridge oil/fuel filters for operating an automobile, a paper filter for the coffee machine, a sand filter bed for the municipal water treatment plant, and so on. In fact, modern society cannot function properly without the benefit of the fluid-particle separation.

Technically, fluid-particle separation involves the removal and collection of a discrete phase of matter (particles) existing in a dispersed or colloidal state in suspension. This separation is most often performed in the presence of a complex medium structure in which physical, physicochemical and/or electrokinetic forces interact. Their analysis requires combined knowledge of fluid mechanics, particle mechanics, solution chemistry, and surface/interface sciences.

Although the industrial equipment classified as fluid-particle separation devices are too numerous to be cited individually, it is generally accepted that the following are the most important categories: (1) **screening**, (2) **thickening/sedimentation**, (3) filtration (**cake filtration** and **deep-bed filtration**), (4) **cycloning**, (5) **centrifugation**, (6) **flotation**, and (7) **membrane filtration** (including **ultrafiltration**).

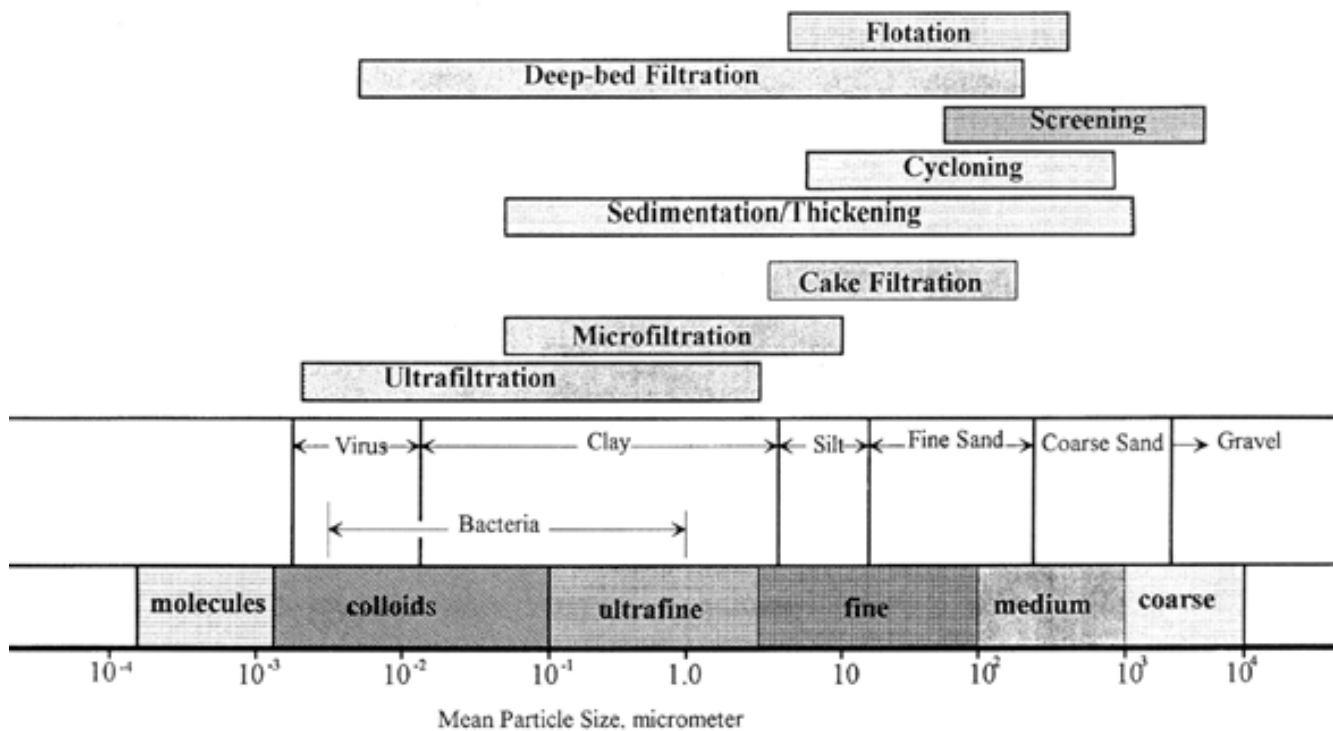
Among these, filtration is the most widely used fluid-particle separation process in industrial applications. Historically, filtration is one of the oldest "tools" used by man. Its modern day application began in 1789 when the French government issued the first patent on a filter as a fluid-particle separation device to Joseph Amy [Orr, 1977]. Recent advances in filtration

technology have focused on the application of various field forces, including hydrodynamic, electric, magnetic, and acoustic field [Muralidhara, 1989], and the use of membranes as the filter medium [Cheryan, 1986]. The most notable developments are cross-flow filtration (using hydrodynamic force to prevent cake formation) and ultrafiltration (using polymeric membranes as the filter medium). These new techniques have led to improved product recovery and quality, decreased energy consumption, and improved handling and transport of the processed materials.

## 62.1 Equipment

The most important criterion for the selection of equipment for a given application of fluid-particle separation is the particle size of the system. Figure 62.1 shows the general range of applicability of major types of equipment in terms of the particle size and representative materials involved. Of course, this representation is an oversimplification of the selection process, as many other factors are not considered. For example, the solid concentration in the feed mixture (suspension) can influence the choice of equipment type. In general, the deep-bed filtration is best for treating dilute slurry with solid concentration less than 1%, whereas cake filtration is the method of choice for slurries having solid concentration greater than 1%.

**Figure 62.1** Equipment selection for fluid-particle separation based on particle size.



It should also be pointed out that the various filtration processes (deep-bed filtration, cake filtration, microfiltration, and ultrafiltration) cover nearly the entire range of particle size.

Therefore, the term *filtration* is often used as a synonym to represent the field of *fluid-particle separation*. The most commonly used filtration equipment is given in [Table 62.1](#).

**Table 62.1** Filtration Equipment

Discontinuous Filters	Semicontinuous Filters	Continuous Filters
Plate and frame filter press Leaf filter Tray filter	Rotary pan filter Semicontinuous belt filter Automatic filter press Electrical precipitator	Drum filter Rotary disk filter Vacuum belt filter Rotary disk cross-flow filter Rotating cylinder cross-flow filter

A detailed procedure for equipment selection for a given requirement in fluid-particle separation can be found in *Perry's Chemical Engineers' Handbook* [[Perry and Green, 1984](#)].

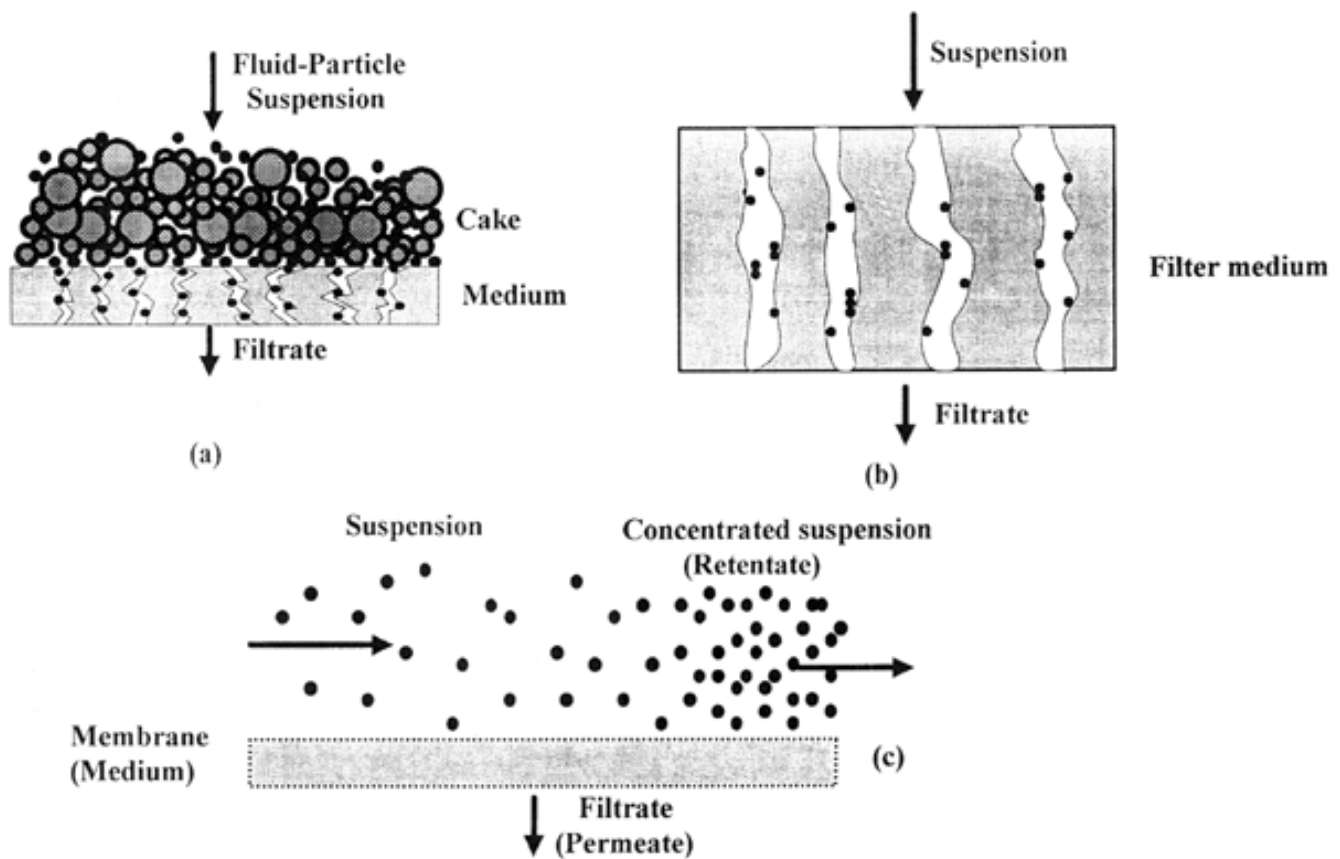
## 62.2 Fundamental Concept

There are two general types of operations for separating solid particulate matter from a fluid phase. In the first type the separation is accomplished by moving the particles through a constrained fluid phase. The particle movement is induced by a body force, such as gravity or electrical potential. For example, in sedimentation and thickening, the solid particles settle due to a difference in density between solid and fluid under the influence of gravitational or centrifugal acceleration. In an electrical precipitator, particles are attracted toward an oppositely charged surface due to an externally imposed electrical potential difference.

In the second type of operation, exemplified by the filtration process, the separation is accomplished by contacting the fluid-particle suspension with a porous medium (see [Fig. 62.2](#)). The porous medium acts as a semipermeable barrier that allows the fluid to flow through its capillary channels and retains the solid particles on its surfaces. Depending on the mechanism for arrest and accumulation of particles, this type of separation can be further divided into two classes [[Perry and Green, 1984](#)]: deep-bed filtration and cake filtration.

**Figure 62.2** Mechanisms of filtration: (a) cake filtration, (b) deep-bed filtration, (c) cross-flow filtration (membrane filtration–ultrafiltration). (*Source*: McCabe, W. L., Smith, J. C., and Harriott, P. 1993. *Unit Operations of Chemical Engineering*, p. 1003. McGraw-Hill, New York. With permission.)

**Figure 62.2**



Deep-bed filtration is also known by terms such as *blocking filtration*, *surface filtration*, and *clarification* [see Fig. 62.2(b)]. This type of filtration is preferred when the solid content of the suspension is less than 1%. In such an operation, a deep bed of packing material (e.g., sand, diatomite, or synthetic fibers) is used to capture the fine solid particles from a dilute suspension. The particles to be removed are several orders of magnitude smaller than the size of the packing material, and they will penetrate a considerable depth into the bed before being captured. The particles can be captured by several mechanisms [Tien, 1989]: (1) the direct-sieving action at the constrictions in the pore structure, (2) gravity settling, (3) Brownian diffusion, (4) interception at the solid-fluid interfaces, (5) impingement, and (6) attachment due to electrokinetic forces.

Cake filtration is the most commonly used industrial process for separating fine particles from a fluid-particle suspension. In cake filtration the filtered particles are stopped by the surface of a filter medium (a porous barrier) and then piled upon one another to form a cake of increasing thickness [see Fig. 62.2(a)]. This cake of solid particles forms the "true" filtering medium. In the case of liquid filtration a filter cake containing filtrate (the liquid) trapped in the void spaces among the particles is obtained at the end of the operation. In many instances where the recovery of the solids is the ultimate objective, it is necessary that the liquid content in the cake be as low as possible. In order to reduce the liquid content, the cake is subjected to a deliquoring process by applying desaturating forces to the cake. These forces can be mechanical, hydrodynamic, electrical, or acoustic in nature [Muralidhara, 1989].

When the mean particle size is less than a few micrometers, the conventional cake filtration operation becomes ineffective, primarily due to the formation of high-resistance filter cake. To overcome this obstacle, cross-flow filtration (often coupled with ultrafiltration) is used to limit the cake growth. In the cross-flow configuration (e.g., in continuous ultrafiltration), the fluid-particle suspension flows tangentially to the filter medium rather than perpendicularly to the medium as in conventional filtration. The shear forces of the flow in the boundary layer adjacent to the surface of the medium continuously remove a part of the cake and thus prevent the accumulation of solid particles on the medium surface. In this manner the rate of filtration can be maintained at a high level to ensure a cost-effective operation.

## 62.3 Design Principles

---

### Cake Filtration

In the design of a cake filtration process, the pressure drop,  $\Delta p$ , the surface area of the cake,  $A$ , and the filtration time,  $t$ , are important parameters to be determined. As the filtration proceeds, particles retained on the filter medium form a filter cake (see Fig. 62.3). For an incompressible cake the pressure drop,  $\Delta p$ , across the filter cake and filter medium can be expressed as

$$\Delta p = p_a - p_b = \left( \frac{\alpha m_c}{A} + R_m \right) \mu u \quad (62.1)$$

where  $\mu$  is the viscosity of the filtrate,  $u$  is the velocity of the filtrate,  $m_c$  is the total mass of solids in the cake,  $R_m$  is the filter-medium resistance, and  $\alpha$  is defined as the **specific cake resistance**. The specific cake resistance depends on particle size, shape, and distribution. It is also a function of porosity of filter cake and pressure drop. For incompressible cakes  $\alpha$  is independent of the pressure drop and the position in the filter cake.

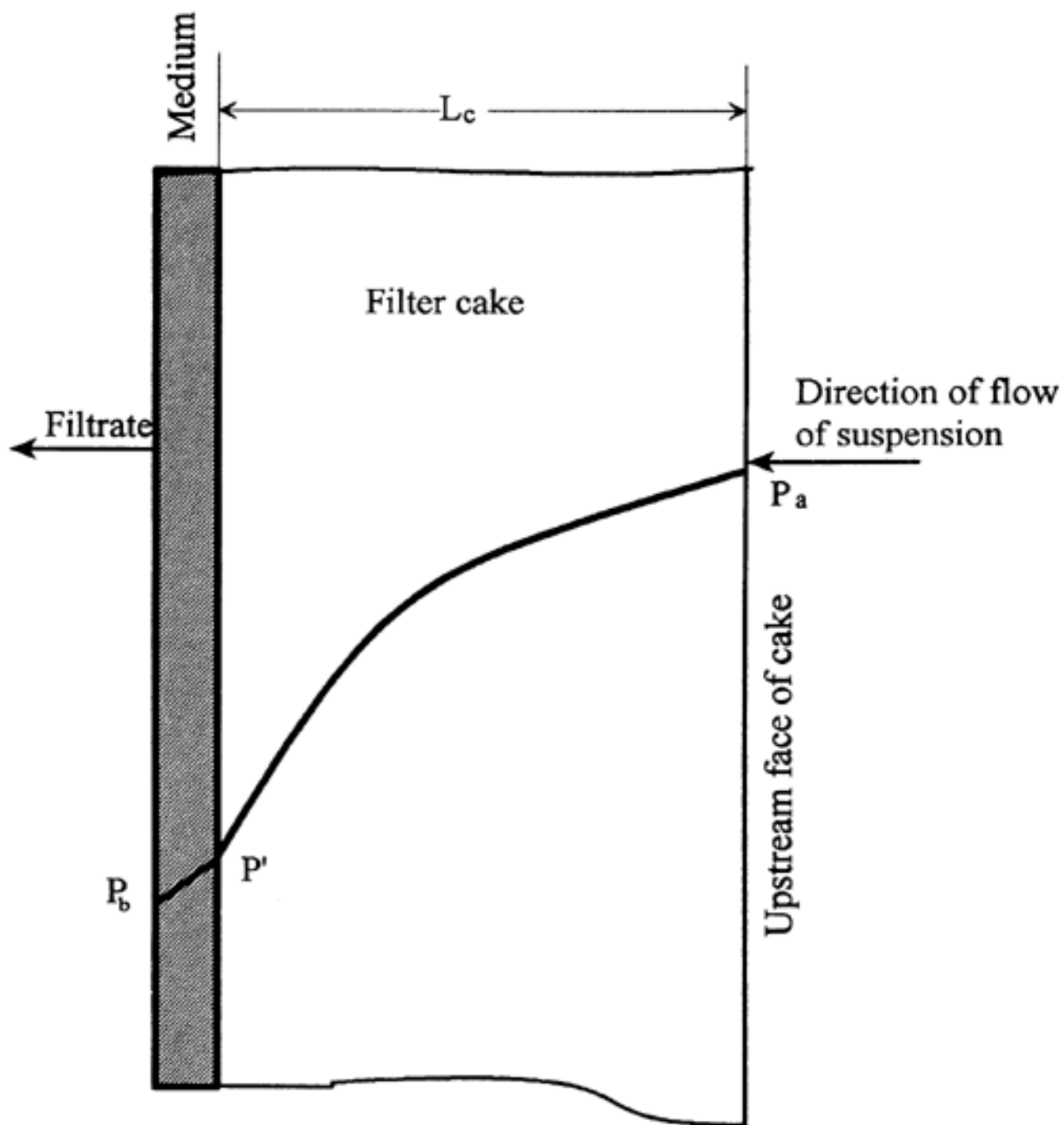
For convenience of data analysis Eq. (62.1) is usually rewritten as follows:

$$\frac{dt}{dV} = \frac{\mu}{A(-\Delta p)} \left[ \frac{\alpha c V}{A} + R_m \right] \quad (62.2)$$

where  $t$  is the filtration time,  $V$  is the volume of filtrate, and  $c$  is the mass of solid per unit filtrate volume. In order to use this equation for design of a cake filtration operation, the specific cake resistance and filter-medium resistance must first be determined by performing a filtration test. The procedure of such a test is discussed in the following example.

**Example 62.1.** The results of five cake filtration tests are plotted in Fig. 62.4. These data will be

**Figure 62.3** Section through filter medium and cake, showing pressure gradients.



used to determine the specific cake resistance and filter medium resistance.

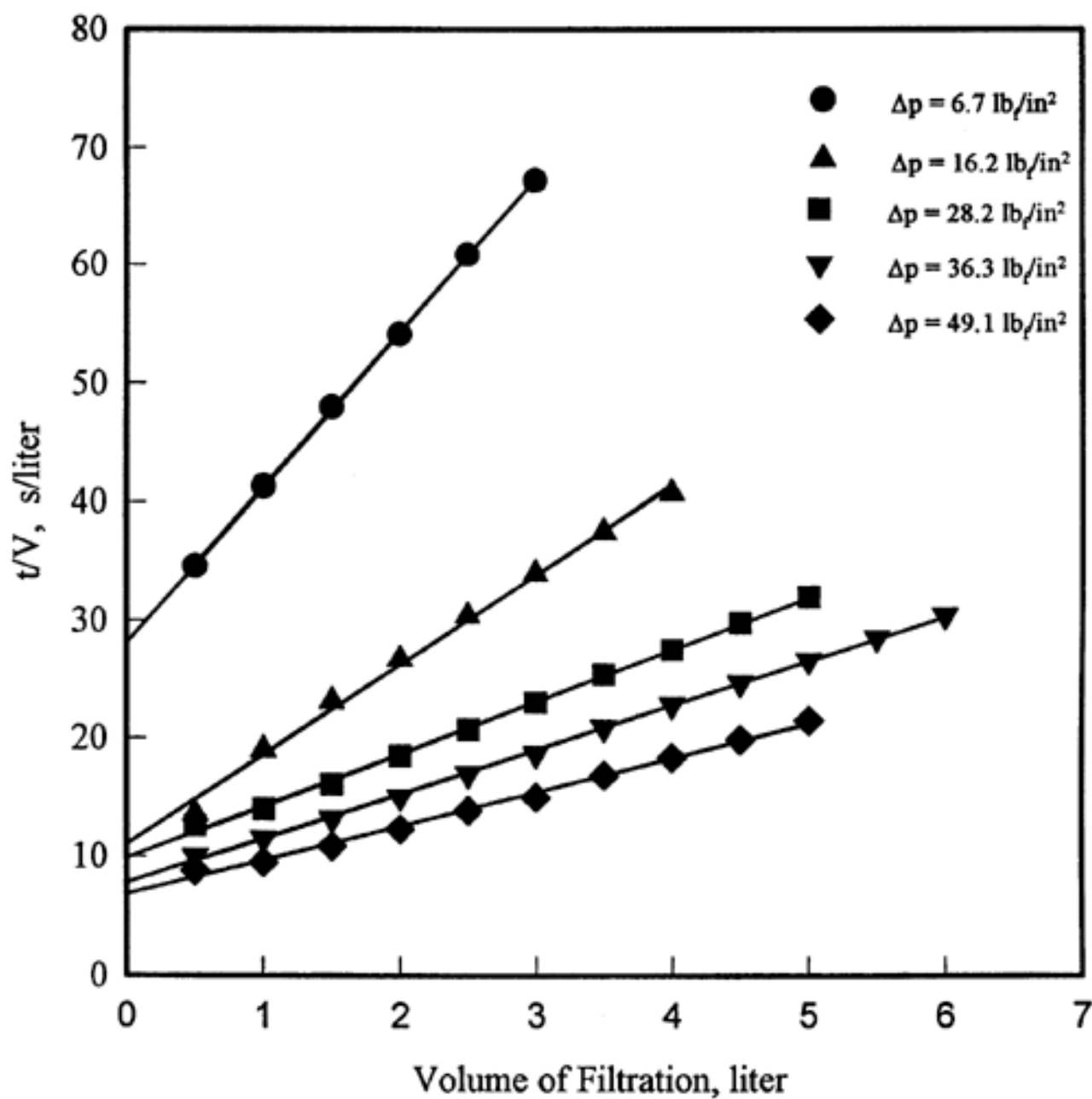
**Solution.** Integration of Eq. (62.2) for the case of constant pressure drop gives

$$\frac{t}{V} = \left( \frac{K_c}{2} \right) V + \frac{1}{q_o} \quad (62.3)$$

where

$$K_c = \frac{\mu c \alpha}{A^2 \Delta p} \quad \text{and} \quad \frac{1}{q_o} = \frac{\mu R_m}{A \Delta p} \quad (62.4)$$

**Figure 62.4** Plot of  $t/V$  versus  $V$ .



The test results yield a linear relation when  $t/V$  is plotted against  $V$  with a slope of  $(K_c/2)$  and an intercept of  $(1/q_o)$ . Using Eq. (62.4), the values of  $\alpha$  and  $R_m$  can be calculated; the results are given in Table 62.2.

**Table 62.2** Values of  $R_m$  and  $\alpha$

Test Number	$\Delta p$ , lbf/ft <sup>2</sup>	$R_m$ , ft <sup>-1</sup> · 10 <sup>-10</sup>	$\alpha$ , ft/lb · 10 <sup>-11</sup>
1	965	1.98	1.66
2	2330	2.05	2.23
3	4060	2.78	2.43
4	5230	2.84	2.64
5	7070	3.26	2.80

Source: McCabe, W. L., Smith, J. C., and Harriott, P. 1993. *Unit Operations of Chemical Engineering*, 5th ed. McGraw-Hill, New York. With permission.

Figure 62.5 is a logarithmic plot of  $\alpha$  versus  $\Delta p$ . From this plot an empirical correlation for  $\alpha$  as a function of pressure drop is obtained:

$$\alpha = 2.85 \cdot 10^{10} \Delta p^{0.26} \quad (62.5)$$

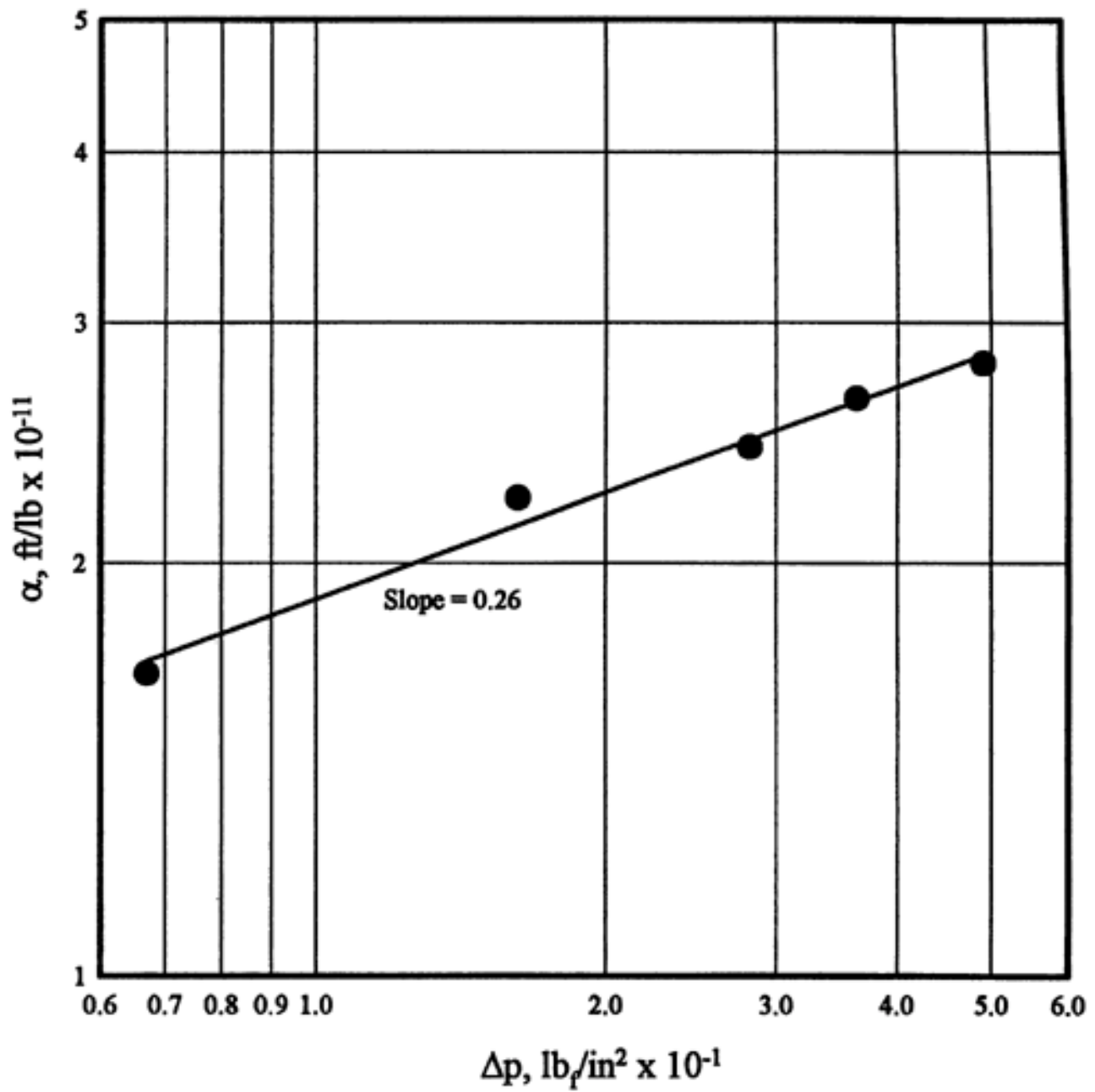
Equations (62.2) and (62.5) can be used for design calculations for a filtration operation with incompressible and slightly compressible filter cakes. For the case of highly compressible cakes the effect of variations in cake porosity on specific cake resistance must be considered [Tiller and Shirato, 1964].

## Ultrafiltration

Ultrafiltration is a membrane process capable of separating or collecting submicrometer-size particles and macromolecules from a suspension or solution. It has been widely used to concentrate or fractionate a solution containing macromolecules, colloids, salts, or sugars. The ultrafiltration membrane can be described as a sieve with pore size ranging from molecular dimension to a few micrometers. It is usually polymeric and asymmetric, designed for high productivity (permeation flux) and resistance to plugging. Ultrafiltration membranes are made commercially in sheet, capillary, and tubular forms. All of these types have commercial niches in which they dominate, but generally there is an overlap in the usage of the different configurations.



**Figure 62.5** Log-log plot of  $\alpha$  versus  $\Delta p$ .



In the design of the ultrafiltration process, either batch operation or continuous operation (employing a cross-flow configuration) can be used. In the batch operation the retentate is returned to the feed tank for recycling through the filter unit. It is the fastest method of concentrating a given amount of material, and it also requires the minimum membrane area.

In order to determine the membrane surface area,  $A$ , for the ultrafiltration process, the following three parameters are required: flux,  $J$ , which is a measure of the membrane productivity; permeate,  $V_p$ , which is the amount of material that has passed through the membrane; and retentate,  $V_R$ , which is the amount of material that has been retained by the membrane. During the batch ultrafiltration operation, flux decreases because of an increase in concentration in the recycled stream. Furthermore, the phenomenon of concentration polarization tends to cause a higher concentration at the membrane surface than that in the bulk. Therefore, an average flux should be used in the design. The average flux,  $J_{av}$ , can be estimated by the following equation:

$$J_{av} = J_f + 0.33(J_i - J_f) \quad (62.6)$$

where  $J_f$  is the final flux at the highest concentration and  $J_i$  is the initial flux. The material balance gives

$$V_f = V_r + V_p \quad (62.7)$$

where  $V_f$ ,  $V_r$ , and  $V_p$  are volume of feed, retentate, and permeate, respectively. The volume concentration ratio is defined as

$$\text{VCR} = V_f/V_r \quad (62.8)$$

The membrane area can be expressed as

$$A = (V_f - V_p)/J_{av} \quad (62.9)$$

These equations are to be used to estimate the membrane surface area as illustrated in the following example.

**Example 62.2.** Estimate the surface area required for concentrating milk at a product rate of 2000 L/h using a batch ultrafiltration process. The filter is operated at 120°F with a flow velocity of 1.0 m/s. The volume concentration ratio (VCR) is 4.0, and the flux varies from 50 L/m<sup>2</sup>/h at the beginning to 20 L/m<sup>2</sup>/h at the end of the operation.

**Solution.**

From Eq. (62.6),  $J_{av} = 20 + 0.33(50 - 20) = 29.9 \text{ L/m}^2/\text{h}$

From Eq. (62.8),  $V_f = 4.0 \times 2000 = 8000 \text{ L/h}$

From Eq. (62.7),  $V_p = 8000 - 2000 = 6000 \text{ L/h}$

From Eq. (62.9),  $A = 6000/29.9 = 200.7 \text{ m}^2$

If a continuous operation is adopted (e.g., using a feed-and-bleed mode of operation), the system

would be operated at a concentration equivalent to the final concentration in a batch operation [Cheryan, 1986]. Therefore, the final flux value,  $20 \text{ L/m}^2/\text{h}$ , should be used in the design calculation,

$$A(\text{continuous operation}) = 6000/20 = 300 \text{ m}^2$$

which gives an estimated membrane surface area for the continuous operation nearly 50% larger than that for the batch operation.

## 62.4 Economics

---

The cost for a given fluid-particle separation process varies widely. The cost for purchasing an industrial filtration equipment can vary from \$500/m<sup>2</sup> to as much as \$50 000/m<sup>2</sup> of filter area. Such a large variation in cost is due to a wide variety of individual features and materials of construction required by specific applications. A good source of information on the cost of common industrial filtration equipment can be found in *Perry's Handbook* [Perry and Green, 1984].

The process cost for an ultrafiltration plant ranges from 600 to 1200/m<sup>2</sup> of membrane area. These costs appear to be comparable to the conventional filtration equipment. However, if the cost is expressed in terms of dollar per unit quantity of feed or product, the ultrafiltration process becomes more expensive. For example, in treating oily waste water, the total process cost of ultrafiltration is over \$3.00 per 1000 gallons of water [Cheryan, 1986], which is more than an order of magnitude higher than the cost of conventional separation technologies. On the other hand, there are certain separations, such as the production of ultrapure water or the recovery of protein from cheese whey, that cannot be accomplished by any conventional methods. In these instances, particularly in the case of high-valued products, the added costs for ultrafiltration are justified.

### Defining Terms

**Cake filtration:** The separation of particles is effected by contacting the fluid-particle suspension with a porous filter medium (made of cloth, synthetic fibers, or metals). The filter medium allows the fluid to flow through its pores while it retains the particles on its surface to form a cake. As filtration proceeds, the cake of solid particles grows in thickness and becomes the "true" filtering medium.

**Centrifugation:** Centrifugation is a separation process based on the centrifugal force either to hold the material in it or to let the material to pass through it. Separation is achieved due to the difference in density.

**Cycloning:** Cycloning is a centrifugal separation process. The feed is introduced tangentially into the cylindrical portion of a cyclone, causing it to flow in a tight conical vortex. The bulk of the fluid leaves upward through a pipe located at the center of the vortex. Solids particles are thrown to the wall and discharged with a small portion of the fluid through the bottom apex of the cyclone.

**Deep-bed filtration:** In this type of filtration a deep bed of packing materials, such as sand, diatomite, or synthetic fibers, is used as the filter medium. The particles are captured within the packed bed while the fluid passes through it.

**Flotation:** Flotation is a gravity separation process based either on the use of a dense medium in which the desired particles will float or on the attachment of gas bubbles to particles, which are then carried to the liquid surface to be separated.

**Membrane filtration:** In membrane filtration a thin permeable film of inert polymeric material is used as the filter medium. The pore size of the membrane ranges from molecular dimension to a few micrometers. It is widely used to collect or fractionate macromolecules or colloidal suspensions. It is also applied to beverage filtration and preparation of ultrapure water.

**Screening:** Screening is an operation by which particles are introduced onto a screen of a given aperture size to separate particles of different sizes.

**Specific cake resistance:** Specific cake resistance is the resistance of a filter cake having unit weight of dry solids per unit area of filtration surface.

**Thickening/sedimentation:** Thickening/sedimentation is a gravity-settling process that removes the maximum quantity of liquid from a slurry and leaves a sludge for further processing.

**Ultrafiltration:** Ultrafiltration is a special type of membrane filtration. It is used for concentration and purification of macromolecular solutes and colloids in which the solution is caused to flow under pressure parallel to a membrane surface (in a cross-flow configuration). Solutes (or submicrometer particles) are rejected at the semipermeable membrane while the solvents and small solute molecules pass through the membrane.

## References

- Cheryan, M. 1986. *Ultrafiltration Handbook*. Technomic, Lancaster, PA.
- McCabe, W. L., Smith, J. C., and Harriott, P. 1993. *Unit Operations of Chemical Engineering*, 5th ed., p. 994–1077. McGraw-Hill, New York.
- Muralidhara, H. S. (Ed.) 1989. *Solid/Liquid Separation*. Battelle Press, Columbus, OH.
- Orr, C. 1977. *Filtration Principles and Practices*, 2nd ed. Marcel Dekker, New York.
- Perry, R. H. and Green, D. W. (Ed.) 1984. *Perry's Chemical Engineers' Handbook*, 6th ed. Chapters 19 and 20. McGraw-Hill, New York.
- Tien, C. 1989. *Granular Filtration of Aerosols and Hydrosols*. Butterworths, Stoneham, MA.
- Tiller, F. M. and Shirato, M. 1964. The role of porosity in filtration: VI, new definition of filtration resistance. *AIChE J.* 10(1):61–67.

## Further Information

An excellent in-depth discussion on the theory and practice of fluid-particle separation is presented in *Solid-Liquid Separation*, third edition, by Ladislav Svarovsky, Butterworths, London, 1990.

The proceedings of the annual American Filtration and Separation Society meeting and the biannual World Filtration Congress document new developments in all aspects of fluid-particle separation.

There are four major journals covering the field of fluid-particle separation:

*Fluid-Particle Separation Journal*. Published by the American Filtration and Separation Society, P.O. Box 1530, 13712 Country Club Dr., Northport, AL 35476.

*Particulate Science and Technology: An International Journal*. Taylor & Francis, 1101 Vermont Avenue, Suite 200, Washington, DC 20005.

*Filtration & Separation*. Published by Elsevier, Amsterdam, New York.

*Separations Technology*. Published by Butterworth-Heinemann, 80 Montrale Avenue, Stoneham, MA 02180.

Corder, W. C., Hanson, S. P. "Other Separation Processes"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Other Separation Processes

---

- 63.1 Sublimation
- 63.2 Diffusional Separations
- 63.3 Adsorptive Bubble Separation
- 63.4 Dielectrophoresis
- 63.5 Electrodialysis

**William C. Corder**

*CONSOL, Inc.*

**Simon P. Hanson**

*CONSOL, Inc.*

For the purposes of this publication, "other separation processes" will be confined to sublimation, diffusional separations, adsorptive bubble separation, dielectrophoresis, and electrodialysis.

Sublimation is the transformation of a substance from the solid into the vapor state without formation of an intermediate liquid phase. Desublimation is the reverse. **Sublimation processes** may be used to separate solids not easily purified by more common techniques.

Diffusional separation processes may be employed when the substances to be separated are quite similar (e.g., the separation of isotopes). Gaseous diffusion, thermal diffusion, pressure diffusion, and mass diffusion fit this category. Gaseous diffusion provides separation by selectively permitting constituents of a gas mixture to flow through extremely small holes in a barrier or membrane. Molecules of species in the mixture for which the mean free path is smaller than the holes will pass through (i.e., separate), whereas the bulk flow of the gas will not. Thermal diffusion depends on a temperature gradient for separation and applies to liquids and gases.

Use of pressure diffusion for separation of gaseous mixtures has been largely confined to the laboratory, although considerable work has been done on the use of the gas centrifuge for separating isotopes of uranium. The separation method depends on the imposition of a pressure gradient. Mass diffusion involves the separation of a gas mixture by diffusion into a third component, or **sweep gas**. The sweep gas establishes a partial-pressure gradient by introducing it as a gas through a porous wall and then letting it condense as a liquid at another region of the process. The sweep gas sweeps the less diffusible component along with it, thus separating this

component of the mixture.

Adsorptive-bubble separation utilizes the selective attachment of materials onto the surfaces of gas bubbles passing through a solution or, commonly, a suspension [Lemlich, 1966]. The bubbles rise to create a foam or froth which is swept away, allowing the collapse of the bubbles and recovery of the adsorbed. Additives commonly are included in the suspension to depress the adsorption of some species and enhance the attachment of others. Probably the greatest use of adsorptive bubble separation is in the minerals-processing industries.

Dielectrophoresis (DEP), related electrophoresis, and electrodialysis all depend on electric gradients to effect a separation. Electrophoresis involves the motion of charged particles in a uniform electric field, whereas DEP applies to neutral, polarizable matter in a nonuniform electric field. DEP works best with larger-than-molecular-sized particles. Electrodialysis is the removal of electrolytes from a solution through an ion-selective membrane by means of an applied electric potential gradient. The primary application of electrodialysis has been the desalination of seawater and brackish water. Most commercial membranes were developed for this purpose.

## 63.1 Sublimation

---

Sublimation has been used most often for separation of a volatile component from other components which are essentially nonvolatile. There has been some interest in separating mixtures of volatile components by sublimation [Gillot and Goldberger, 1969].

The vapor pressure of the subliming components must be greater than their partial pressures in the gas phase in contact with the solid. Usually, the solid must be heated and the gaseous environment in contact with the solid must be controlled. Vacuum operation and the use of nonreactive gas or **entrainer** are means of controlling the gaseous environment. The use of vacuum operation has been the most common commercial approach. For pure substances or mechanical mixtures containing only one volatile component, there is no theoretical limit to the purity of the product obtained in a sublimation process.

In a sublimation process comprised of a sublimer (to volatilize the sublimable components) and a condenser (to recover them), the loss per pass of entrainer gas through the system for a system consisting of two sublimable components is

$$\text{Percent loss} = \frac{r(P_{AC} + P_{BC})/(P_{AS} + P_{BS})}{(1 + r) - [(P_{AC} + P_{BC} - \Delta P)/(P_{AS} + P_{BS})]} \times 100 \quad (63.1)$$

where  $r$  is the ratio of moles of inert gas (either unavoidable, as in a vacuum operation, or intentional, as with an entrainer) to the moles of solids sublimed (i.e.,

$r = P_1/(P_{AS} + P_{BS}) = (P - P_{AS} - P_{BS})/(P_{AS} + P_{BS})$ .  $P_A$  and  $P_B$  are the vapor pressures of components  $A$  and  $B$ , subscripts  $S$  and  $C$  refer to the sublimer and condenser,  $P_1$  is the partial pressure of inert gas,  $\Delta P$  is the total pressure drop between sublimer and condenser, and  $P$  is the total pressure in the sublimer.

To calculate the **yield per pass**, a material balance must be made on the sublimer and condenser using the calculated loss per pass. For vacuum sublimation where recycle of gas is not possible, the yield of condensed solids is simply the percent loss calculated by Eq. (63.1) subtracted from 100



## 63.2 Diffusional Separations

---

Gaseous diffusion requires many stages to result in a clean separation. It is probably economically feasible only for large-scale separation of heavy isotopes (e.g., uranium). The flow rate of the gases to be separated across the gaseous diffusion barrier is described by the equations for Knudsen and Poiseuille flow, the combination of which provides

$$N_T = \frac{a}{\sqrt{M}}(P_F - P_B) + \frac{b}{\mu}(P_F^2 - P_B^2) \quad (63.2)$$

where  $N_T$  is the molar flow rate of gas per unit area of barrier,  $a$  and  $b$  are functions of temperature and barrier properties,  $M$  is the molecular weight of the gas,  $P_F$  and  $P_B$  are the high and low side pressures on the barrier, and  $\mu$  is the gas viscosity. A comprehensive treatment of the design equations for gaseous diffusion is provided by Pratt [1967].

As with gaseous diffusion, thermal diffusion requires many stages to provide a clean separation. Thus, its use has been confined to isotope separations. The preferred equipment for thermal diffusion evolved into the thermal diffusion or thermogravitational column. In the column, the fluid mixture has a horizontal temperature gradient imposed on it. Thermal convection currents create a countercurrent flow effect providing a large number of individual separation stages in a single piece of equipment. The fundamental equations for separation by thermal diffusion are heavily dependent on the physical properties of the mixture, the properties of the separating column, and the temperatures imposed.

Gas centrifuges and separation nozzles are the tools used to achieve separation by pressure diffusion. In both, a high pressure gradient is created to segregate the lighter components from the heavier ones. The gas centrifuge is the more highly developed. High energy requirements for operation have confined the use of the technology to the separation of mixtures that are very hard to separate (e.g., isotopes). The maximum separative work a gas centrifuge can accomplish per unit time [Olander, 1972] is

$$\Delta U = \frac{\pi Z C D}{2} \frac{(\Delta M V^2)^2}{(2RT)} \quad (63.3)$$

where  $\Delta U$  is the separation in moles/unit time,  $Z$  is the length of the centrifuge,  $C$  is the molar density,  $D$  is the diffusivity of the gas,  $\Delta M$  is the mass difference between the gases to be separated,  $V$  is the peripheral velocity of the centrifuge,  $T$  is the absolute temperature, and  $R$  is the gas constant.

Mass diffusion can be conducted in continuous countercurrent flow columns. The sweep gas moves in a closed cycle between liquid and vapor and (returns to) liquid. The less diffusible component is enriched at the bottom of the column, and the more diffusible one is enriched at the top of the column. A treatment of the theory is provided in Pratt [1967].

## 63.3 Adsorptive Bubble Separation

---

Adsorptive bubble separation techniques are most often employed for the removal of small amounts of either liquid or solids from large amounts of liquid. In its most common use (i.e., **flotation**), fine particulates are removed from liquid. Adsorptive bubble separations can be conducted in staged cells or in columns, the former being the more common type of equipment used. The mechanisms influencing the effectiveness of the separation are adsorption, bubble size and formation, foam overflow and drainage, foam coalescence, and foam breaking. Perry [1984] provides equations applicable to each mechanism. The techniques, in addition to recovery of minerals and coal, have industrial applications in the areas of pollution control, papermaking, food processing, and the removal of organic materials from water.

## 63.4 Dielectrophoresis

---

Separation by DEP is dependent on the fact that, in a nonuniform electric field, a net force will act on even neutral particles in a fluid. Particles will behave differently depending on their polarizability. The net force will direct different particles to regions of varying field strength. A series of equations describing the technique is presented in Pohl [1978]. Application of DEP is reasonably widespread and includes electrofiltration to remove particles from a fluid medium, orientation and separation of biological materials (e.g., cells), and imaging processes.

## 63.5 Electrodialysis

---

The ion-selective membrane used in electrodialysis is key. Ion-selective membranes are identified as anionic or cationic depending on the ion permitted passage through the membrane. Their electrochemical behavior is characterized by a conductivity, a transference number, and a transport number. The conductivity is simply the amount of current which passes through the membrane per unit of imposed electric potential gradient. The transference number is the amount of any substance transported through the membrane per unit of current. The transport number is the fraction of the current carried by a particular ionic species. The conductivity, transference number, and transport number are similarly defined for the solution in contact with the membrane.

If the transport number for an ionic species in solution differs from that inside a membrane, separation will occur when an electric current passes through the membrane. The effectiveness of an ion-selective membrane is quantified by the *permselectivity*, as defined [Winger *et al.*, 1957] for anionic and cationic membranes, respectively, by the equations

$$\Psi_a = \frac{t_a^m - t_a^s}{1 - t_a^s} \quad (63.4)$$

$$\Psi_c = \frac{t_c^m - t_c^s}{1 - t_c^s} \quad (63.5)$$

where  $\Psi$  is the permselectivity,  $t$  is the transport number, subscripts  $a$  and  $c$  designate anion or cation, and superscripts  $m$  and  $s$  designate membrane or solution. At low solute concentrations, the transport number for anions in an anionic membrane, or cations in a cationic membrane, approaches unity. As the concentration of solute increases, membrane conductivity increases, but the transport number decreases. This is due to compression of the diffuse charge double layer within the membrane's pores such that passage of excluded electrolyte species cannot be prevented.

Electrodialysis is a selective transport process that directly uses electrical energy to effect separation of charged species from a solvent. The theoretical thermodynamic minimum energy to effect a separation [Spiegler, 1958] is given by

$$U = 2RT(C_i - C_{od}) \left[ \frac{\log \left( \frac{c_i}{c_{oc}} \right)}{\left( \frac{c_i}{c_{oc}} - 1 \right)} - \frac{\log \left( \frac{c_i}{c_{od}} \right)}{\left( \frac{c_i}{c_{od}} - 1 \right)} \right] \quad (63.6)$$

where  $R$  is the ideal gas constant,  $T$  is the absolute temperature,  $C$  is the concentration in equivalents, and the subscripts  $i$ ,  $oc$ , and  $od$  indicate the inlet, concentrate outlet, and dilute outlet streams. The direct use of electric energy is most effective at low concentrations since energy is expended only on removing the contaminant. The actual energy consumed is usually 10 to 20 times the theoretical.

Electrodialysis is competitive with other separation processes for low concentration feeds, or for situations in which the specificity of electrodialysis makes it uniquely suitable.

## Defining Terms

**Entrainer:** Nonreactive, gaseous diluent used to assist in the sublimation process.

**Flotation:** The process of removing fine particulate matter from aqueous solution by attachment of the particulate to air bubbles.

**Sublimation process:** Either one or both of a combination of sublimation and desublimation steps in equipment designed for that purpose.

**Sweep gas:** The vapor introduced in mass diffusion separations which moves through the process and assists in the separation.

**Yield per pass:** The amount of the desirable constituent recovered per pass of the entrainer gas through the sublimation process, usually expressed as a percent of the total.

## References

Gillot, J. and Goldberger, W. M. 1969. *Chem. Eng. Prog. Symp. Ser.* 65(91): 36–42.

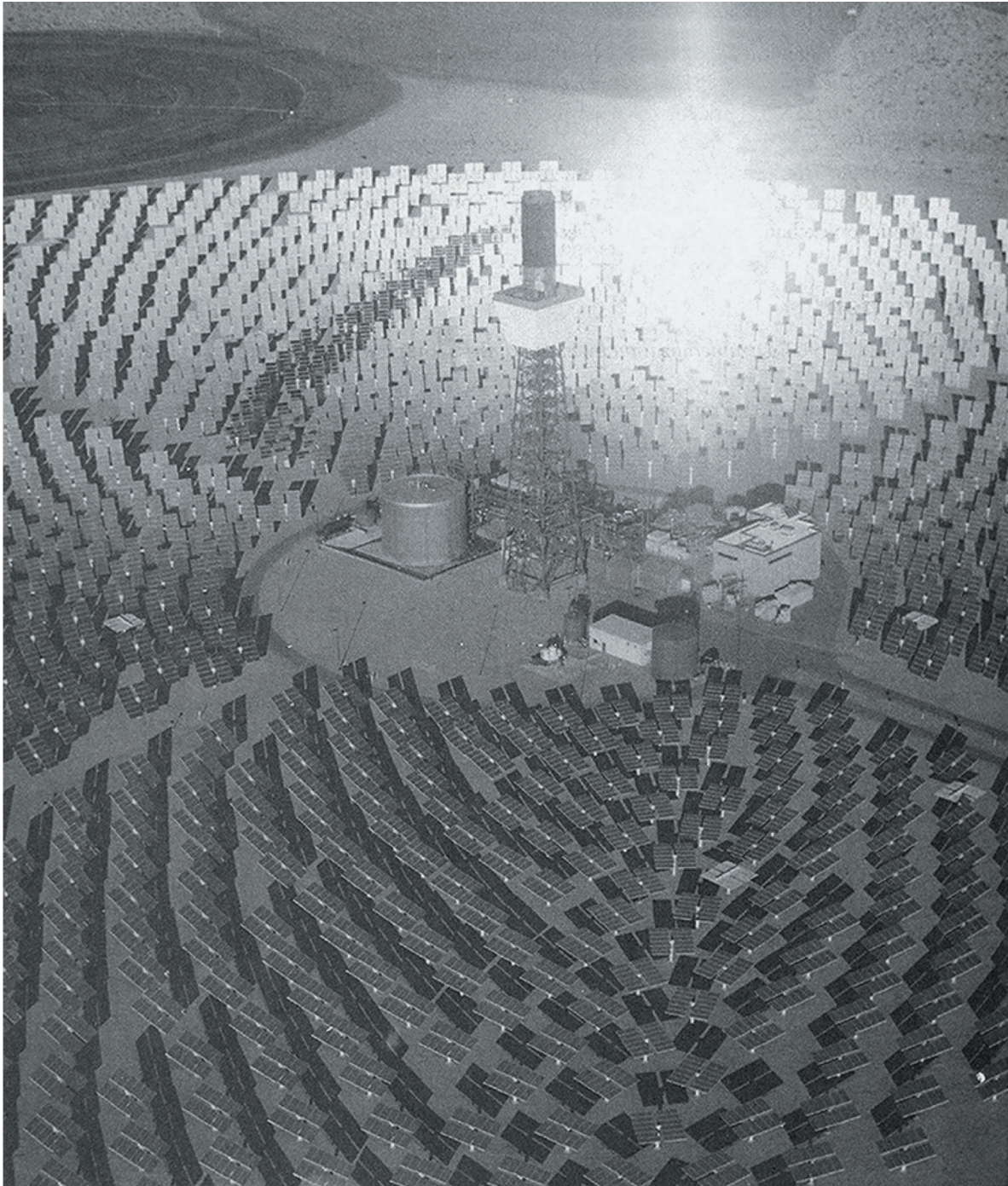
- Lemlich, R. (Ed.) 1972. *Adsorptive Bubble Separation Techniques*. Academic Press, New York.
- Olander, D. R. 1972. Technical basis of the gas centrifuge. *Adv. Nucl. Sci. Technol.* 6:105–174.
- Perry. 1984. *Perry's Chemical Engineers Handbook*. McGraw-Hill, New York.
- Pohl, H. A. 1972. *Dielectrophoresis: The Behavior of Matter in Non-uniform Electric Fields*. Cambridge University Press, New York.
- Pratt, H. R. C. 1967. *Countercurrent Separation Processes*. Elsevier, Amsterdam.
- Spiegler, K. S. 1958. Transport processes in ionic membranes. *Trans. Faraday Soc.* 54:1408.
- Winger, A. G., Bodamer, G. W., and Kunin, R. 1957. Some electrochemical properties of new synthetic ion exchange membranes. *J. Electrochem. Soc.* 100:178.

### **Further Information**

- Wilson, J. R. (Ed.). 1960. *Demineralization by Electrodialysis*. Butterworths Scientific Publications, London.
- Fuerstenau, D. W. (Ed.) 1962. *Froth Flotation*. AIME, New York.

Kreith, K. "Fuels and Energy Conversion"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000





The Mojave Desert is the site of the world's most advanced emission-free central receiver solar thermal power plant, known as Solar Two. Solar Two is designed to demonstrate that molten salt technology is the most efficient and cost-effective method to convert solar energy to electrical energy on a large scale. Molten salt central receiver plants are especially valuable because they can collect thermal energy during the day and store it to produce electricity at night.

The project uses more than 1900 heliostats, or mirrors, to focus the sun's rays on a central tower, where heat is captured in molten salt. When needed, the heat from the molten salt is used to create steam, which drives a conventional steam turbine to produce electricity. Solar Two, using new molten salt technology, is expected to serve as a 10-megawatt blueprint for future commercial solar thermal power plants. For additional information on the central receiver design of Solar Two, see pages 670 and 671. (Courtesy of Southern California Edison. Photo: Southern California Edison.)

# IX

## Fuels and Energy Conversion

---

**Frank Kreith**

*University of Colorado*

- 64 **Fuels** *S. M. A. Moustafa*  
Coal • Oil • Natural Gas • Important Products of Crude Oil Refining
- 65 **Solar Power Systems** *J. M. Chavez, E. M. Richards, and A. Van Arsdall*  
Solar Thermal Systems • Photovoltaic Systems • Biomass Systems
- 66 **Internal Combustion Engines** *A. A. Kornhauser*  
Basics of Operation • Engine Classifications • Spark Ignition Engines • Compression Ignition Engines • Gas Exchange Systems • Design Details • Design and Performance Data for Typical Engines
- 67 **Gas Turbines** *L. S. Langston and G. Opdyke, Jr.*  
Gas Turbine Usage • Gas Turbine Cycles • Gas Turbine Components
- 68 **Nuclear Power Systems** *D. M. Woodall*  
Nuclear Power Applications • Nuclear Power Fundamentals • Economics of Nuclear Power Systems
- 69 **Power Plants** *M. M. El-Wakil*  
The Rankine Cycle • The Turbine • The Condenser • The Condenser Cooling System • The Feedwater System • The Steam Generator • Cycle and Plant Efficiencies and Heat Rates
- 70 **Wind Turbines** *A. Swift and E. Moroz*  
Fundamentals • Power Regulation and Control • Energy Capture Estimation • Stand-Alone Applications • Cost of Energy Calculations • Environmental and Social Cost Issues • Summary
- 71 **Hydraulic Turbines** *R. E. A. Arndt*  
General Description • Principles of Operation • Factors Involved in Selecting a Turbine
- 72 **Steam Turbines** *G. Shibayama and R. D. Franceschinis*  
Types of Steam Turbines • Impulse versus Reaction Turbines • Thermodynamics • Stop and Control Valves • Water Induction Protection • Generators • Turbine Generator Auxiliaries
- 73 **Cogeneration** *D. H. Cooke*  
Cogeneration Fundamentals • Examples of Cogeneration
- 74 **Electric Machines** *E. K. Stanek*  
Induction Machines • Synchronous Machines • DC Machines

ENERGY IS THE MAINSTAY of an industrial society. At present, most useful energy is derived from fossil fuels, mainly coal, oil, and natural gas. But, due to the finiteness of these resources, increased interest has recently developed in using solar energy, wind energy, and water power.

This section covers the basic elements of energy generation and conversion technologies. These include internal combustion engines, gas turbines, nuclear power, wind turbines, hydraulic turbines, steam turbines, and solar energy. In recent years, energy efficiency has become important, and one of the best ways to derive more useful energy from a given fuel source is to

combine electric power generation with heat production in cogeneration systems. This is also covered in this section, as are the principles of the electric machines necessary to convert the shaft power from rotating turbines and other prime movers into electricity.

The contributors to this section have done an excellent job of providing concise and practical coverage of a large and important topic. The author(s) of each chapter had to cover a large field in a few pages. As a result, the information is presented in condensed form and is confined mainly to principles without the details of engineering know-how and design. In most cases, the information presented will be complete enough to provide the reader with useful background; when this is not the case, the references will point the way to additional information.



Moustafa, S. M. A. "Fuels"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 64

## Fuels

---

### 64.1 Coal

Coal Class • Coal Analysis • The Heating Value (Btu) • Agglomerating Character • Fixed Carbon (FC) Limit • Coal Heating Value • Percent Volatile Matter (VM)

### 64.2 Oil

Oil Formation • Oil Composition • Crude Oil Refining • Properties of Petroleum Derivatives • Flash Point • Pour Point

### 64.3 Natural Gas

### 64.4 Important Products of Crude Oil Refining

**Safwat M. A. Moustafa**

*California Polytechnic University*

Materials that possess chemical energy are known as *fuels*. The general categories of fuels are fossil fuels, nuclear fuels, and renewable energy sources. Fossil fuels release their chemical energy during combustion. Nuclear fuels are those that release their chemical energy by nuclear reaction.

There are three general classes of fossil fuels—coal, oil, and natural gas. Other fuels, such as shale oil, tar-sand, and fossil-fuel derivatives are commonly lumped under one of the three main fossil fuel categories.

Fossil fuels were produced from fossilized carbohydrate compounds with the chemical formula  $C_x(H_2O)_y$ . These compounds were produced by living plants in the photosynthesis process. After the plants died, the carbohydrates were converted by pressure and heat, in the absence of oxygen, into hydrocarbon compounds with a general chemical formula of  $C_xH_x$ . Although hydrocarbon compounds are composed of only carbon and hydrogen, in some molecules the same number of hydrogen atoms can be arranged in various structures to produce compounds that are strikingly different in chemical and physical properties.

## 64.1 Coal

---

Coal is the most abundant fossil fuel. It is thought to be fossilized vegetation. At least 20 ft of compacted vegetation is necessary to produce a 1 ft seam of coal. The compacted vegetation—in the absence of air and in the presence of high pressure and temperature—is converted into peat (a very low-grade fuel), then brown coal, then lignite, then subbituminous coal, then **bituminous coal**, and finally anthracite coal. As the aging process progresses, the coal becomes harder, the

hydrogen and oxygen content decrease, the moisture content usually decreases, and the carbon content increases. Coal is normally found in the earth's crust. The average seam thickness in the U.S. is approximately 1.65 m.

The American Society for Testing Materials (ASTM) has classified coal into four major classes (Table 64.1) according to the length of its aging process—the oldest being anthracitic coal, followed by bituminous coal, subbituminous coal, and lignitic coal.

**Table 64.1** ASTM Classification of Coals (ASTM D388)

Class	Group	Fixed Carbon Limit % (dry mineral matter-free)		Volatile Matter Limit % (dry mineral matter-free)		Calorific Value Limit, Btu/lb (moist mineral matter-free)		Agglomerating Character
		= or > than	< than	> than	= or < than	= or > than	< than	
1. Anthracitic	a. Metaanthracite	98	—	—	2	—	—	Nonagglomerating
	b. Anthracite	92	98	2	8	—	—	
	c. Semianthracite	86	92	8	14	—	—	
2. Bituminous	a. Low-volatile bituminous coal	78	86	14	22	—	—	Common agglomerating
	b. Medium-volatile bituminous coal	69	78	22	31	—	—	
	c. High-volatile A bituminous coal	—	69	31	—	14 000	—	
	d. High-volatile B bituminous coal	—	—	—	—	13 000	14 000	Agglomerating
	e. High-volatile C bituminous coal	—	—	—	—	10 500	13 000	
3. Subbituminous	a. Subbituminous A coal	—	—	—	—	10 500	11 500	Nonagglomerating
	b. Subbituminous B coal	—	—	—	—	9 500	10 500	
	c. Subbituminous C coal	—	—	—	—	8 300	9 500	
4. Lignitic	a. Lignite A	—	—	—	—	6 300	8 300	
	b. Lignite B	—	—	—	—	—	6 300	

Source: ASTM, *Standards on Gaseous Fuels, Coal and Coke*.

Selection of a coal for a particular application involves consideration of its specific chemical and physical properties, as well as general factors involving storage, handling, or pulverizing. Factors related to furnace design include volume, grate area, and the amount of radiant heating surface.

## Coal Class

The higher-rank coals are classified according to fixed carbon on the dry basis, whereas the lower-rank coals are classified according to heating value (Btu) on the moist basis.

## Coal Analysis

The two basic coal analyses are the proximate analysis and the ultimate analysis. In any coal seam there are two components that can show significant variation throughout the seam—the moisture and the ash. The ash fraction varies because ash is essentially the inorganic matter deposited with the organic material during the compaction process. The moisture content of the coal varies significantly, depending on the exposure to groundwater before mining and during the transportation and the storage before the coal is burned.

The proximate analysis gives the mass fraction of fixed carbon (FC), volatile matter (VM), and

ash of the coal. The analysis can be made by weighing, heating, and burning of small samples of coal. A powdered coal sample is carefully weighed and then heated to 110°C (230°F) for 20 minutes. The sample is then weighed again; the mass loss divided by the original mass gives the mass fraction of moisture ( $M$ ) in the sample. The remaining sample is then heated to 954°C (1750°F) in a closed container for 7 minutes, after which the sample is weighed. The resulting mass divided by the original mass is equal to the mass fraction of volatile matter in the sample.

The sample is then heated to 732°C (1350°F) in an open crucible until it is completely burned. The residue is then weighed; the final weight divided by the original weight is the ash fraction ( $A$ ). The mass fraction of fixed carbon is obtained by subtracting the moisture, volatile matter, and ash fraction from unity. In addition to the FC, the VM, the  $M$ , and the  $A$ , most proximate analyses list separately the sulfur mass fraction ( $S$ ) and the higher heating value (HHV) of the coal.

The ultimate coal analysis is a laboratory analysis that lists the mass fraction of carbon ( $C$ ), hydrogen ( $H_2$ ), oxygen ( $O_2$ ), sulfur ( $S$ ) and nitrogen ( $N_2$ ) in the coal along with the higher heating value (HHV). It also lists the moisture,  $M$ , and ash,  $A$ . This analysis is required to determine the combustion-air requirements for a given combustion system and the size of the draft system for the furnace.

## The Heating Value (Btu)

The heating value represents the amount of chemical energy in a pound mass of coal (Btu/lbm) on the moist basis. There are two heating values—a higher or gross heating value (HHV) and a lower or net heating value (LHV). The difference between these two values is essentially the latent heat of vaporization of the water vapor present in the exhaust products when the fuel is burned in dry air. In actual combustion systems, this includes the water present in the as-burned fuel (the moisture) and the water produced from the combustion of hydrogen, but it does not include any moisture that is introduced by the combustion of air.

## Agglomerating Character

Coals are considered agglomerating if, in a test to determine the amount of volatile matter, they produce either a coherent button that will support a 500 g weight without pulverizing or a button that shows swelling or cell structure.

## Fixed Carbon (FC) Limit

Two formulas used for calculating the fixed carbon limit, on a mineral matter-free (mm-free) basis, are the Parr formula,

$$\text{Dry, mm-free FC} = \frac{FC - 0.15S}{100 - (1.08A + 0.55S)} \times 100 \quad (64.1)$$

and the approximate formula,

$$\text{Dry, mm-free FC} = \frac{\text{FC}}{10 - (M + 1.1A + 0.1S)} \times 100 \quad (64.2)$$

## Coal Heating Value

Two formulas are used for calculating the coal heating value: the Parr formula,

$$\text{Moist, mm-free Btu} = \frac{\text{Btu} - 50S}{100 - (1.08A + 0.55S)} \times 100 \quad (64.3)$$

and the approximate formula,

$$\text{Moist, mm-free Btu} = \frac{\text{Btu}}{100 - (1.1A + 0.1S)} \times 100 \quad (64.4)$$

## Percent Volatile Matter (VM)

The formula used for calculating coal VM value is as follows:

$$\text{Dry, mm-free VM} = 100 - \text{dry, mm-free FC} \quad (64.5)$$

## 64.2 Oil

---

### Oil Formation

Petroleum or crude oil is thought to be partially decomposed marine life. It is normally found in large domes of porous rock. Crude oils are normally ranked in three categories, depending on the type of residue left after the lighter fractions have been distilled from the crude. Under this system the petroleum is classified as paraffin-based crude, asphalt-based crude, or mixed-based crude.

### Oil Composition

Crude oil is composed of many organic compounds, yet the ultimate analysis of all crude oils is fairly constant. The carbon mass fraction ranges from 84 to 87%, and the hydrogen mass fraction from 11 to 16%. The sum of the oxygen and nitrogen mass fractions ranges from 0 to 7%, and the sulfur mass fraction ranges from 0 to 4%.

## Crude Oil Refining

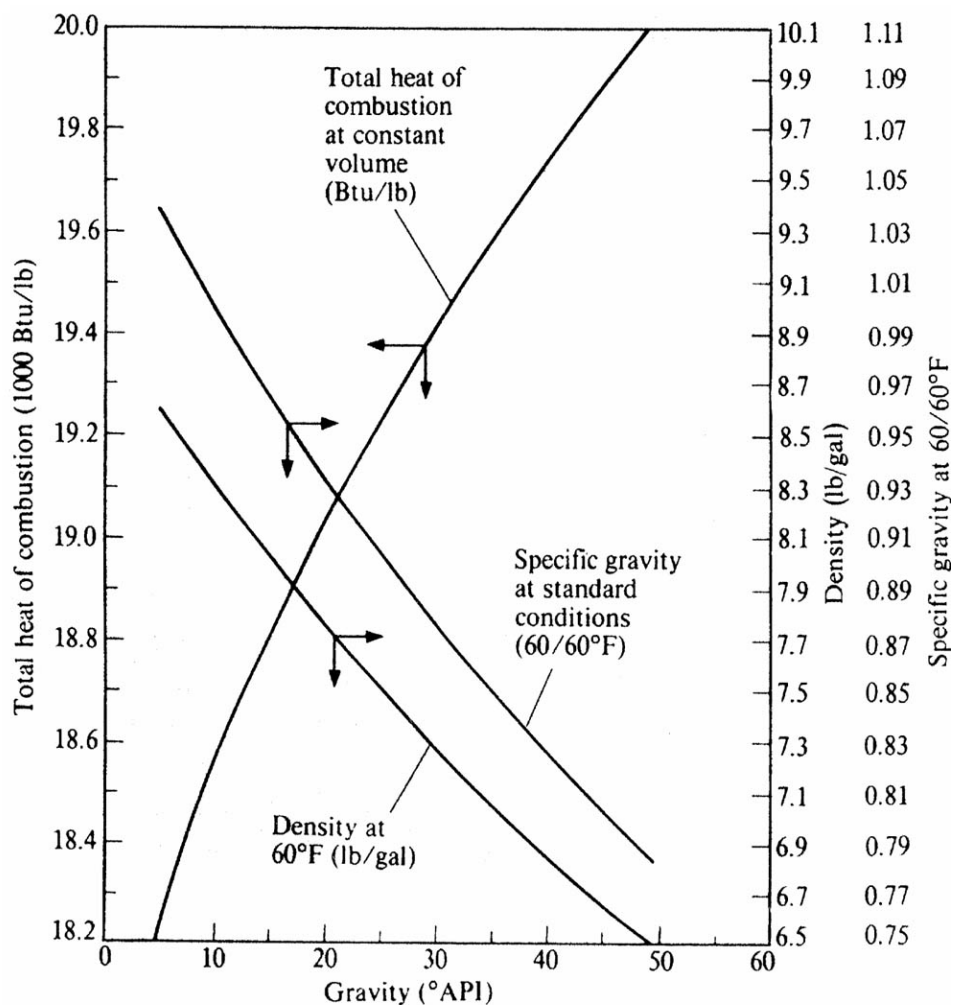
Crude oils are more valuable when defined as petroleum products. Distillation separates the crude oil into fractions equivalent in the boiling range to **gasoline**, **kerosene**, gas oil, lubricating oil, and a residual. Thermal or **catalytic cracking** is used to convert kerosene, gas oil, or residual to gasoline, lower-boiling fractions, and a residual coke. The major finished products are usually blends of a number of stocks, plus additives.

## Properties of Petroleum Derivatives

The main variables of petroleum derivatives are the heating value, the specific gravity ( $s$ ), the flash point, and the pour point. The heating value (usually the higher heating value) is reported in units of either kilojoules per kilogram (or Btu per pound mass) or kilojoules per liter (or Btu per gallon).

The heating value of crude oil and petroleum products is shown as a function of the specific gravity in Fig. 64.1. The heating value (on unit mass basis) of petroleum derivatives increases as the specific gravity of the product decreases.

**Figure 64.1** Properties of petroleum derivatives. (Source: Bobcock and Wilcox Company. 1972. *Steam—Its Generation and Use*, 38th ed. Bobcock and Wilcox Company, New York. With permission.)



The specific gravity of any liquid is the density of the liquid divided by the density of water at 60°F (15.6°C). The specific gravity of crude oil is usually between 0.80 and 0.97. The corresponding °API (American Petroleum Institute) gravity is 45 to 15 degrees. The relationship between the specific gravity and the °API is as follows:

$$^{\circ}\text{API} = \frac{141.5}{\text{Specific gravity at } 60/60^{\circ}\text{F}} - 131.5 \quad (64.6)$$

## Flash Point

The flash point of a liquid fuel is the minimum fluid temperature at which the vapor coming from the fluid surface will just ignite. At a slightly higher temperature, called the *fire point*, the vapors will support combustion. When storing the oil, care should be taken to ensure that its temperature does not exceed its flash point.

## Pour Point

The pour point of a petroleum product is the lowest temperature at which an oil or oil product will flow under standard conditions. It is determined by finding the maximum temperature at which the surface of an oil sample in a standard test tube does not move for 5 s when the test tube is rotated to the horizontal position. The pour point is equal to this temperature plus 5 degrees Fahrenheit.

## 64.3 Natural Gas

---

Natural gas is the only true fossil fuel gas. It is usually trapped in limestone casing on the top of petroleum reservoirs. Reservoir pressures may reach as high as 350 to 700 bar (5000–10 000 lb/in.<sup>2</sup>). Natural gas is primarily composed of methane, with small fractions of other gases.

Natural gas has the highest heating value of all fossil fuels—about 55 800 kJ/kg (24 000 Btu/lbm), or 37 000 kJ/m<sup>3</sup> (1000 Btu/ft<sup>3</sup>) at 1 atm and 20°C (68°F). Natural gas is commonly sold in units of "therms" (1 therm = 100 000 Btu).

There are a number of manufactured fossil fuel gases, including liquified petroleum gas (LPG), synthetic or substitute natural gas (SNG), and primary flash distillate (PFD).

## 64.4 Important Products of Crude Oil Refining

---

1. *Natural gas*. Natural gas is composed mainly of CH<sub>4</sub> and some N<sub>2</sub>, depending on the source. It is separated from crude oil by natural sources. Its principal use is as a fuel gas.
2. *Liquified petroleum gas (LPG)*. LPG is composed of propane and butane. It is usually stripped from "wet" natural gas or from a crude oil cracking operation. LPG is widely used for industrial and domestic applications where natural gas is not available.
3. *Primary flash distillate (PFD)*. PFD is composed of propane and butane dissolved in the gasoline-kerosene range of liquids. It is produced by preliminary distillation of crude oil.
4. *Gasoline*. Gasoline has a boiling range of 300 to 450 K (80 to 350°F). It is produced by primary distillation and reforming processes to improve its performance. Its principal use is

spark-ignition internal combustion engines.

5. *Kerosene*. Kerosene is a paraffinic hydrocarbon with a boiling range of 410 to 575 K ( 280 to 575°F ). It is produced by distillation and cracking of crude oil. Its principal uses are in agricultural tractors, lighting, heating, and aviation gas turbines.
6. *Gas oil*. Gas oil is a saturated hydrocarbon with a boiling range of 450 to 620 K ( 350 to 660°F ). It is composed of saturated hydrocarbons produced by distillation and hydrodesulfurization of crude oil. Its principal uses are in diesel fuel, heating, and furnaces and as a feed to cracking units.
7. *Diesel fuel*. **Diesel fuel** is a saturated hydrocarbon with a boiling range of 450 to 650 K (350 to 680°F ). It is composed of saturated hydrocarbons produced by distillation cracking of crude oil. Its principal uses are in diesel engines and furnace heating. The ASTM grades for diesel fuel are as follows:
  - Grade 1-D. A volatile distillate fuel. Suitable for engines in service requiring frequent speed and load changes.
  - Grade 2-D. A distillate fuel oil of lower volatility. Suitable for engines in industrial and heavy mobile service.
  - Grade 4-D. A fuel oil for low- and medium-speed engines.
8. *Fuel oil*. Fuel oil has a boiling range of 500 to 700 K (440 to 800°F ). It is composed of residue of primary distillation, blended with distillates. Its primary use is in large-scale industrial heating.
9. *Lubricating oil*. There are three types of lubricating oils—mainly aromatic, mainly aliphatic, and mixed—which are manufactured by vacuum distillation of primary distillation residue-solvent extraction.
10. *Wax*. Wax is composed of paraffins produced by chilling residue of vacuum distillation. It is used primarily in toilet preparation, food, candles, and petroleum jelly.
11. *Bitumen*. Bitumen is produced from residue of vacuum distillation or by oxidation of residue from primary distillation. It is used for road surfacing and waterproofing.

## Nomenclature

<i>A</i>	Percentage of ash
API	American Petroleum Institute
Btu	Coal heating value, Btu per pound on moist basis
<i>C</i>	Mass fraction of carbon
FC	Percentage of fixed carbon
H <sub>2</sub>	Hydrogen
HHV	Higher heating value
<i>M</i>	Percentage of moisture
N <sub>2</sub>	Nitrogen
O <sub>2</sub>	Oxygen



<i>S</i>	Percentage of sulfur
<i>s</i>	Specific gravity
VM	Percentage of volatile matter

## Defining Terms

**Bituminous coal:** Soft coal containing large amounts of carbon. It has a luminous flame and a great deal of smoke.

**Catalytic cracking:** A refinery process that converts a high-boiling fraction of petroleum (gas oil) to gasoline, olefin feed for alkylation, distillate, fuel oil, and fuel gas by use of a catalyst and heat.

**Diesel fuel:** Fuel for diesel engines obtained from the distillation of crude oil. Its quality is measured by cetane number.

**Gasoline:** Light petroleum product obtained by refining crude oil with or without additives, blended to form a fuel suitable for use in spark-ignition engines. Its quality is measured by octane rating. The four classes are leaded regular (87–90 octane), unleaded regular (85–88 octane), mid-grade unleaded (88–90 octane), and premium (greater than 90 octane).

**Kerosene:** Colorless low-sulfur oil products used in space heaters, cook stoves, and water heaters.

**Octane:** A rating scale used to grade gasoline according to its antiknock properties. Also, any of several isometric liquid paraffin hydrocarbons,  $C_8H_{18}$ . Normal octane is a colorless liquid found in petroleum boiling at  $124.6^\circ C$ .

## References

- Bobcock and Wilcox Company. 1972. *Steam—Its Generation and Use*, 38th ed. Bobcock and Wilcox Company, New York.
- Francis, W. 1965. *Fuels and Fuel Technology*. Pergamon Press, New York.
- Harker, J. H. and Allen, D. A. 1972. *Fuel Science*. Oliver & Boyd, Edinburgh, UK.
- Johnson, A. J. and Auth, G. A. 1951. *Fuel and Combustion Handbook*. McGraw-Hill, New York.
- Popovich, M. and Hering, C. 1959. *Fuels and Lubricants*. John Wiley & Sons, New York.

## Further Information

American Gas Association

American Petroleum Institute

National Petroleum Council

U.S. Department of the Interior publications *International Petroleum Annual* and *U.S. Petroleum and Natural Gas Resources*

Chavez, J. M., Richards, E. M., VanArsdall, A. "Solar Power Systems"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Solar Power Systems

---

65.1 Solar Thermal Systems

65.2 Photovoltaic Systems

65.3 Biomass Systems

**James M. Chavez**

*Sandia National Laboratories*

**Elizabeth M. Richards**

*Sandia National Laboratories*

**Anne VanArsdall**

*Sandia National Laboratories*

Solar power systems are usually classified by technology—solar thermal and photovoltaic systems are the principal types. Photovoltaic systems use the energy in sunlight directly to produce electricity; in solar thermal power systems, the sun heats a transfer medium, such as oil, which is used to generate steam to drive a turbine for electricity. Biomass power technology is included in this discussion because the leaves of plants are natural solar collectors; combustion of biomass—produced by using or storing solar energy—produces heat for power.

Experimentation with solar systems goes back into the inventors' labs of the past century. However, in the U.S., serious research into solar power systems only began in the early 1970s. At that time, the Department of Energy, its national laboratories, and many industrial firms with which they contracted began to explore solar energy as an alternative to fossil fuels because of a political embargo on oil. The advantage seen at the time was energy security; the added benefit, now a primary one, was solar energy's lack of environmental pollution.

From this research and development, commercially attractive photovoltaic, solar

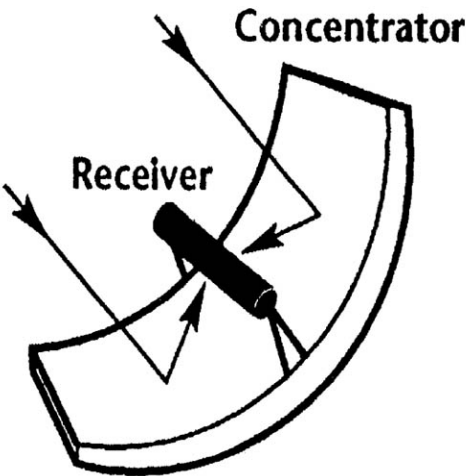
thermal, and biomass systems have emerged. Each has distinct characteristics and advantages, and all depend on and are affected by the sunlight and/or natural resources of an area. The systems discussed here represent the state of the art in solar systems today; they include *parabolic trough systems*, *central receiver "power towers,"* *concentrating dish systems*, *stand-alone photovoltaic systems*, and *biomass conversion power plants*. The major potential for solar power systems, including biomass, is as a supplement to utility power and as a provider of electricity in areas far from the grid. The power such systems can produce ranges from about 1 watt to hundreds of megawatts—from power for a water pump, to power for a village, to power for cities.

### 65.1 Solar Thermal Systems

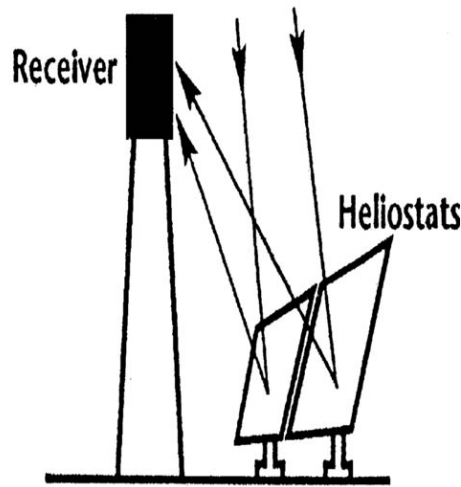
The schematic that follows illustrates the concepts behind the three major types of solar thermal systems. (See [Figs. 65.1](#), [65.2](#), and [65.3](#).)

	Trough	Power Tower	Dish
Solar concentration	80 suns	800 suns	3000 suns
Operating temperature	350°C	560°C	800°C
Annual efficiency	10 to 14%	15 to 20%	24 to 28%

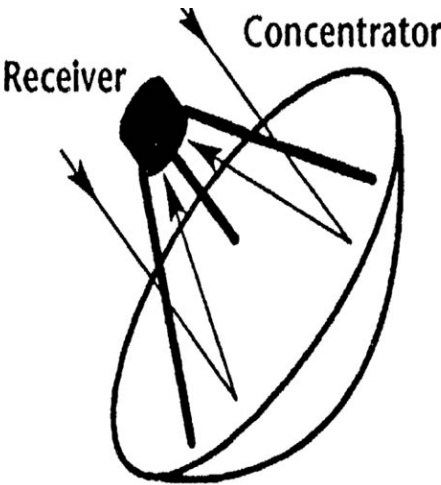
**Figure 65.1** Trough system.  
(Courtesy of Sandia National Laboratories.)

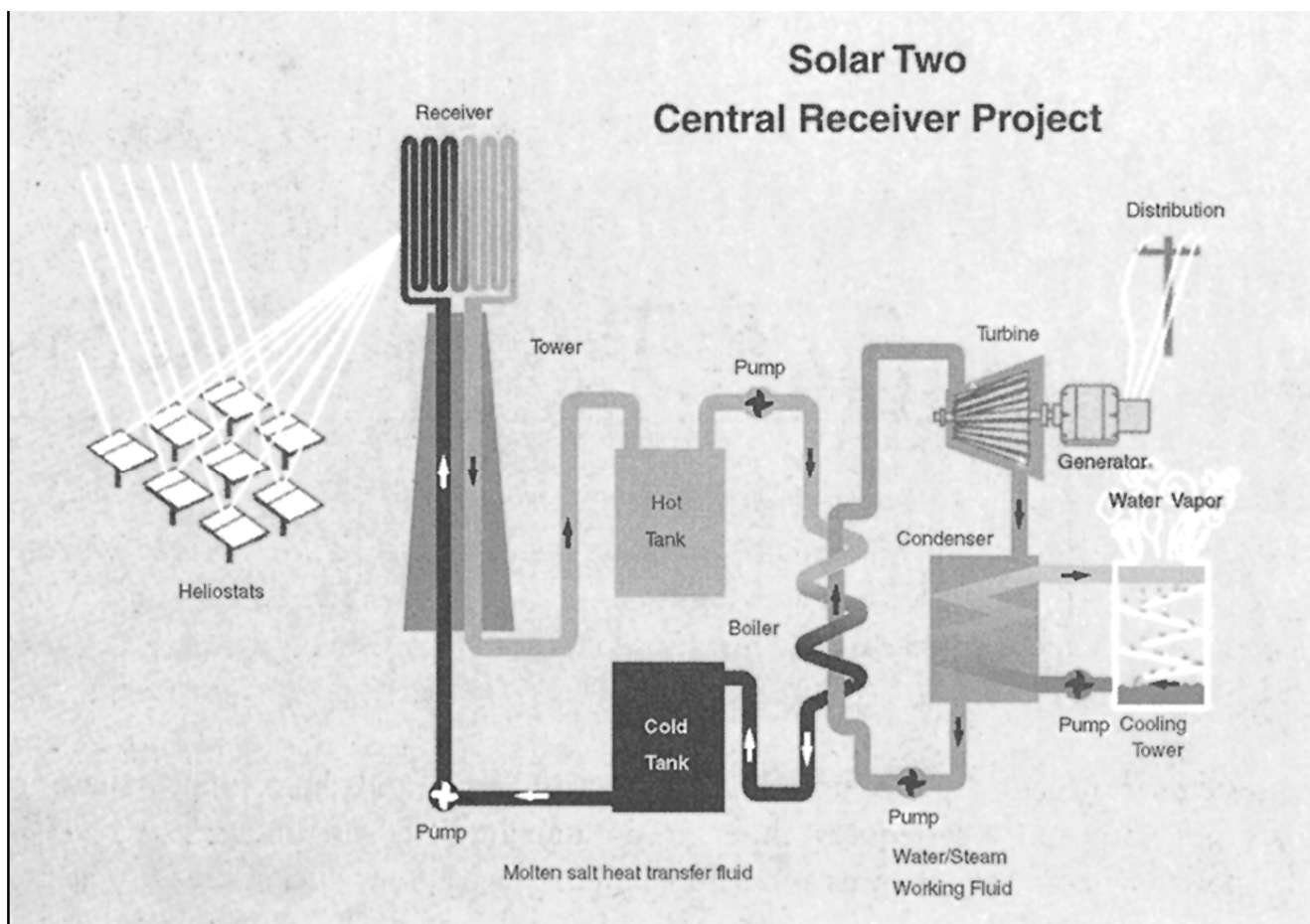


**Figure 65.2** Central receiver.  
(Courtesy of Sandia National Laboratories.)



**Figure 65.3** Dish system.  
(Courtesy of Sandia National Laboratories.)





### CENTRAL RECEIVER DESIGN

Solar One, the test pilot power plant which preceded Solar Two by more than ten years, as well as solar plants in France and Spain, validated the technical feasibility of the central receiver design. The technology has little environmental impact and is widely accepted by the public. Because it is not encumbered by emission-control and fuel costs, which are subject to fluctuation, a central receiver plant's expenses are fairly stable.

The central receiver system of Solar Two is made up of several specially designed subsystems. The most important subsystems are the heliostats (mirrors), the receiver, and the thermal storage. The diagram above illustrates a field of heliostats that are positioned to concentrate sunlight onto a central receiver on top of a 300-foot-tall tower. "Cold" molten salt (550°F) from an insulated storage tank is pumped to the top of the tower and heated to 1050°F by the concentrated sunlight. The hot salt then flows to the insulated storage tank. When electric power is needed, the hot liquid salt is pumped to a conventional steam-generating system to produce superheated steam, which drives a turbine/generator. The salt, now cooled back down to 550°F, returns from the steam

generator to the "cold" tank to be reused. Efficient storage is the unique advantage of this system: electric power production can occur at times other than when solar energy is being collected. About 36% of the solar energy absorbed by the central receiver is dispatched as electricity.

It is anticipated that the additional feasibility testing by the Solar Two central receiver project will lead to a group of commercial solar power plants based on the same technology. There are no major technological impediments to scaling the design up to a 100–200 MW power plant. Enough information has been gathered, and construction of such plants may begin as early as 1998. (Courtesy of Southern California Edison.)



Shown here are just a few of the more than 1900 heliostats used in the Solar Two project. Each heliostat consists of 12 slightly concave mirrored surfaces. The angle of each heliostat is adjusted by a computer to reflect the most sunlight possible to the stationary central receiver. (Courtesy of Southern California Edison. Photo: Southern California Edison.)

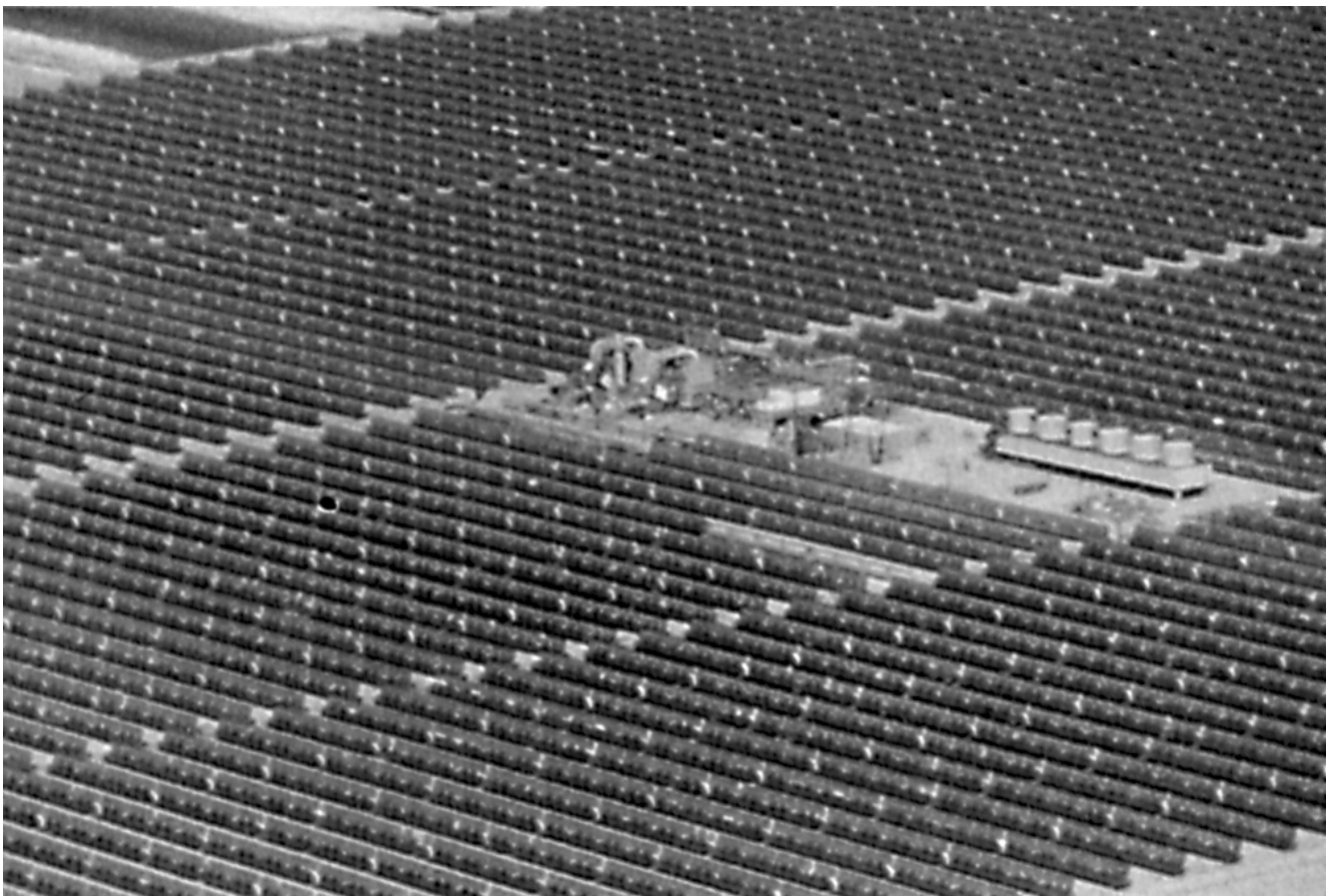
*Trough-electric systems* use parabolic trough concentrators to focus the sunlight on a glass-encapsulated tube that runs along the focal line of the collector, as shown in [Fig. 65.1](#). The troughs are located so that they typically track the sun from east to west to keep the sun aligned on the receiver tube. The solar concentration and temperature of



the working fluid, usually an oil, are relatively low in such systems. The oil (or other fluid) is heated by concentrated sunlight to 300 to 400°C in the receiver before going to a heat exchanger, which transfers the heat to water. The water is boiled and used to drive a conventional Rankine cycle turbine generator. The working fluid is pumped through the entire plant, and the electricity is produced at a single power block. The optimum size for such a system is about 200 megawatts-electric.

Nine trough solar electric generating systems, called *SEGS*, are currently operating in southern California (see [Fig. 65.4](#)); they deliver 354 megawatts-electric to Southern California Edison's power grid. These systems were installed between 1985 and 1991—each larger than the one before it. The larger plants generate progressively higher solar field temperatures, have better economies of scale and greater conversion efficiencies, and have demonstrated that they can meet and, in some cases, even exceed projected outputs.

**Figure 65.4** Solar thermal SEGS plant. (Courtesy of Sandia National Laboratories.)



A major challenge facing these plants is to reduce the expense for operating and maintenance, which represents about a quarter of the cost of the electricity they produce. Operation and maintenance are expected to be similar for all three types of solar thermal systems. With each newer solar electric generating system, performance has improved; any design changes have been prompted by the desire for more cost-effective energy production. The systems have also increased in size—SEGS I is 14 megawatts-electric; SEGS IX is 80. The newer plants are designed to operate at 10 to 14% annual efficiency and to produce electricity for \$.08 to \$.14 a kilowatt-hour, depending on interest rates and tax incentives. Maintenance costs are estimated to be about \$.02 a kilowatt-hour.

*Central receiver power towers* are not as commercially mature as trough systems. (See [Fig. 65.5](#).) In a power tower configuration, two-axis tracking mirrors, called *heliostats*, reflect solar energy onto a receiver that is mounted on top of a tower in the center of the field of heliostats. To keep sunlight concentrated on the receiver at all times, each heliostat tracks a position in the sky midway between the receiver and the sun. A number of different working fluids have been evaluated for power towers including water/steam, sodium nitrate salts, and air. However, in the U.S., modern power towers decouple solar collection from power generation by using a nitrate salt as the working fluid, allowing the solar energy to be collected when the sun shines, storing it, and, thus, being able to produce power on demand with a conventional turbine-generator. This feature increases the operating time (capacity factor) of the system.

The advantage of using salt instead of water—one of the possible working fluids used in the past—is that it is a liquid at the operating temperature of the turbine-generator used in the power block. The process is for cold salt (290°C) to be pumped out of a cold storage tank and through the thermal receiver, where solar flux at 800 suns' intensity heats the salt. The heated salt (about 560°C) is then delivered to a hot salt storage tank, and electricity is produced by removing the hot salt from the storage tank, passing it through a steam generator, and delivering steam to a turbine-generator. The salt, which has cooled down, is returned to the cold storage tank. Power towers with nitrate salts constitute the only solar thermal technology in which thermal storage makes sense economically.

The optimal size for power towers is in the 100- to 300-megawatt range. Studies have estimated that such solar systems can produce electrical power at a cost of \$.06 to \$.11 per kilowatt-hour. The current demonstration plant is Solar Two, a power tower system in southern California that improves on an earlier plant. Solar One used water as the heat transfer medium and experienced intermittent operation because of the effect of cloud transients and lack of thermal storage.

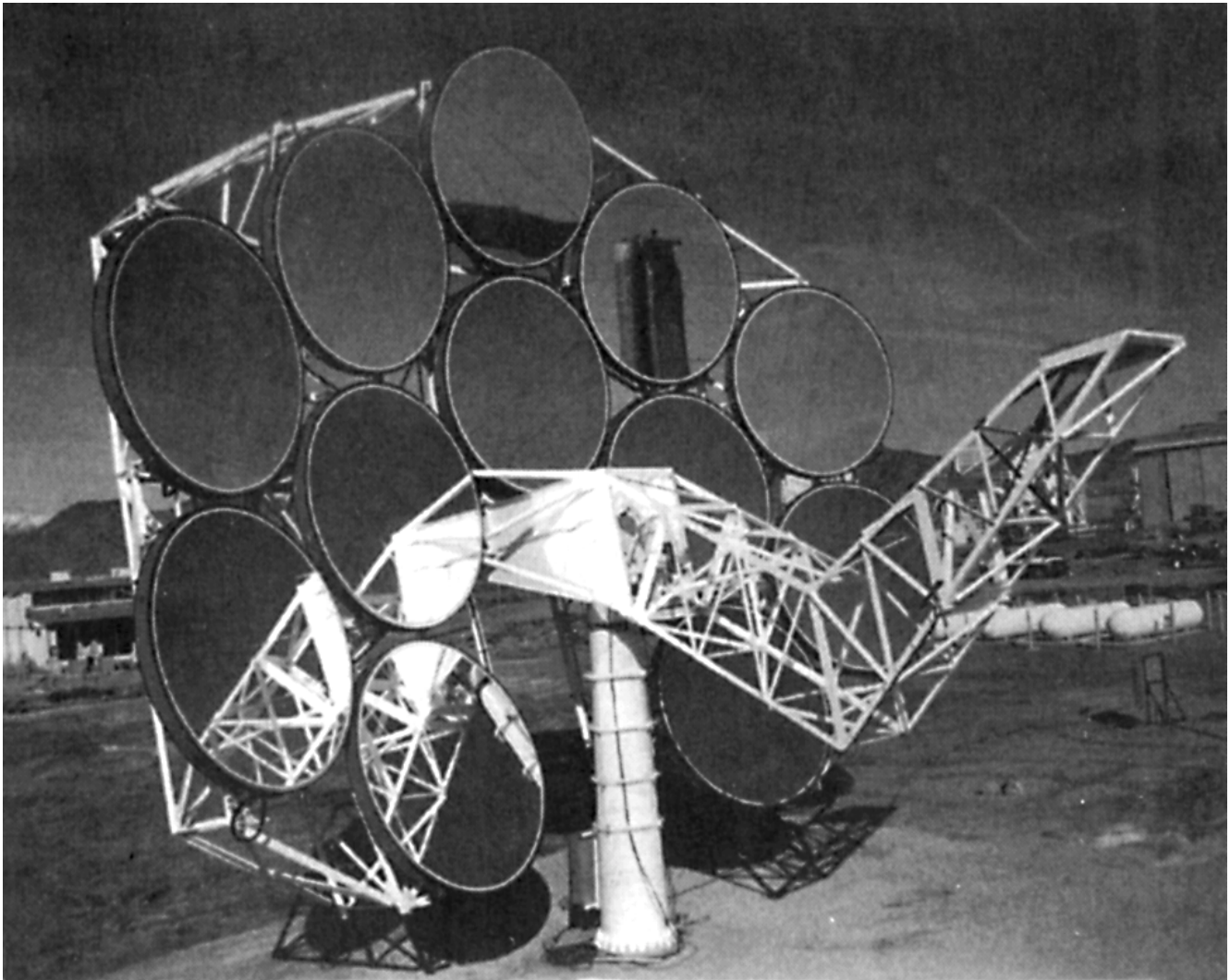
*Dish/Stirling systems* are composed of a mirrored parabolic concentrator or dish, a thermal receiver, and a heat engine/generator ([Fig. 65.6](#)). The system tracks the sun and reflects the solar energy onto the focal point of the dish, where the receiver absorbs it. The absorbed heat is then transferred to the heater head of an engine/generator that is externally fired. Stirling engines are most often used in these systems.



**Figure 65.5** Solar One was the first central receiver to be demonstrated. (Courtesy of Sandia National Laboratories.)



**Figure 65.6** Dish/Stirling system. (Courtesy of Sandia National Laboratories.)



The major issues for developers are the reliability of operating a receiver at high-temperature and high-heat flux conditions, the lifetime, and efficiency over time of Stirling engines. Heat fluxes of 50 to 75 watts per square centimeter and temperatures of 700 to 800°C are common in thermal receivers, which are typically either direct illumination, heat pipes, or pool boilers. The types of Stirling engines being compared in these systems are kinematic and free piston.

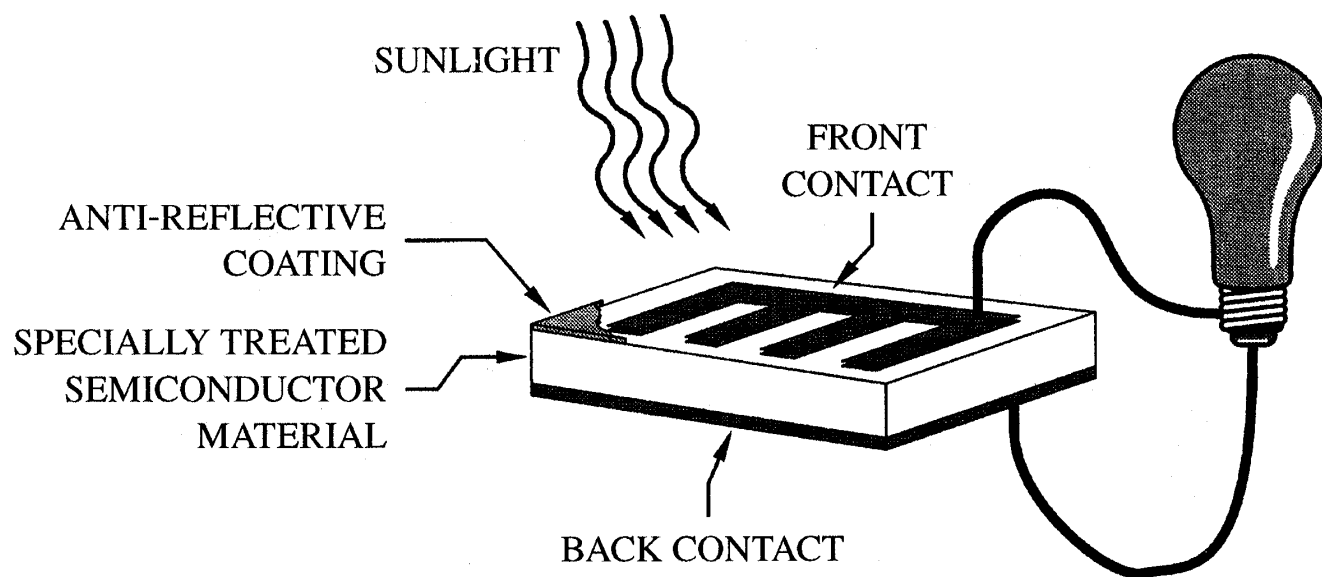
The best size for dish/Stirling systems is about 25 kilowatts-electric, primarily because of the available engines and the wind loading on the concentrator. Dish systems are modular and can be combined into power plants to produce from a few hundred kilowatts to about 50 megawatts-electric. Electrical power from these systems, when combined to produce utility-scale power, is projected to be \$.06 to \$.11 per kilowatt-hour with mass production.

## 65.2 Photovoltaic Systems

Photovoltaics is the direct conversion of light into electricity, an effect observed in the early 1800s but not seriously considered as a usable technology until the 1950s. The first solar cells were developed for the space effort in the U.S.; there were few of them and they were very expensive. Photovoltaics entered the U.S. laboratories for development as a possible fuel source at the same time as solar thermal technologies—early in the 1970s.

Photovoltaic solar cells are made of semiconductor materials (usually silicon) that are pressed into a thin wafer specially treated to form an electric field—positive on one side and negative on the other. When light energy strikes the cell, electrons are knocked loose from the atoms in the semiconductor material. The electrons can be captured as electric current if electrical conductors are attached to the positive and negative sides. The electricity can be used to power a load, such as a light bulb or water pump. A typical four-inch silicon solar cell produces about one-and-a-half watts of electricity in bright noon sunshine. See [Fig. 65.7](#).

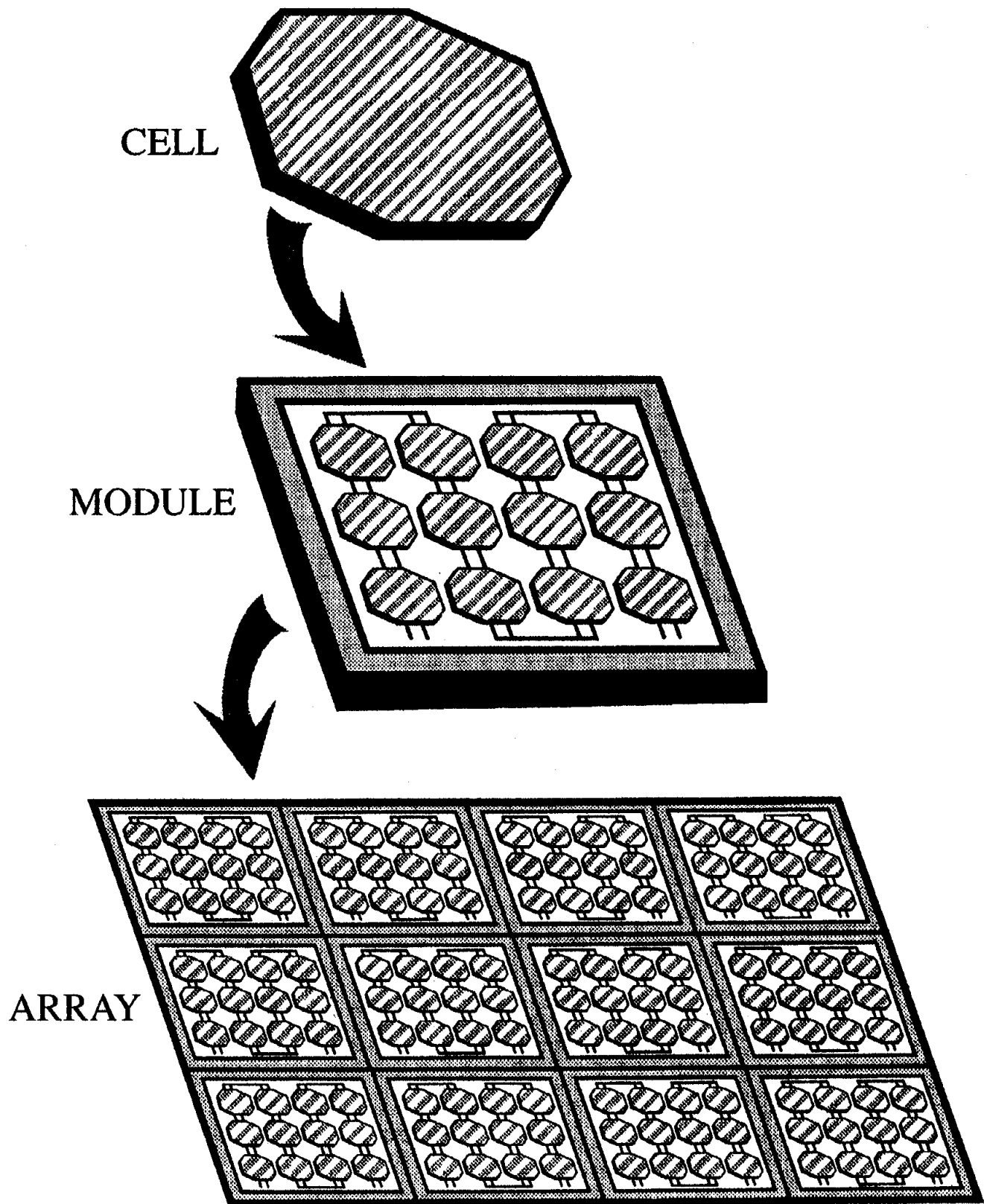
**Figure 65.7** Photovoltaic cell. (Courtesy of Sandia National Laboratories.)



A photovoltaic system consists of solar cells electrically connected to each other in a support structure to form a module. Modules are designed to supply electricity at a certain voltage, commonly 12 volts. The current produced is directly dependent on how much light strikes the module. Two or more modules can be wired together to form an array, as shown in [Fig. 65.8](#). In general, the larger the area of a module or array is, the more electricity will be produced. Photovoltaic modules and arrays produce DC electricity. They can be connected in both series and parallel to produce any required voltage and current combinations.



**Figure 65.8** Photovoltaic system. (Courtesy of Sandia National Laboratories.)



A photovoltaic system involves an array or arrays combined with a number of other components, collectively known as the *balance of system*. These components vary according to the types of service required and whether the system is needed only during hours of sunlight or also at night. Some form of energy storage is needed for the systems to operate at night.

Photovoltaics is a well-established, proven solar technology with a substantial international industrial network. These systems power everything from individual light bulbs to entire villages in every region of the world; their modular construction facilitates expansion of the systems as needed. These systems operate with little servicing and no refueling, they produce no emissions to harm the environment, and they are silent. See [Figs. 65.9](#) and [65.10](#).

**Figure 65.9** A small photovoltaic system. (Courtesy of Sandia National Laboratories.)



**Figure 65.10** A photovoltaic system used in disaster relief. (Courtesy of Sandia National Laboratories.)



Some concerns about photovoltaics involve its relatively high initial cost, the need to have an inverter if AC power is required, and the need to purchase and maintain batteries in the systems if power is needed at night. Photovoltaic systems are sometimes used in a hybrid configuration with one or more power sources that are not sun dependent, such as a wind-powered or diesel generator.

### **65.3 Biomass Systems**

---

The biomass fuel known most commonly is simply firewood, wood from trees that use sunlight to grow. This common "solar" source of energy is everywhere and makes up some 15% of the world's energy—much greater in developing countries, where it can reach 27%. Plant matter and its derivatives, such as dung, likewise constitute

biomass.

Burning biomass—such substances as sawdust, rice hulls, organic wastes—in a boiler to generate electricity is understood in principle. But the challenge for modern research is to find a way of efficiently preparing and then burning these products, because boiler systems may not run as well on biomass fuels as on coal. In addition, pollution control in burning biomass is one of the challenges facing researchers in the field. A secondary but equally important issue involves collection of the biomass products to burn and transportation of these products to the biomass plant.

A major possible source for biomass fuels is municipal refuse. Here, separating out combustible materials from other wastes poses a problem, and several demonstration plants across the country are experimenting with different ways to accomplish the sorting. As early as the 1970s, it was estimated that 2 quads of energy could be produced from the 200 millions tons of organic municipal wastes in the U.S.

Agricultural wastes are another large source of biomass fuel. In a given year, field crops constitute 400 million tons of waste, which would produce 4 quads of energy. In a similar vein, residues of lumber mills can be compacted and used much like coal in boiler plants. And forests that must be cleared can provide wood to be chipped and processed for burning. Lands that are waste or depleted can be planted with crops suitable for processing into biomass for energy. An obvious drawback of biomass compared to coal is that its mass density is much less—the plant matter has moisture that has to dry out and it is less compact than coal. However, its ash content is lower and it has many fewer toxic chemicals than coal. In fact, the ash that is recovered from a biomass plant can be used for fertilizer.

A strategy currently being considered involves producing electricity—or producing electricity and heat at the same time—using gasified biomass with advanced conversion technologies.

## References

- Hill, R. 1994. Commercializing photovoltaic technology. *Mech. Eng.* 116(8):80–83.
- Kitani, O. and Hall, C. (Eds.) 1989. *The Biomass Handbook*. Gordon and Breach. Newark, NJ.
- Mancini, T. R., Chavez, J. M., and Kolb, G. J. 1994. Solar thermal power today and tomorrow. *Mech. Eng.* 116(8):74–79
- Schurr, S. H. (Ed.) 1979. *Energy in America's Future—The Choice Before Us*. Johns Hopkins University Press, Baltimore, MD.
- Sheppard L. and Richards, E. 1993. *Solar Photovoltaics for Development*



*Applications*. Sandia National Laboratories, SAND93-1642, Albuquerque, NM.

U.S. Department of Energy, Solar Thermal and Biomass Power Division. 1993. *Solar Thermal Electric Five Year Plan*. U.S. Department of Energy, Washington, DC.

## Further Information

Baxter, L. 1992. Ash deposition during biomass and coal combustion: A mechanistic approach. *Biomass and Bioenergy*. IV:85–102.

Falcone, P. K. 1986. *A Handbook for Solar Central Receiver Design*. Sandia National Laboratories, SAND86-8009, Albuquerque, NM.

Hustad, J. and Sønju, O. 1991. Biomass combustion in international energy agency countries. *Biomass and Bioenergy*. II:239–261.

Johansson, T. B., Kelly, H., Reddy, A., and Williams, R. H. 1993. *Renewable Energy, Sources for Fuels and Electricity*. Island Press, Washington, DC.

Lotker, M. 1991. *Barriers to Commercialization of Large-Scale Solar Electricity: Lessons Learned from the LUZ Experience*. Sandia National Laboratories, SAND91-7014, Albuquerque, NM.

Sandia Photovoltaic Design Assistance Center. 1991. *Stand-Alone Photovoltaic Systems: A Handbook of Recommended Design Practices*. Sandia National Laboratories, SAND87-7023, Albuquerque, NM.

Stevens, J. W., Thomas, M. G., Post, H. N., and VanArsdall, A. 1993. *Photovoltaic Systems for Utilities*. Sandia National Laboratories, SAND90-1378, Albuquerque, NM.

Stine, W. B. and Diver, R. B. 1994. *A Compendium of Dish Stirling Technology*. Sandia National Laboratories, SAND93-7026, Albuquerque, NM.

Thomas, M. G., Post, H. N., and VanArsdall, A. 1994. *Photovoltaics Now*, rev. ed. Sandia National Laboratories, SAND88-3149, Albuquerque, NM.

U.S. Department of Energy, Solar Thermal and Biomass Power Division. 1992. *Solar Thermal Electric and Biomass Power Program Overviews*. FY 1990–1991.



Kornhauser, A. A. "Internal Combustion Engines"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Internal Combustion Engines

---

## 66.1 Basics of Operation

## 66.2 Engine Classifications

## 66.3 Spark Ignition Engines

Idealized and Actual Cycles • Combustion, Fuels, and Emissions • Control • Advantages

## 66.4 Compression Ignition Engines

Idealized and Actual Cycles • Combustion, Fuels, and Emissions • Control • Advantages

## 66.5 Gas Exchange Systems

4-Stroke Intake and Exhaust • 2-Stroke Scavenging • Supercharging and Turbocharging

## 66.6 Design Details

Engine Arrangements • Valve Gear • Lubrication • Cooling

## 66.7 Design and Performance Data for Typical Engines

### **Alan A. Kornhauser**

*Virginia Polytechnic Institute and State University*

An internal combustion (i.c.) engine is a heat engine in which the thermal energy comes from a chemical reaction within the working fluid. In external combustion engines, such as steam engines, heat is transferred to the working fluid through a solid wall and rejected to the environment through another solid wall. In i.c. engines, heat is released by a chemical reaction in the working fluid and rejected by exhausting the working fluid to the environment.

Internal combustion engines have two intrinsic advantages over other engine types:

1. They require no heat exchangers (except for auxiliary cooling). Thus, weight, volume, cost, and complexity are reduced.
2. They require no high temperature heat transfer through walls. Thus, the maximum temperature of the working fluid can exceed maximum allowable wall material temperature.

They also have some intrinsic disadvantages:

1. Practically, working fluids are limited to air and products of combustion.
2. Nonfuel heat sources (waste heat, solar, nuclear) cannot be used.
3. There is little flexibility in combustion conditions because they are largely set by engine requirements. This can make low-emissions combustion hard to attain.

The advantages far outweigh the disadvantages. I.c. engines comprise more individual units and more rated power than all other types of heat engines combined.

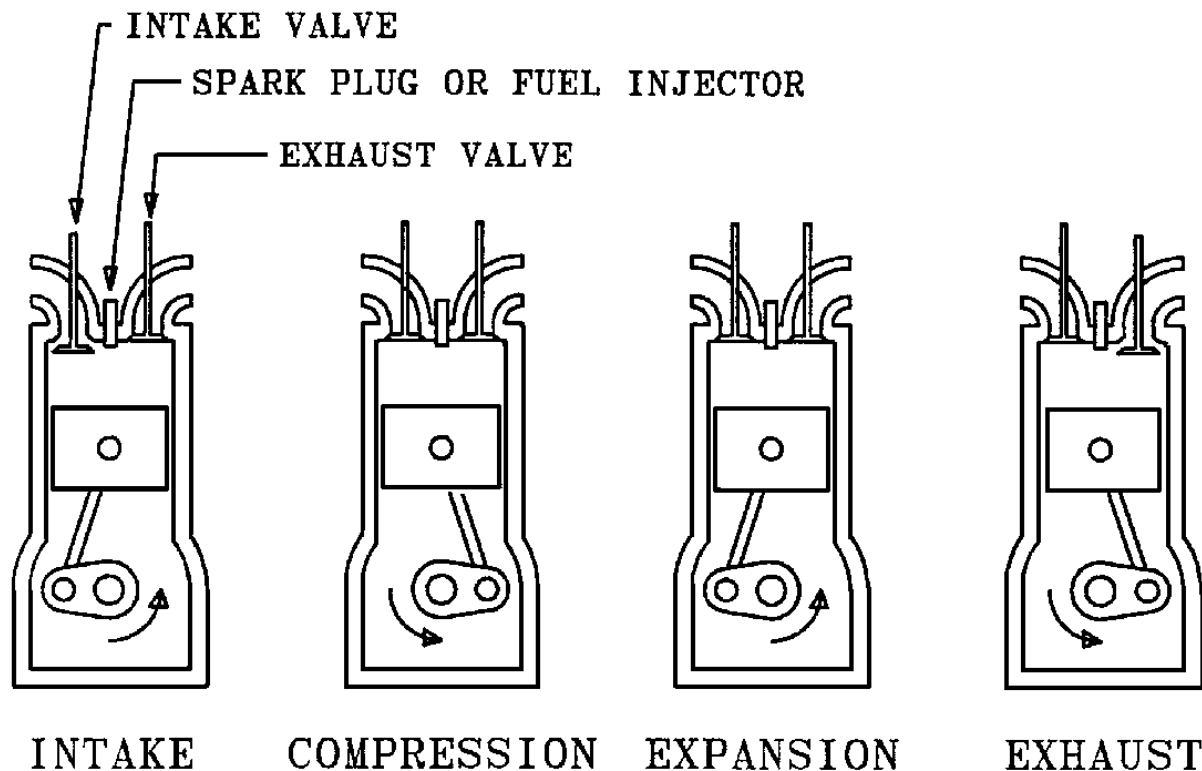
According to the definition given above, i.c. engines include reciprocating types, rotary (Wankel) types, and gas turbines. In customary usage, however, the title "internal combustion" is used only for the first two of these three types. A more proper designation might be "positive displacement internal combustion" engines. These are the engines described in this chapter.

## 66.1 Basics of Operation

The basic operation of an i.c. engine is shown in [Fig. 66.1](#). The typical engine cycle is divided into four steps:

1. *Intake*. Engine working volume increases. Intake valve opens to admit air or air/fuel mixture into the working volume.
2. *Compression*. Engine working volume decreases. Valves are closed, and the air or mixture is compressed. Work is done on the working fluid.
3. *Combustion and expansion*. Air/fuel mixture burns and releases chemical energy. If fuel was not admitted previously, it is injected at this point. Pressure and temperature inside the working volume increase dramatically. Working volume increases, and work (much greater than that of compression) is done by the working fluid.
4. *Exhaust*. Engine working volume decreases. Exhaust valve opens to expel combustion products from the working volume.

**Figure 66.1** Operating cycle for a 4-stroke i.c. engine.



The engine shown is a 4-stroke reciprocating type; details would vary for 2-stroke or rotary engines.

## 66.2 Engine Classifications

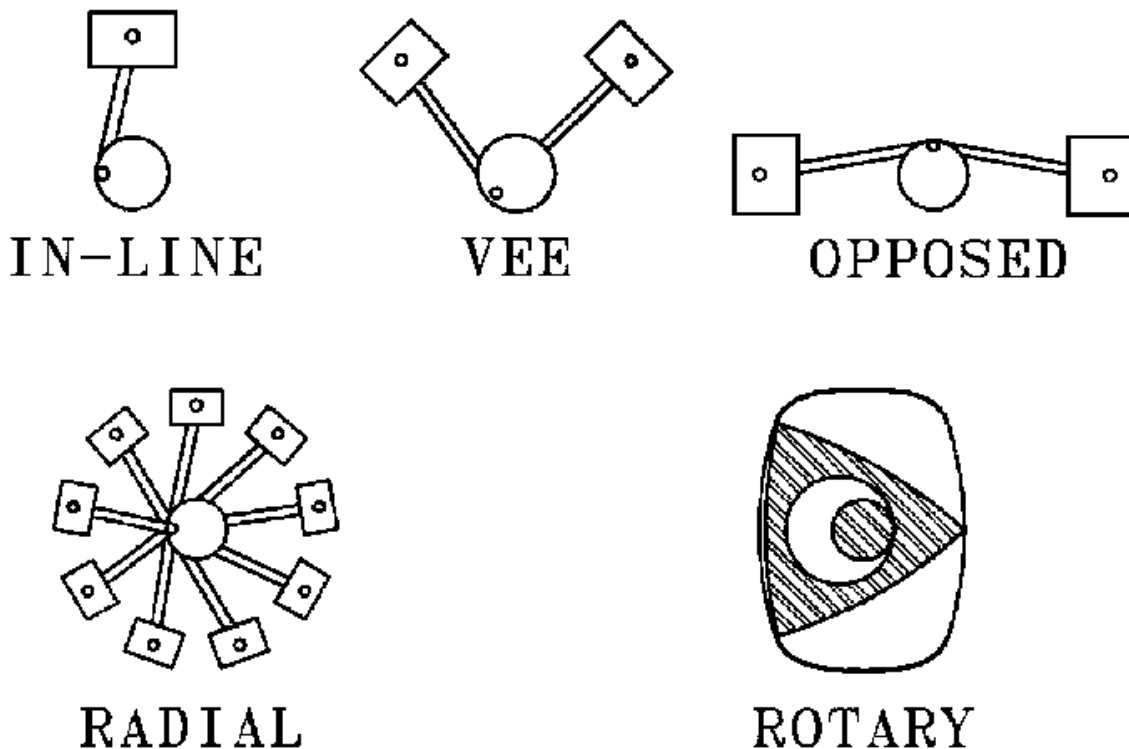
---

I.c. engines can be classified in various ways. Some important classifications are:

*Spark ignition/compression ignition.* In **spark ignition** (s.i., gasoline, petrol, or Otto) engines, the fuel is either mixed with the air prior to the intake stroke or shortly after inlet valve closure. An electric spark ignites the mixture. In **compression ignition** (c.i., oil, or diesel) engines, the fuel is injected after the compression process. The high temperature of the compressed gas causes ignition.

*4-stroke/2-stroke.* In **4-stroke** engines, the working cycle is as shown in [Fig. 66.1](#). A complete 4-stroke cycle takes two crankshaft revolutions, with each stage (intake, compression, expansion, exhaust) comprising about 180°. A complete 2-stroke cycle takes only one crankshaft revolution. In a **2-stroke** engine intake and exhaust strokes are eliminated: gas exchange occurs when the piston is near bottom center position between the expansion and compression strokes. Because the piston does not provide pumping action, some external device is required to ensure that fresh air or mixture replaces the combustion products.

**Figure 66.2** Engine arrangements.



*Mechanical layout.* Various mechanical layouts are shown in Fig. 66.2. Reciprocating i.c. engines use multiple piston-cylinder arrangements driving a single crankshaft. The total number of cylinders per engine ranges from 1 to 20 or more, with 1, 4, 6, and 8 the most common. The cylinders can be arranged in line, in a vee, radially, or horizontally opposed. Rotary i.c. engines use an approximately triangular rotor which revolves eccentrically in a lobed stator. The spaces between the rotor and the stator go through essentially the same processes shown in Fig. 66.1. A single rotor-stator pair is thus equivalent to three cylinders. Additional rotor-stator pairs can be stacked on a single shaft to form larger engines.

*Intake system.* In **naturally aspirated** engines, the pumping action of the piston face draws air into the cylinder. In crankcase **scavenged** engines, the pumping action of the back side of the piston in the crankcase forces air into the cylinder. In **supercharged** engines, a compressor, typically driven off the crankshaft, forces air into the cylinder. In **turbocharged** engines, the compressor is driven by a turbine which recovers work from the exhaust gas.

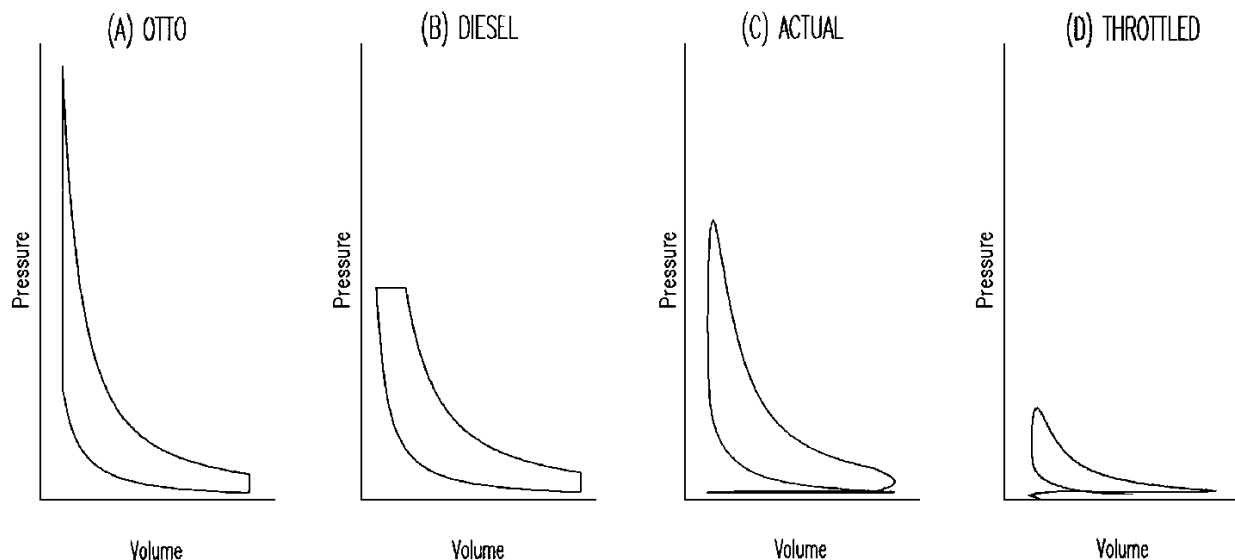
Besides the major classifications above, engines can be classified by valve number and design (2, 3, or 4 valves per cylinder; rocker arm or overhead cam; cross-, loop-, or uniflow-scavenged), by fuel addition method (carbureted, fuel injected), by combustion chamber shape (tee, ell, flat, wedge, hemisphere, bowl-in-piston), and by cylinder wall cooling method (air, water).

## 66.3 Spark Ignition Engines

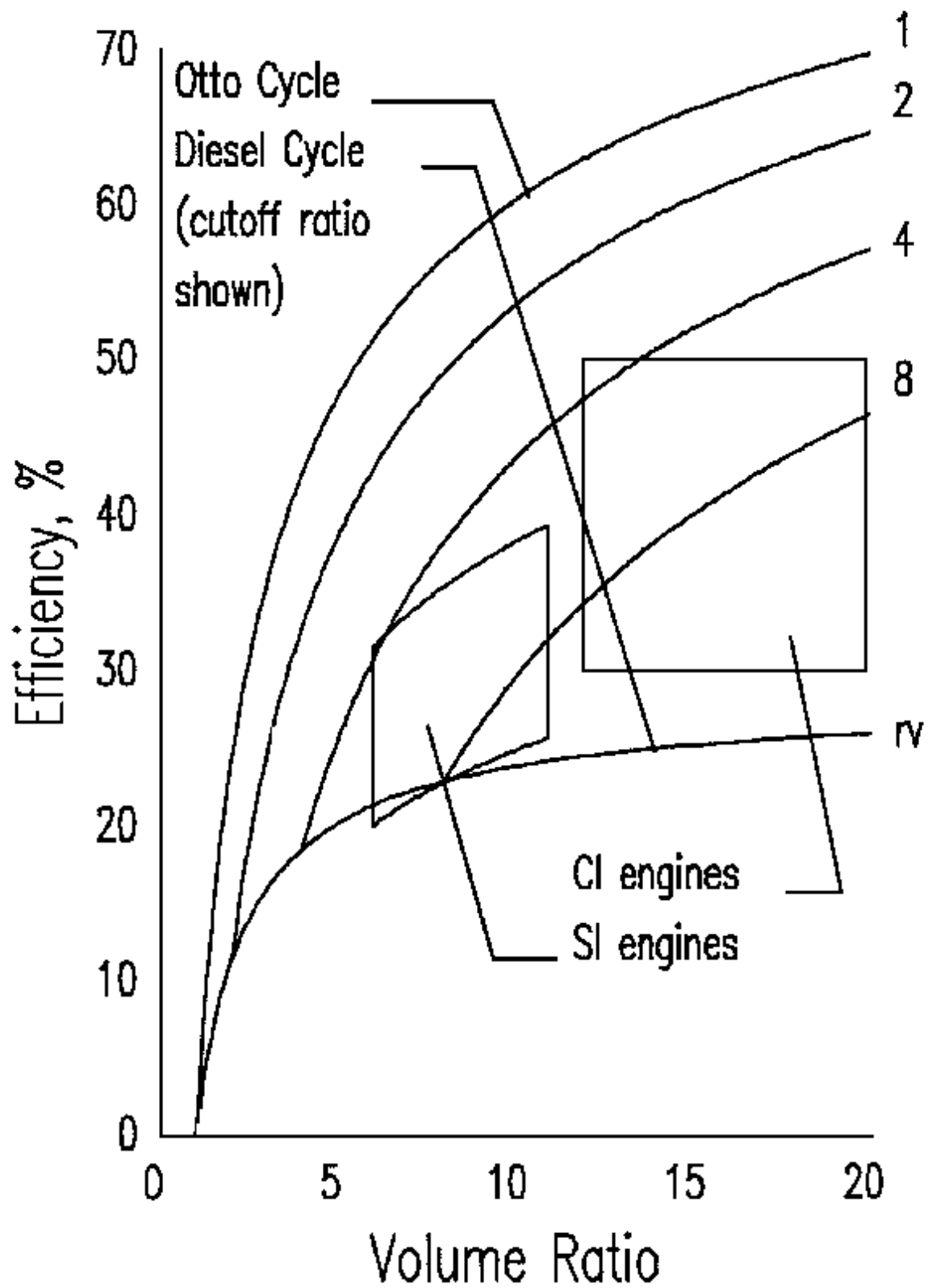
### Idealized and Actual Cycles

The spark ignition (s.i.) engine can be idealized as an **Otto** cycle using an ideal gas with constant specific heat [Fig. 66.3(a)]. The Otto cycle consists of isentropic compression, constant volume heating (simulating combustion), isentropic expansion, and constant volume cooling (simulating intake and exhaust). The **thermal efficiency** of an Otto cycle (Fig. 66.4) is  $\eta_t = 1 - r_v^{1-\gamma}$ , where  $r_v$  is the **compression ratio** and  $\gamma$  is the gas specific heat ratio.

**Figure 66.3** Pressure-volume diagrams for ideal and actual engine cycles.



**Figure 66.4** Efficiency of ideal cycles and actual engines.



The actual engine "cycle" [Fig. 66.3(c)] differs from the Otto cycle: (1) in that heat transfer occurs during compression and expansion; (2) in that combustion takes place gradually during compression and expansion rather than instantaneously; (3) in the presence of intake and exhaust processes; and (4) in the variation in gas composition and gas specific heat. For a given  $r_v$ , the efficiency of a typical s.i. engine is considerably lower than that of the ideal cycle (Fig. 66.4), and actual engine  $r_v$  is limited by combustion **knock**.

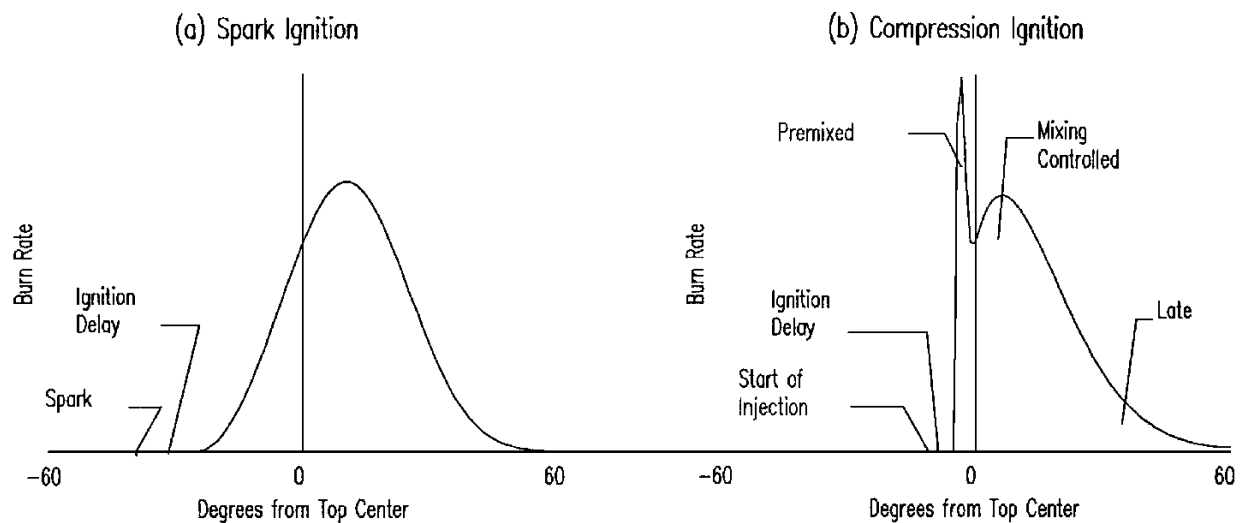
## Combustion, Fuels, and Emissions

In an s.i. engine, air and vaporized fuel are generally premixed before they enter the cylinder.

**Equivalence ratio** ( $\phi$ ) generally ranges from about 0.7 to 1.3, with lean mixtures (low  $\phi$ ) giving maximum efficiency and the rich mixtures (high  $\phi$ ) giving maximum power.

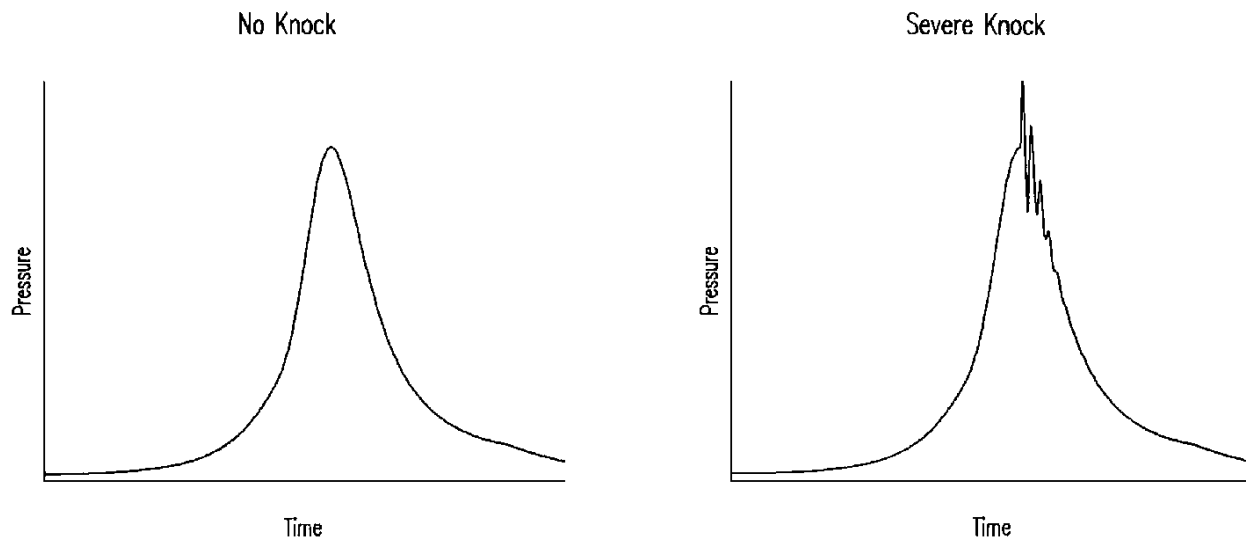
The mixture is heated by compression, but not enough to cause autoignition. Combustion is initiated by an electric spark. If the engine is operating properly, a turbulent flame front travels smoothly and rapidly across the cylinder space. It takes a 15–25° crank angle for the first 10% of the mixture to burn, while the next 85% is burned within an additional 35–60° [Fig. 66.5(a)]. The low numbers of these ranges correspond to  $\phi \approx 1$ , high turbulence combustion chambers, and low engine speeds. The high numbers correspond to rich or lean mixtures, low turbulence combustion chambers, and high engine speeds. To time the heat release optimally, the spark is typically discharged 5° to 40° before top center, with this advance automatically varied according to engine speed and load. Because higher engine speeds result in increased turbulence and, thus, in increased flame speeds, the total crank angle for combustion increases only slightly as engine speed changes.

**Figure 66.5** Heat release rates for s.i. and c.i. engines.



Under some operating conditions, the flame does not burn smoothly. In these cases the mixture ahead of the flame front is heated by compression and autoignites before the flame arrives. The resulting detonation wave causes an extremely rapid pressure rise (Fig. 66.6) which is noisy and can damage the engine. Because of the noise, the phenomenon is known as *knock*. Knock can be avoided by decreasing the pressure ratio, using more knock-resistant fuels, increasing flame speed, retarding the spark, and designing the combustion chamber to ensure that the last mixture burned is in the coolest part of the cylinder.

**Figure 66.6** Effect of knock on cylinder pressure.



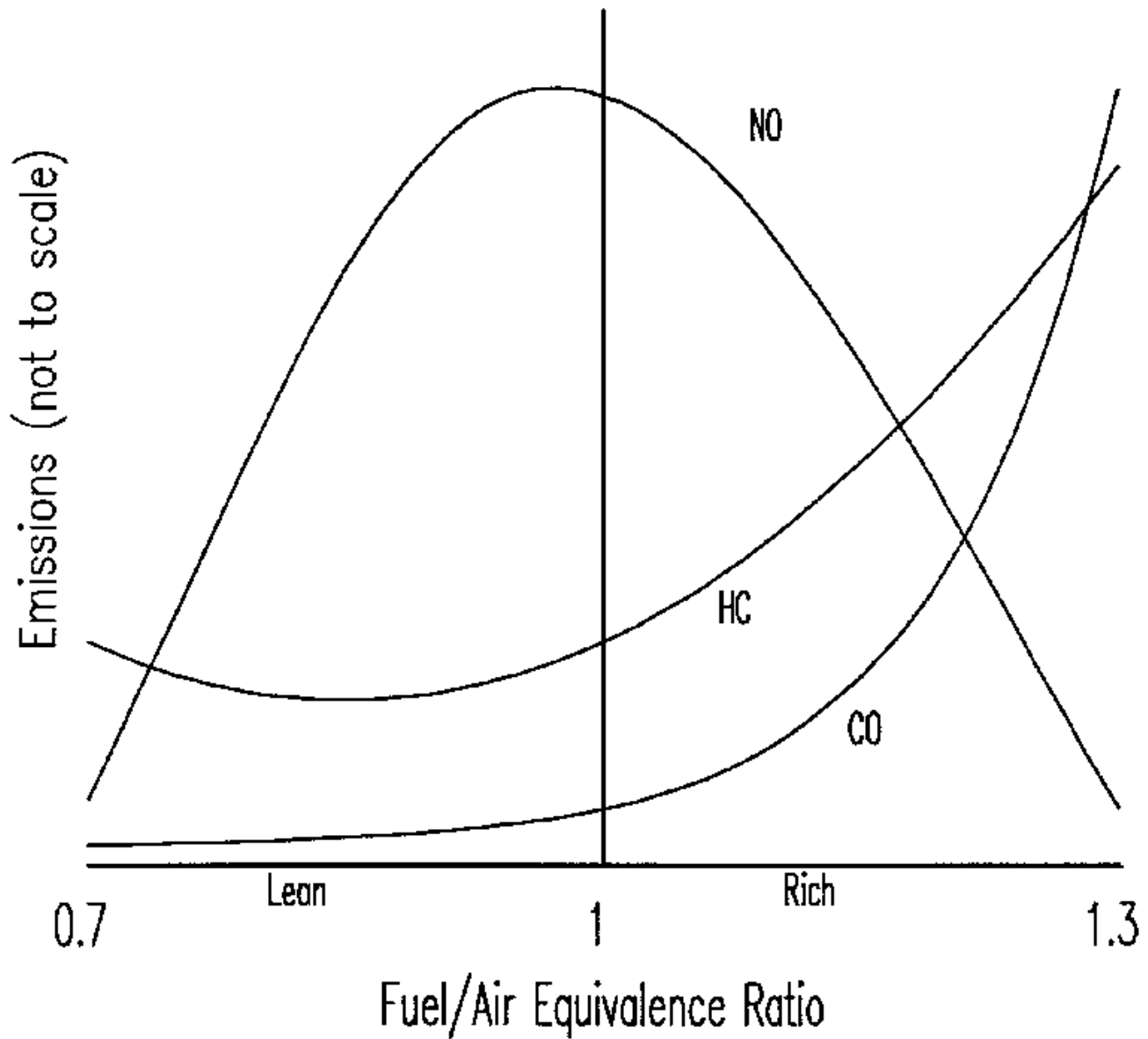
S.i. engines are usually fueled with gasoline, alcohol, or natural gas, but can use other liquid or gaseous fuels. It is important that any s.i. fuel be resistant to autoignition. This resistance is expressed in terms of the **octane number** of the fuel, based on an empirical scale on which iso-octane has been assigned a rating of 100 and *n*-heptane a rating of zero. Typical gasolines have octane numbers in the 85–105 range; these octane numbers are usually obtained with the aid of additives. Liquid s.i. engine fuels must be adequately volatile to evaporate fully prior to ignition, but not so volatile as to cause problems with storage and transfer.

Besides carbon dioxide and water, the combustion process in s.i. engines produces several pollutants (Fig. 66.7): carbon monoxide (CO), unburned hydrocarbons (HCs), and nitric oxide (NO). Large amounts of CO are formed as an equilibrium product in rich mixtures, while smaller amounts remain in the products of lean mixtures due to chemical kinetic effects. HCs are left over from the combustion of rich mixtures and from flame quenching at walls and crevices in lean mixtures. NO is formed from air at high temperatures, and the chemical kinetics allow it to remain as the burned gas cools. Most contemporary engines meet CO and HC emissions standards by running lean and using **catalytic converters** in their exhaust systems to complete the combustion. NO is typically reduced by limiting flame temperature through lean operation and exhaust gas



recirculation. Catalytic converters that can simultaneously control NO, CO, and HC are also in use, but they require careful control of the air/fuel ratio.

**Figure 66.7** Effect of equivalence ratio on s.i. engine emissions.



## Control

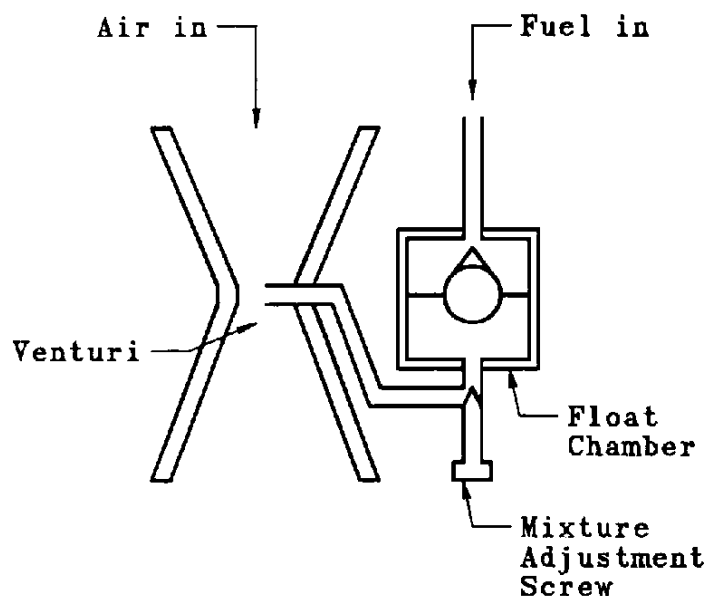
The control system of an s.i. engine must govern engine output, but it must also regulate equivalence ratio and spark timing. Since the ranges of  $\phi$  and spark timing over which the engine will run smoothly are limited, engine output is varied by reducing air flow while holding  $\phi$  and

timing essentially constant. Control of  $\phi$  and timing is directed toward maximizing efficiency and minimizing emissions at a given speed and torque.

Engine output is usually controlled by throttling the intake air flow with a butterfly-type throttle valve. This reduces net output by reducing the heat release and increasing the pumping work [Fig. 66.3(d)]. Other methods (late intake valve closing, shutting down cylinders of multicylinder engines) have been tried to reduce output with less efficiency penalty, but they are not widely used.

There are two basic methods of mixing fuel and air for s.i. engines: carburetion and fuel injection. A **carburetor** [Fig. 66.8(a)] provides intrinsic control of  $\phi$  by putting fuel and air flow through restrictions with the same differential pressure. The intrinsic control is imperfect because air is compressible while liquid fuels are not. Various corrective methods are used to provide near-constant lean  $\phi$  over most of the air flow range, with enrichment to  $\phi > 1$  for starting and maximum power operation. The manifold between the carburetor and the cylinder(s) must be arranged so that fuel evaporates fully (and is evenly distributed among the cylinders). Due to the difficulty in obtaining low emissions levels, no contemporary U.S. production automobiles use carburetors.

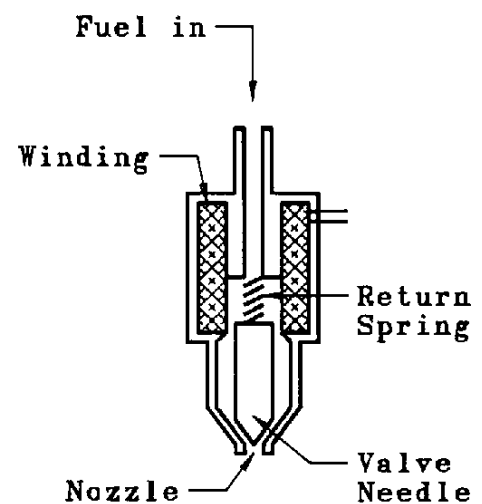
**Figure 66.8** S.i. engine fuel addition devices.



## Simple Carburetor

Only Main Jet Shown

(a)



## Fuel Injector

(b)

A **fuel injector** [Fig. 66.8(b)] injects a spray of fuel into the air stream. In throttle body injection, a single injector serves for multiple cylinders; in port injection (more common), each

cylinder has its own injector. Port injectors are timed to spray fuel while the inlet valve is closed, to allow evaporation time. They are typically controlled by digital electronics. The volume of injected fuel is controlled in response to various measurements, including speed, inlet manifold vacuum, and exhaust oxygen concentration. As for carburetors, the mixture is kept lean, except for starting and maximum power. Figure 66.8 shows carburetors and fuel injectors used for liquid fuels; the arrangements for gaseous fuels are similar.

The high voltage (10–25 kV) for the ignition spark is provided either by interrupting current through a choke or discharging a capacitor. The spark advance is typically regulated in response to engine speed and manifold vacuum; high speeds and high vacuums require more advance. The switching required for spark generation and control can be done either mechanically or electronically. Some electronically controlled engines incorporate vibrational knock sensors to retard the spark if required. For older designs, control is almost entirely mechanical, with the necessary adjustments to the fuel addition and ignition systems made through pressure-driven diaphragms, centrifugal speed sensors, and linkages. On newer designs, control is mainly through electronic sensors, digital electronics, and solenoid actuators.

## Advantages

Relative to compression ignition engines, s.i. engines have higher mass and volume power density, lower first cost, greater fuel availability (for automotive use), and wider speed range. Emissions are lower with use of a catalytic converter. The advantages of s.i. engines become more pronounced for smaller sizes.

## 66.4 Compression Ignition Engines

---

### Idealized and Actual Cycles

The compression ignition (c.i.) engine can be idealized as a diesel cycle using an ideal gas with constant specific heat [Fig. 66.3(b)]. The **diesel** cycle consists of isentropic compression, constant pressure heating (simulating combustion), isentropic expansion, and constant volume cooling (simulating intake and exhaust). The thermal efficiency of a diesel cycle (Fig. 66.4) is  $\eta_t = 1 - r_\nu^{1-\gamma} (r_c^\gamma - 1) / (r_c - 1) / \gamma$ . The **cutoff ratio**,  $r_c$ , idealizes the volume ratio over the fuel addition period.

The actual engine cycle [Fig. 66.3(c)] differs from the diesel cycle (1) in that heat transfer occurs during compression and expansion; (2) in that combustion takes place at varying rather than constant pressure; (3) in that combustion continues after the end of fuel addition; (4) in the presence of intake and exhaust processes; and (5) in the variation in gas composition and gas specific heat. For a given  $r_\nu$  and  $r_c$ , the efficiency of a typical s.i. engine is considerably lower than that of the ideal cycle (Fig. 66.4). The pressure-volume diagrams for actual s.i. and c.i. engines are quite similar.

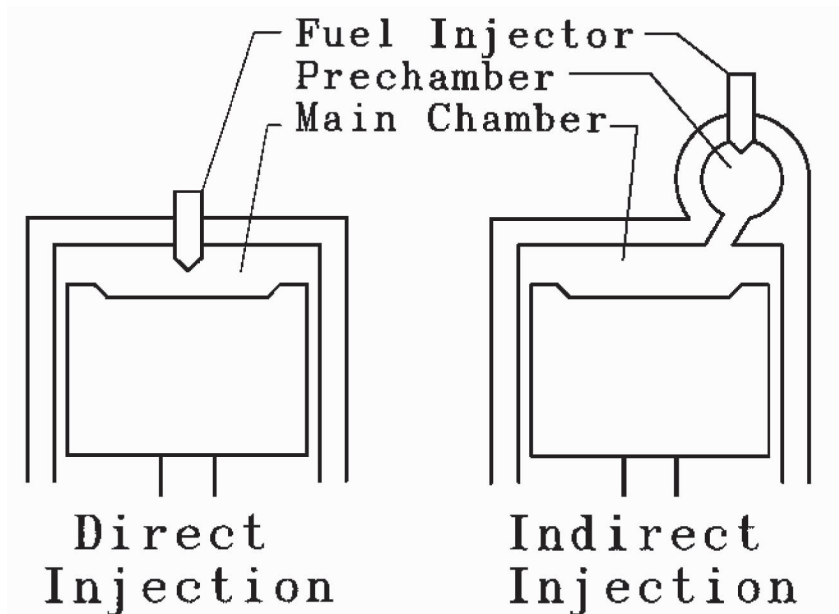
## Combustion, Fuels, and Emissions

In a c.i. engine, air is compressed before fuel is added. Fuel is injected as a fine spray beginning slightly before the volume reaches a minimum and ignites after coming in contact with the hot air. Overall equivalence ratio ( $\phi$ ) generally ranges from about 0.15–0.8, with lean mixtures (low  $\phi$ ) corresponding to idle and low power, and rich mixtures corresponding to full power with considerable smoke emission. Since the fuel and air are not premixed, combustion takes place at near  $\phi = 1$ , no matter what the overall  $\phi$ .

C.i. engine combustion takes place in four stages [Fig. 66.5(b)]. In the ignition delay period, fuel evaporates, mixes with the air, and reacts slowly. In the premixed combustion phase, the fuel which evaporated and mixed during the delay period burns rapidly in a process similar to that in s.i. knock. In the mixing–controlled combustion phase, a diffusion flame exists at the boundary between a rich atomized fuel-air mixture and the remaining air in the cylinder. In the late combustion phase, the pockets of fuel which so far have escaped the flame are consumed. Since the premixed combustion has a rapid pressure rise which causes rough operation, it is desirable to minimize the amount of fuel vaporized before it begins. This is done by minimizing ignition delay time and evaporation rate during that time. Minimum delay is obtained by injecting at the optimum time (10–15° before top center), with high cylinder wall temperature, high compression ratio, and high cetane number fuel. Indirect injection (see below) gives little premixed combustion.

The combustion process in c.i. engines does not speed up with increased turbulence as much as the process in s.i. engines does. For large, low-speed engines, combustion is adequate when the fuel is injected directly into the center of a relatively quiescent combustion chamber. For medium-size, medium-speed engines, the combustion chamber must be designed for increased turbulence in order for combustion to take place in the time available. In small, high-speed engines, combustion is initiated in a small, hot, highly turbulent prechamber. The fuel and burned gases from the prechamber then expand into the main combustion chamber and combine with the remaining air. Engines with a single chamber are known as **direct injection** (d.i.) engines, while those with a prechamber are known as **indirect injection** (i.d.i.) engines (Fig. 66.9).

**Figure 66.9** C.i. engine combustion chamber types.



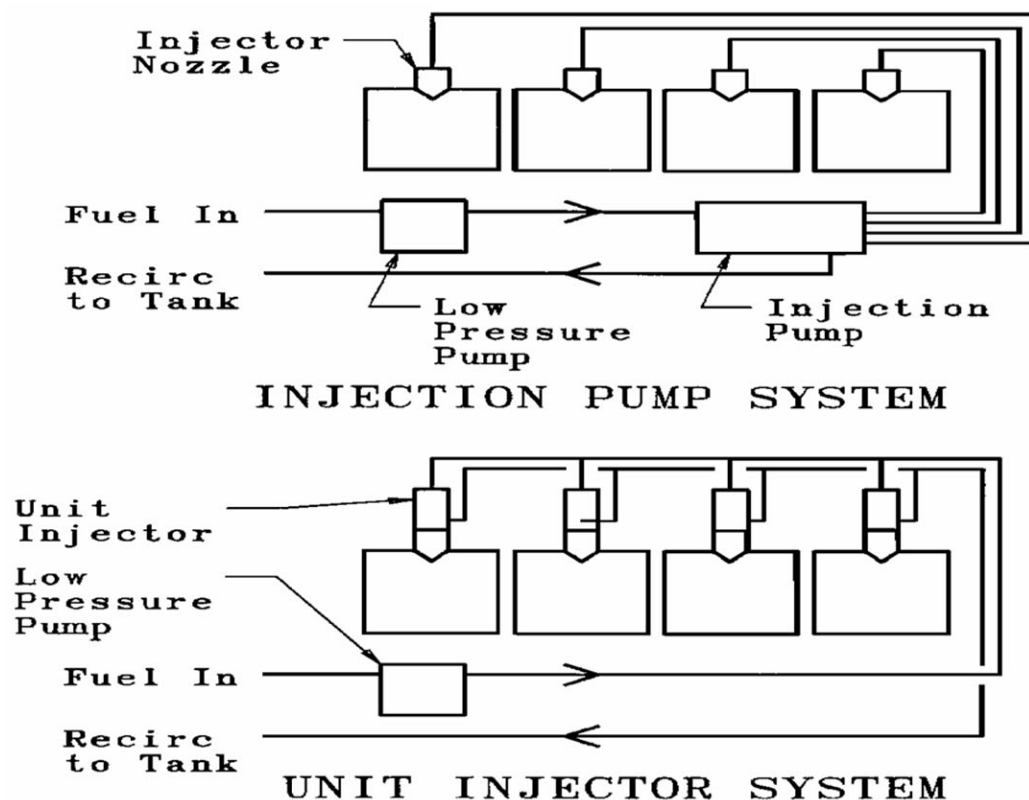
C.i. engines are fueled with petroleum oils consisting of longer-chain molecules than those in gasolines. Depending on the engine design, oils ranging from crude to kerosene can be used. It is important that any c.i. fuel have adequate autoignition properties. The ignition quality is expressed in terms of the **cetane number** of the fuel, based on an empirical scale on which *n*-hexadecane (cetane) has been assigned a rating of 100 and heptamethylnonane (iso-cetane) a rating of 15. Typical c.i. fuels have cetane numbers in the 30–60 range. For heavy, low-cost c.i. fuels, high pour point can be a problem. Some of these fuels must be heated before they can be pumped.

Besides carbon dioxide and water, the combustion process in c.i. engines produces several pollutants, the most important of which are soot (carbon plus hydrocarbons), nitric oxide (NO), and nitrogen dioxide (NO<sub>2</sub>). Carbonaceous soot is formed by fuel pyrolysis in rich regions near the flame front. Hydrocarbons then adsorb onto the soot particles during expansion and exhaust. Soot emissions are highest at high loads. NO and NO<sub>2</sub> are formed from air at high temperatures, and the chemical kinetics allow them to remain as the burned gases cool. Carbon monoxide and gaseous hydrocarbon emissions from diesel engines are relatively small.

## Control

C.i. engines are not throttled, but are controlled by regulating the amount of fuel injected and the injection timing. Because of the required high injection pressures, almost all diesel engine fuel injectors are mechanically rather than electrically driven. Two types of systems are used: injection pump and unit injector (Fig. 66.10). In an **injection pump** system, a central pump timed to the camshaft delivers fuel to nozzles located at each cylinder. The pump typically has individual barrels for each cylinder, but a single barrel with a fuel distributor is also used. In a **unit injector** system, there is a pump and nozzle on each cylinder, driven by a shaft running over all the cylinder heads.

**Figure 66.10** C.i. fuel systems.



The injection start and duration are varied according to engine load and operating conditions. In the past, the control was generally accomplished through purely mechanical means and consisted mainly of increasing the injection duration in response to increased torque demand. In recent years, however, electronically controlled injector pumps have become common. In these units, the power is supplied mechanically, but fuel delivery is controlled by unloading solenoids. Electronic control allows fuel delivery to be adjusted in response to engine operating conditions and is useful in achieving low emissions.

Since c.i. engines are not throttled at reduced load, they do not provide engine braking. For heavy vehicle use, c.i. engines are often fitted with auxiliary compression brakes which increase engine pumping work.

## **Advantages**

Relative to spark ignition engines, c.i. engines have higher thermal efficiency at full load and much higher thermal efficiency at low load. They also are capable of using inexpensive fuels such as heavy fuel oil.

## **66.5 Gas Exchange Systems**

---

The torque of an internal combustion engine is primarily limited by the mass of air that can be captured in the cylinder. Intake system design is therefore a major factor in determining the torque for a given engine displacement. Since residual exhaust gas takes up space that could be used for fresh charge, exhaust system design also affects engine output. Exhaust design is also driven by the need to muffle the noise generated by the sudden flow acceleration at exhaust valve opening.

## **4-Stroke Intake and Exhaust**

Air flow into a naturally aspirated 4-stroke engine is optimized by reducing charge temperature, reducing flow friction in the intake system, reducing residual exhaust gas, and tuning and extended valve opening.

It has been found experimentally that engine air flow and torque are inversely proportional to the square root of the stagnation temperature of the air entering the cylinder. In s.i. engines, torque is increased by the cooling effect of fuel evaporation. This effect is much larger with alcohol fuels, which are therefore used in many racing cars. Torque is adversely affected by heat transfer to the intake air from the hot cylinder and exhaust manifold. In carbureted s.i. engines, some heat transfer is necessary to prevent fuel from puddling in the intake manifold.

According to both experiment and theory, engine torque is proportional to the pressure of the air entering the cylinder. This pressure is increased by minimizing the intake pressure drop. Intake valves are thus made as large as possible. High performance engines utilize two or three intake valves per cylinder to maximize flow area. Intake piping is normally designed for minimum pressure drop. However, in carbureted s.i. engines, the intake manifold is often designed for optimal fuel evaporation and distribution rather than for minimum flow friction. Intake air cleaners are designed for minimum flow resistance consistent with adequate dirt

removal.

Cylinder pressure just prior to exhaust valve opening is much higher than atmospheric, but pressure falls rapidly when the exhaust valve opens. Back pressure due to exhaust system pressure drop increases the concentration of burned gas in the charge and thus reduces torque. The effect of exhaust system pressure drop is less than that of intake system pressure drop, but two exhaust valves per cylinder are sometimes used to minimize pressure drop. Exhaust system pressure drop is usually increased by the use of a muffler to reduce exhaust noise.

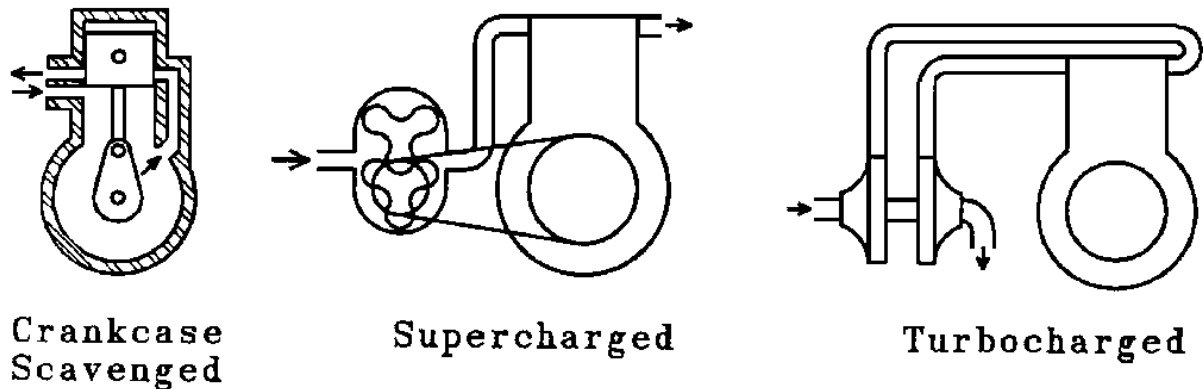
Good intake and exhaust system design makes use of the dynamic effects of gas acceleration and deceleration. Most engine designs incorporate open periods well over  $180^\circ$  of crank angle: intake valves open  $5\text{--}30^\circ$  before top center and close  $45\text{--}75^\circ$  after bottom center; exhaust valves open  $40\text{--}70^\circ$  before bottom center and close  $15\text{--}35^\circ$  after top center. (The longer valve open times correspond to high performance engines.) Valves are thus open when piston motion is in the opposite direction from the desired gas flow. At high engine speeds, the correct flow direction is maintained by gas inertial effects. At low engine speeds, an extended valve open period is detrimental to performance. Some engines incorporate variable valve timing to obtain optimum performance over a range of speeds.

The extended valve open period is generally used in combination with intake and exhaust **tuning**. Intake systems are often acoustically tuned as organ-pipe resonators, Helmholtz resonators, or more complex resonating systems. By tuning the intake to 3, 4, or 5 times the cycle frequency, pressure at the intake valve can be increased during the critical periods near valve opening and closing. Such tuning is usually limited to diesel, port fuel injected, or one carburetor per cylinder engines because the design of intake manifolds for other carbureted engines and throttle-body injected engines is dominated by the need for good fuel distribution. The tuning penalizes performance at some speeds away from the design speed. Branched exhaust systems are tuned so that, at design speed, expansion waves reflected from the junctions arrive at the exhaust valve when it is near closing. Individual cylinder exhaust systems are tuned as organ-pipe resonators. In either case, performance away from the design speed is penalized.

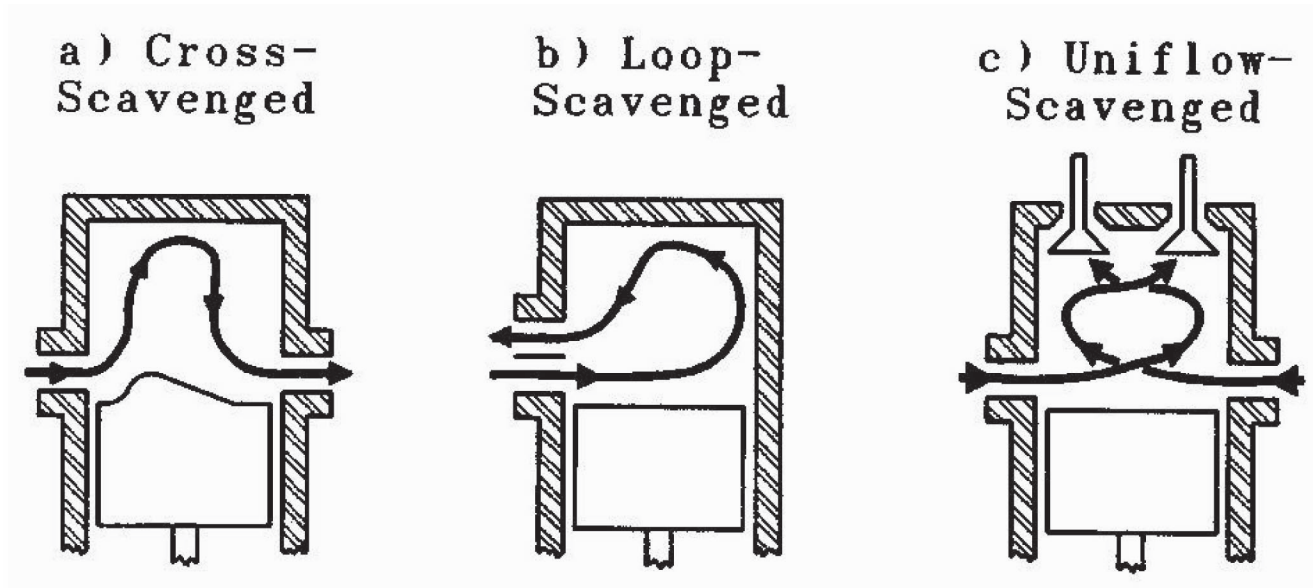
## 2-Stroke Scavenging

In a 2-stroke engine, intake and exhaust take place simultaneously, and some means of air pumping is needed for gas exchange (Fig. 66.11). Small s.i. engines are generally crankcase scavenged—the bottom face of the piston is used to pump the air and oil is generally added to the fuel to lubricate the crank bearings. Larger engines use either rotary superchargers or turbochargers. These allow more freedom in crankcase lubrication. The cylinder and piston are arranged to maximize inflow of fresh charge and outflow of exhaust while minimizing their mixing. Cross-scavenging [Fig. 66.12(a)] and loop-scavenging [Fig. 66.12(b)] require only cylinder wall ports; uniflow scavenging [Fig. 66.12(c)] requires poppet valves as well.

**Figure 66.11** Scavenging and supercharging systems



**Figure 66.12** Scavenging arrangements.



Scavenging spark ignition engines involves a trade-off between residual gas left in the cylinder and air-fuel mixture lost out the exhaust. In compression ignition engines, only air is lost through the exhaust. Two-stroke s.i. engines are thus used mainly where low weight and first cost are of primary importance, while 2-stroke c.i. engines can be built to be suitable for any service.



## Supercharging and Turbocharging

The output of a given i.c. engine can be increased by providing an auxiliary air compressor, or supercharger, to increase the pressure and, thus, the density of the air entering the cylinder intake. Although it actually applies to all types of auxiliary air compression systems, the term *supercharger* is generally used to describe systems driven by the engine output shaft. Air compression systems powered by an exhaust gas-driven turbine are known as *turbochargers*. When supercharging or turbocharging is added to a naturally aspirated engine, the engine is usually modified to reduce its compression ratio. However, the overall compression ratio is increased.

Most shaft-driven superchargers are positive displacement compressors, the most common being the Roots blower (Fig. 66.11). Shaft-driven, positive-displacement superchargers have the advantages of increasing their delivery in proportion to engine speed and of responding almost instantly to speed changes. Their disadvantage is that the use of shaft power to drive the compressor results in decreased overall thermal efficiency. Roots blowers are unacceptably inefficient at pressure ratios greater than about 2.

## 66.6 Design Details

---

### Engine Arrangements

Various engine cylinder arrangements are shown in Fig. 66.2.

In-line engines are favored for applications in which some sacrifice in compactness is justified by mechanical simplicity and ease of maintenance. They are also used where the need for a narrow footprint overrides length and height considerations. The in-line design is most popular for small utility and automobile engines, small truck engines, and very large marine and stationary engines.

Vee engines are used where compactness is important. Vee engines are used for large automobile engines, large truck engines, locomotive engines, and medium-size marine and stationary engines.

Opposed engines are used primarily where low height is important—in rear engine automobiles and for small marine engines meant for below-deck installation. They are also used for some small aircraft engines, where they allow for ease in air cooling and servicing.

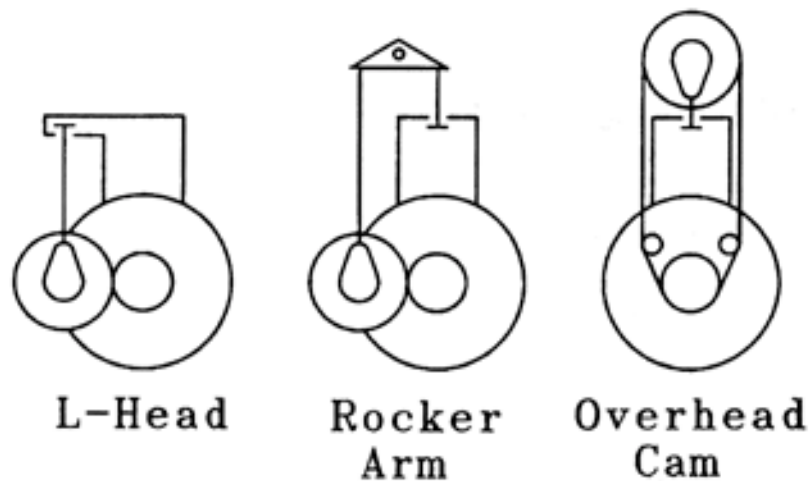
Radial engines are used primarily in aircraft, where their design allows for efficient air cooling.

Rotary (Wankel) engines have been used primarily in sports cars. They have not captured a major share of any market sector.

### Valve Gear

Poppet valves on 4-stroke and uniflow 2-stroke engines fall into one of three categories: valve-in-block, valve-in-head/rocker arm, or valve-in-head/overhead cam. The arrangements are illustrated in Fig. 66.13.

**Figure 66.13** Valve arrangements.



**Valve-in-block** engines are the cheapest to manufacture. In an L-head arrangement, intake and exhaust valves are on the same side of the cylinder; in a T-head engine, they are on opposite sides. The valves are directly driven by a camshaft located in the block, gear or chain driven at half the crankshaft speed. The performance of these engines suffers from the elongated shape of the combustion chamber, and the designs are currently used only for inexpensive utility engines.

Valve-in-head/**rocker arm** engines have the valves installed in the cylinder head while maintaining a camshaft in the block. The design allows compact combustion chambers, but control of valve motion suffers from the slack in the long mechanical drive train. A majority of production automobile engines have this type of valve drive.

Valve-in-head/**overhead cam** engines have valves in the cylinder head directly driven by a camshaft running over all the heads. The design allows both compact combustion chamber design and accurate control of valve motion, but is more expensive to manufacture and more difficult to maintain than rocker arm designs. In recent years, overhead cam designs have become increasingly common in high-performance automobile engines.

## Lubrication

The bearings of most i.c. engines are plain or grooved journal bearings. Rolling contact bearings are rarely used. In the crankshaft bearings, a hydrodynamic film is maintained by rotation; in the piston pin bearings, the maintenance of a film depends on the oscillating nature of the load. For the connecting rod bearings the two effects are combined.

The most critical lubrication areas in an i.c. engine are at the piston rings, which are required to seal the high-pressure gas in the cylinder and prevent excess oil from entering the cylinder. Typical designs have two compression rings to seal the gases and an oil control ring to wipe oil from the cylinder wall. Piston rings ride on a hydrodynamic film at midstroke, but are in a boundary lubrication regime near top and bottom center. Lubrication is aided by good ring (alloy cast iron) and cylinder wall (cast iron or chrome-plated steel) materials.

Low-cost engines are splash lubricated by running the crankshaft partly in an oil pan, but all other engines have force-feed lubrication systems that deliver filtered oil to the bearings, the cylinder walls below the piston, and the valve train. Smaller engines depend on convection from the oil pan to cool the oil, while others have auxiliary oil coolers.

## Cooling

Most large i.c. engines are liquid cooled, and most small engines are air cooled. About a third of the energy input to a typical engine is dissipated through the cooling system. Liquid-cooled engines use either water or an aqueous ethylene glycol solution as coolant. When the glycol is used, it gives lower freezing and higher boiling points, but also increases the viscosity of the coolant. Although some natural-convection cooling systems have been built, most engines have the coolant pumped through numerous passages in the cylinder walls and heads and then into a heat exchanger where the heat is transferred to the environment. Small marine engines are typically cooled directly with water from the environment.

Air-cooled engines have finned external surfaces on their pistons and heads to improve heat transfer and fans to circulate air over the engine. The larger passages needed for air require that the cylinders be more widely spaced than for liquid-cooled engines. While most air-cooled engines are small, many large aircraft engines have been air cooled.

## 66.7 Design and Performance Data for Typical Engines

Design and performance data for various engines are given in [Table 66.1](#).

**Table 66.1** Design and Performance Data for Various Internal Combustion Engines

Application and Type	Cylinders/Arrangement	Displ. (l)	Comp. Ratio	Rated Power (kW)	Rated Speed (rpm)	Mass (kg)
Utility, 2-stroke, s.i., c.s.	1	0.10	9.0	8.9	9000	5.0
Marine, 2-stroke, s.i., c.s.	1	0.13	10.5	7.5	8000	6.6
Utility, 4-stroke, s.i., n.a.	1	0.17	6.2	2.5	3600	13.9
Motorcycle 2-stroke, s.i., c.s.	2/in-line	0.30	7.1	19.4	7000	
Utility, 4-stroke, s.i., n.a.	1	0.45	8.7	11.9	3600	38.2
Motorcycle 4-stroke, s.i., n.a.	2/in-line	0.89	10.6	61	6800	
Automobile 4-stroke, s.i., n.a.	4/in-line	2.2	9.0	73	5200	
Automobile 4-stroke, s.i., t.c.	4/in-line	2.2	8.2	106	5600	

Automobilele 4-stroke, c.i., n.a.	4/in-line	2.3	23	53	4500	
Automobilele 4-stroke, c.i., t.c.	4/in-line	2.3	21	66	4150	
Aircraft 4-stroke, s.i., n.a.	4/opposed	2.8	6.3	48	2300	76
Automobilele 4-stroke, c.i., n.a.	8/vee	5.0	8.4	100	3400	
Truck/Bus, 2-stroke, c.i., t.c.	8/vee	9.5	17	280	2100	1100
Truck/Bus, 4-stroke, c.i., t.c.	6/in-line	10	16.3	201	1900	890
Aircraft 4-stroke, c.i., s.c.	9/radial	30	7.21	1 140	2800	670
Locomotiveve 2-stroke, c.i., t.c.	16/vee	172	16	2 800	950	16 700
Locomotiveve 4-stroke, c.i., t.c.	16/vee	239	12.2	3 400	1000	25 000
Large Marine, 2-stroke, c.i., t.c.	12/in-line	14 500		36 000	87	$1.62 \cdot 10^6$

---

Based on Taylor, C. F. 1985. *The Internal Combustion Engine in Theory and Practice*, 2nd ed. MIT Press, Cambridge, MA.

C.s.: crankcase scavenged; n.a.: naturally aspirated; t.c.: turbocharged.

## Defining Terms

**Carburetor:** Controls fuel-air mixture by flowing air and fuel across restrictions with the same differential pressure.

**Catalytic converter:** Uses catalyst to speed up chemical reactions, normally slow, which destroy pollutants.

**Cetane number:** Empirical number quantifying ignition properties of c.i. engine fuels.

**Compression ignition engine:** Fuel and air compressed separately, ignited by high air compression temperatures.

**Compression ratio:** Ratio of maximum working volume to minimum working volume.

**Cutoff ratio:** Fraction of expansion stroke during which heat is added in diesel cycle.

**Diesel cycle:** Thermodynamic idealization of compression ignition engine.

**Direct injection c.i. engine:** Fuel is injected directly into the main combustion chamber.

**Equivalence ratio:** Fuel/air ratio relative to fuel/air ratio for stoichiometric combustion.

**Indirect injection c.i. engine:** Fuel is injected into a prechamber connected to the main combustion chamber.

**Injection pump:** Delivers metered high-pressure fuel to all fuel injector nozzles of a c.i. engine.

**Fuel injector:** Controls fuel-air mixture by metering fuel in proportion to measured or predicted air flow.

**Knock:** Spark ignition engine phenomenon in which fuel-air mixture detonates instead of burning smoothly.

**Naturally aspirated engine:** Piston face pumping action alone draws in air.

**Octane number:** Empirical number quantifying antiknock properties of s.i. fuels.

**Otto cycle:** Thermodynamic idealization of spark ignition engine.

**Overhead cam engine:** Valves are in head, driven by camshaft running over top of head.

**Rocker arm engine:** Valves are in head, driven from camshaft in block by push rods and rocker arms.

**Scavenging:** Intake/exhaust process in 2-stroke engines.

**Spark ignition engine:** Fuel and air compressed together, ignited by electric spark.

**Supercharged engine:** Shaft-driven air compressor forces air into cylinder.

**Thermal efficiency:** Engine work divided by heat input or lower heating value of fuel used.

**Tuning:** Designing intake and exhaust so that flow is acoustically reinforced at design speed.

**Turbocharged engine:** Air forced into cylinder by compressor driven by exhaust gas turbine.

**Unit injector:** Combination pump and nozzle which delivers metered fuel to a single c.i. engine cylinder.

**Valve-in-block engine:** Low-cost design in which valves are driven directly by a camshaft in the cylinder block.

**2-stroke engine:** One power stroke per cylinder per revolution.

**4-stroke engine:** One power stroke per cylinder per two revolutions.

## References

Benson, R. S. and Whitehouse, N. D. 1989. *Internal Combustion Engines*. Pergamon, New York.  
Cummins, L. C., Jr. 1989. *Internal Fire*, rev. ed. Society of Automotive Engineers, Warrendale, PA.

Heywood, J. B. 1988. *Internal Combustion Engine Fundamentals*. McGraw-Hill, New York.

Obert, E. F. 1968. *Internal Combustion Engines and Air Pollution*, 3rd ed. Harper Collins, New York.

Taylor, C. F. 1985. *The Internal Combustion Engine in Theory and Practice*, 2nd ed. MIT Press, Cambridge, MA.

## Further Information

*Internal Combustion Engines and Air Pollution* by Edward F. Obert and *The Internal Combustion Engine in Theory and Practice* by Charles Fayette Taylor are comprehensive and highly readable texts on i.c. engines. Although they are somewhat dated, they are still invaluable sources of information.

*Internal Combustion Engine Fundamentals* by John B. Heywood is an up-to-date and comprehensive text, but is less accessible to those with no previous i.c. engine background than the texts above.

*Internal Fire* by Lyle C. Cummins, Jr., is a fascinating history of the i.c. engine.

The Society of Automotive Engineers publishes *SAE Transactions* and a wide variety of books and papers on internal combustion engines. For more information contact: SAE, 400 Commonwealth Drive, Warrendale, PA, 15096, USA. Phone (412) 776-4841.

Langston, L. S., Opdyke, Jr. G. "Gas Turbines"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 67

## Gas Turbines

---

### 67.1 Gas Turbine Usage

### 67.2 Gas Turbine Cycles

### 67.3 Gas Turbine Components

Compressors and Turbines • Combustors

#### **Lee S. Langston**

*University of Connecticut*

#### **George Opdyke, Jr.**

*Dykewood Enterprises*

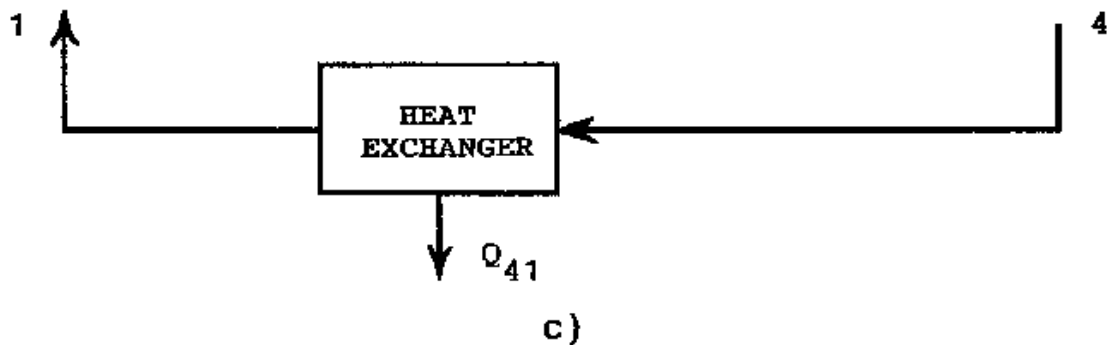
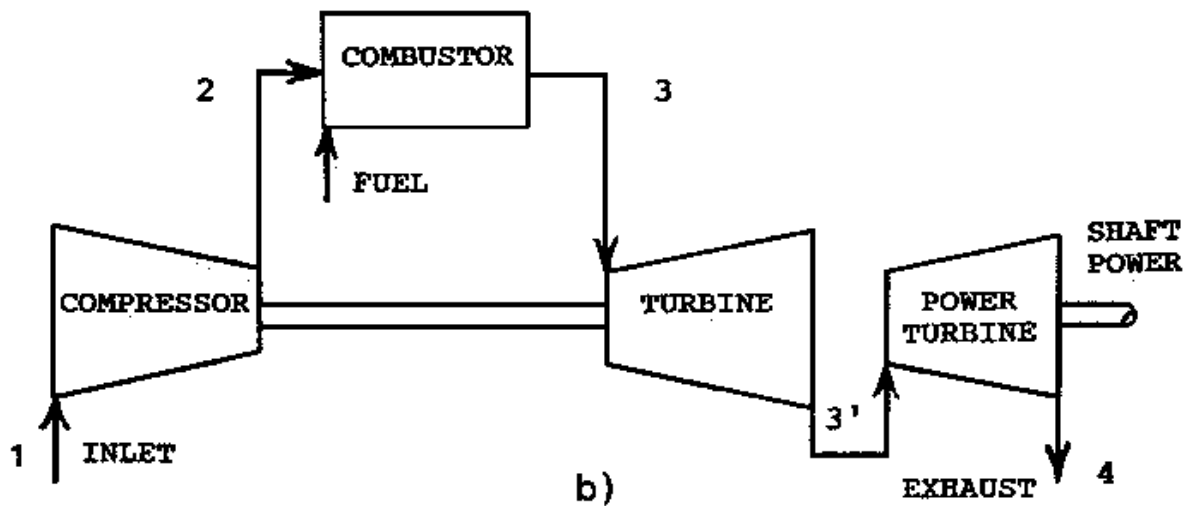
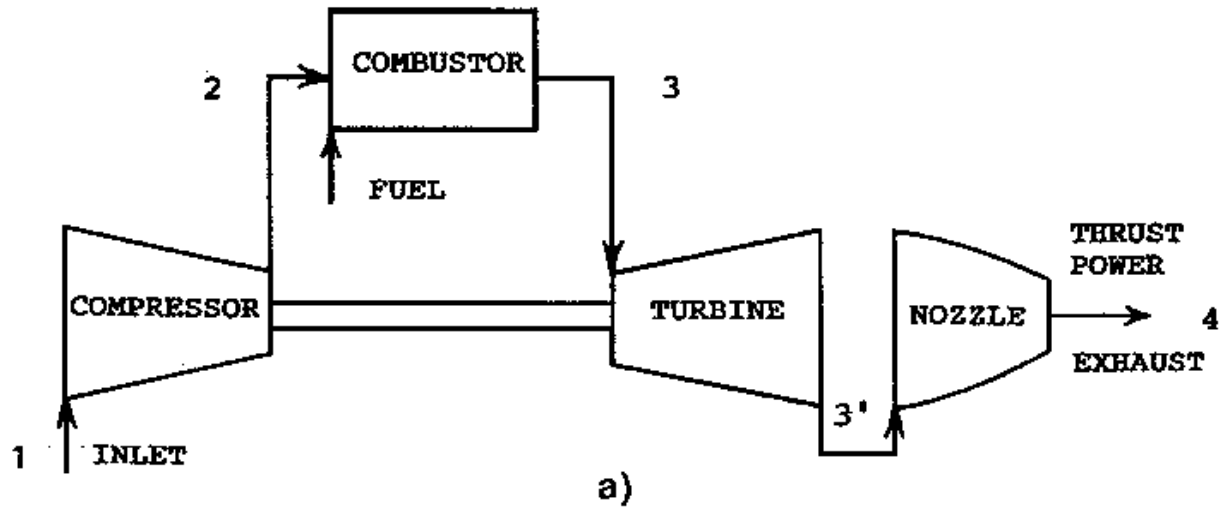
In the history of energy conversion the gas turbine is a relatively new prime mover. The first practical gas turbine used to generate electricity ran at Neuchatel, Switzerland, in 1939 and was developed by the Brown Boveri firm. The first gas turbine–powered airplane flight also took place in 1939, in Germany using the gas turbine developed by Hans P. von Ohain. In England the 1930s invention and development of the aircraft gas turbine by Frank Whittle resulted in a similar British flight in 1941.

The name *gas turbine* is somewhat misleading, for it implies a simple turbine that uses gas as a working fluid. Actually, a gas turbine (as shown schematically in [Fig. 67.1](#)) has a *compressor* to draw in and compress gas (usually air), a *combustor* (or burner) to add fuel to heat the compressed gas, and a *turbine* to extract power from the hot gas flow. The gas turbine is an internal combustion (IC) engine employing a continuous combustion process, as distinct from the intermittent combustion occurring in a diesel or Otto cycle IC engine.

Because the 1939 origin of the gas turbine lies both in the electric power field and in aviation, there has been a profusion of "other names" for the gas turbine. For land and marine applications it is generally called a *gas turbine*, but also a *combustion turbine*, a *turboshaft engine*, and sometimes a *gas turbine engine*. For aviation applications it is usually called a *jet engine*, and various other names (depending on the particular aviation configuration or application) such as *jet turbine engine*, *turbojet*, *turbofan*, *fanjet*, and *turboprop* or *prop jet* (if it is used to drive a propeller). The compressor-combustor-turbine part of the gas turbine ([Fig. 67.1](#)) is commonly termed the *gas generator*.



**Figure 67.1** Gas turbine schematics. (a) Jet engine, a gas turbine (open cycle) used to produce thrust power. (b) Gas turbine (open cycle) used to produce shaft power. (c) Heat exchanger added to (b) yields a closed-cycle gas turbine (combustor becomes another heat exchanger since closed-cycle working fluid cannot sustain combustion).

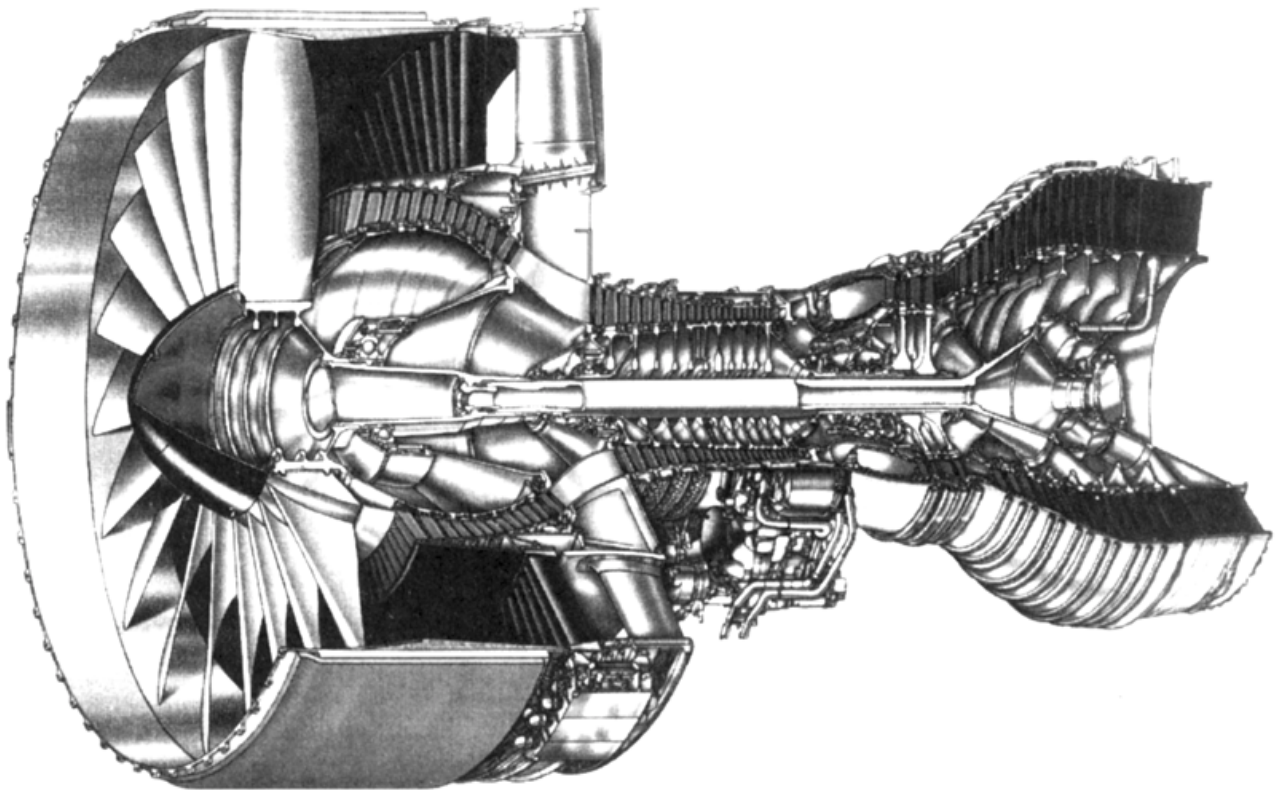


## 67.1 Gas Turbine Usage

In an aircraft gas turbine, all of the turbine power is used to drive the compressor (which may also have an associated fan or propeller). The gas flow leaving the turbine is then accelerated to the atmosphere in an exhaust nozzle [Fig. 67.1(a)] to provide *thrust* or *propulsion power*. Gas turbine or jet engine thrust power is the mass flow momentum increase from engine inlet to exit, multiplied by the flight velocity.

A typical jet engine is shown in Fig. 67.2. Such engines can range from about 100 pounds thrust (lbt)(445 N) to as high as 100 000 lbt, (445 000 N), with dry weights ranging from about 30 lb (134 N) to 20 000 lb (89 000 N). The jet engine of Fig. 67.2 is a *turbofan* engine, with a larger-diameter compressor-mounted fan. Thrust is generated both by air passing through the fan (bypass air) and through the gas generator itself. With a large frontal area, the turbofan generates peak thrust at low (takeoff) speeds, making it most suitable for commercial aircraft. A *turbojet* does not have a fan and generates all of its thrust from air that passes through the gas generator. Turbojets have smaller frontal areas and generate peak thrusts at high speeds, making them most suitable for fighter aircraft.

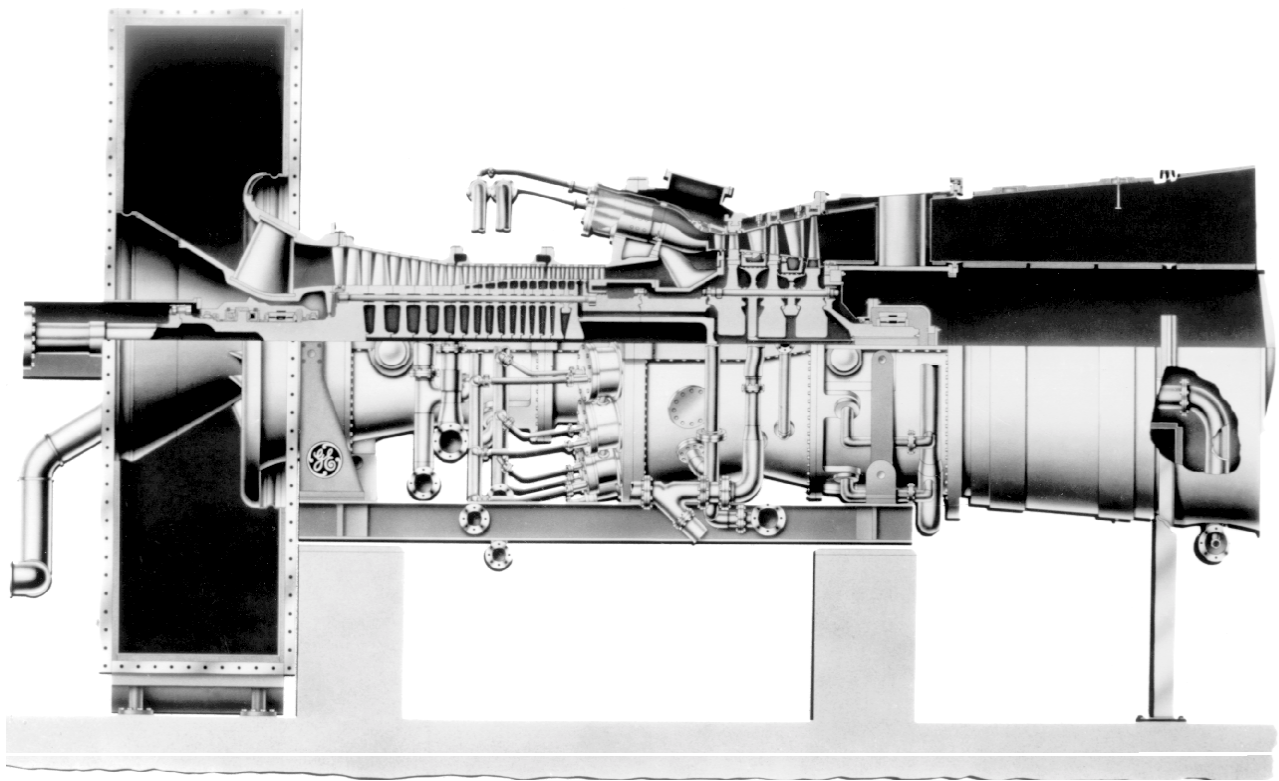
**Figure 67.2** A modern jet engine used to power Boeing 777 aircraft. This is a Pratt and Whitney PW4084 turbofan which can produce 84 000 pounds (374 kN) of thrust (about 63 MW at 168 m/s). It has a 112 in. (2.85 m) diameter front-mounted fan, a flange-to-flange length of 192 inches (4.87 m), and a mass of about 15 000 pounds (6804 kg). (Courtesy of Pratt and Whitney.)



In nonaviation gas turbines, only part of the turbine power is used to drive the compressor. The remainder, the "useful power," is used as output *shaft power* to turn an energy conversion device [Fig. 67.1(b)] such as an electrical generator or a ship's propeller. [The second "useful power" turbine shown in Fig. 67.1(b) need not be separate from the first turbine.]

A typical land-based gas turbine is shown in Fig. 67.3. Such units can range in power output from 0.05 MW to as high as 240 MW. The unit shown in Fig. 67.3 is also called an *industrial* or *frame* machine. Lighter-weight gas turbines derived from jet engines (such as in Fig. 67.2.) are called *aeroderivative* gas turbines and are most frequently used to drive natural gas line compressors and ships and to provide peaking and intermittent power for electrical utility applications.

**Figure 67.3** A modern land-based gas turbine used for electrical power production and for mechanical drives. This General Electric MS7001FA gas turbine is rated at 168 MW and is about 44 feet (13.4 m) in length and weighs approximately 377000 pounds (171000 kg). Similar units have been applied to mechanical drive applications up to 108200 hp. (Courtesy of General Electric.)



Some of the principle advantages of the gas turbine are as follows:

1. It is capable of producing large amounts of useful power for a relatively small size and weight.
2. Since motion of all its major components involves pure rotation (i.e., no reciprocating motions as in a piston engine), its mechanical life is long and the corresponding maintenance costs are relatively low.

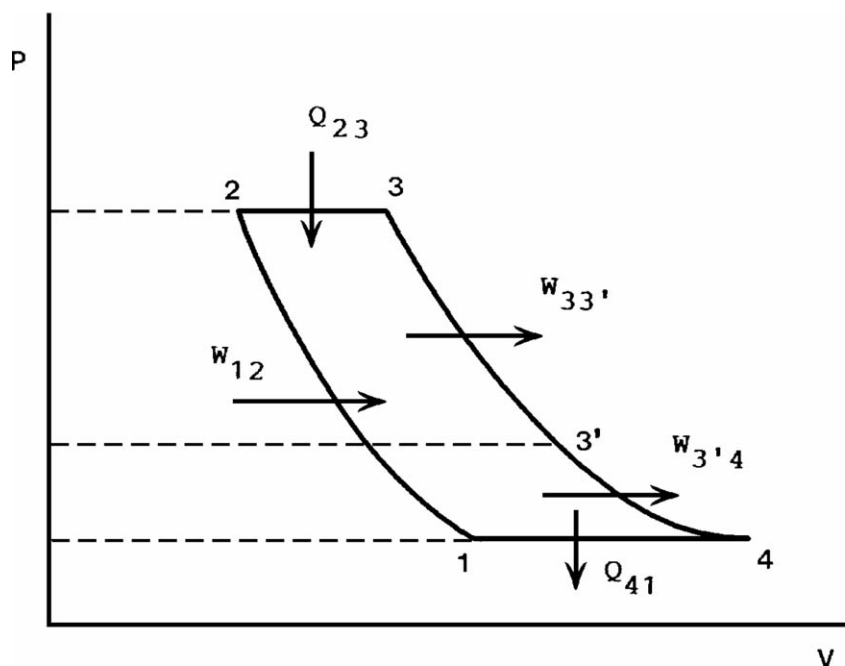
3. Although the gas turbine must be started by some external means (a small external motor or another gas flow source, such as another gas turbine), it can be brought up to full load conditions in minutes, in contrast to a steam turbine plant, whose start-up time is measured in hours.
4. A wide variety of fuels can be utilized. Natural gas is commonly used in land gas turbines, whereas light distillate (kerosene-like) oils power aircraft gas turbines. Diesel oil or specially treated residual oils can also be used, as well as combustible gases derived from blast furnaces, refineries, and coal gasification.
5. The usual working fluid is atmospheric air. As a basic power supply, the gas turbine requires no coolant (e.g., water).

In the past, one of the major disadvantages of the gas turbine was its lower thermal efficiency (hence, higher fuel usage) when compared to other IC engines and to steam turbine power plants. However, during the last 50 years, continuous engineering development work has pushed the lower thermal efficiency (18% for the 1939 Neuchatel gas turbine) to present levels of 40% for simple cycle operation and above 55% for combined cycle operation. Even more fuel-efficient gas turbines are in the planning stages, with simple cycle efficiencies predicted as high as 45 to 47% and combined cycles in the 60% range. These projected values are significantly higher than values for other prime movers, such as steam power plants.

## 67.2 Gas Turbine Cycles

The ideal gas Brayton cycle (1876, and also proposed by Joule in 1851)—shown in graphic form in Fig. 67.4 as a pressure-volume diagram—is an idealized representation of the properties of a fixed mass of gas (working fluid) as it passes through a gas turbine in operation (Fig. 67.1).

**Figure 67.4** Brayton cycle pressure ( $P$ ) versus volume ( $V$ ) diagram for a unit mass of working fluid (e.g., air), showing work ( $W$ ) and heat ( $Q$ ) inputs and outputs.



A unit mass of ideal gas (e.g., air) is compressed isentropically from point 1 to point 2. This represents the effects of an ideal adiabatic compressor [Fig. 67.1(a) and 67.1(b)] and any isentropic gas flow deceleration for the case of an aviation gas turbine in flight [Fig. 67.1(a)]. The ideal work necessary to cause the compression is represented by the area between the pressure axis and the isentropic curve 1-2.

The unit gas mass is then heated at constant pressure from 2 to 3 in Fig. 67.4 by the exchange of heat input,  $Q_{23}$ . This isobaric process is the idealized representation of heat addition caused by the combustion of injected fuel into the combustor in Fig. 67.1(a) and 67.1(b). The mass flow rate of the fuel is very much lower than that of the working fluid (roughly 1:50), so the combustion products can be neglected as a first approximation.

The unit gas mass is then isentropically expanded (lower pressure and temperature and higher volume) from 3 to 4 in Fig. 67.4, where  $P_4 = P_1$ . In Fig. 67.1 this represents flow through the turbine (to point 3') and then flow through the exit nozzle in the case of the jet engine [Fig. 67.1(a)], or flow through the power turbine in Fig. 67.1(b) (point 4). In Fig. 67.4 the area between the pressure axis and the isentropic curve 3-3' represents the ideal work,  $W_{33'}$ , derived from the turbine. This work has to be equal to the ideal compressor work  $W_{12}$  (the area bounded by curve 1-2). The ideal "useful work,"  $W_{3'4}$ , in Fig. 67.4 (area bounded by the isentropic curve 3'-4) is that which is available to cause output *shaft power* [Fig. 67.1(b)] or *thrust power* [Fig. 67.1(a)].

The Brayton cycle is completed in Fig. 67.4 by a constant pressure process in which the volume of the unit gas mass is decreased (temperature decrease) as heat  $Q_{41}$  is rejected. Most gas turbines operate in an *open cycle* mode, in which, for instance, air is taken in from the atmosphere (point 1 in Figs. 67.1 and 67.4) and discharged back into the atmosphere (point 4), with exiting working fluid mixing to reject  $Q_{41}$ . In a *closed cycle* gas turbine facility the working fluid is continuously recycled by ducting the exit flow (point 4) through a heat exchanger [shown schematically in Fig. 67.1(c)] to reject heat  $Q_{41}$  at (ideally) constant pressure and back to the compressor inlet (point 1). Because of its confined, fixed mass working fluid, the closed cycle gas turbine is *not* an internal combustion engine, so the combustor shown in Fig. 67.1 is replaced with an input heat exchanger supplied by an external source of heat. The latter can take the form of a nuclear reactor, the fluidized bed of a coal combustion process, or some other heat source.

The ideal Brayton cycle thermal efficiency,  $\eta_B$ , can be shown [Bathie, 1984] to be

$$\eta_B = 1 - \frac{T_4}{T_3} = 1 - \frac{1}{r^{(k-1)/k}} \quad (67.1)$$

where the absolute temperatures  $T_3$  and  $T_4$  correspond to the points 3 and 4 in Fig. 67.4,  $k = c_p/c_v$  is the ratio of specific heats for the ideal gas, and  $r$  is the compressor pressure ratio,

$$r = \frac{P_2}{P_1} \quad (67.2)$$

Thus, the Brayton cycle thermal efficiency increases with both the turbine inlet temperature,  $T_3$ , and compressor pressure ratio,  $r$ . Modern gas turbines have pressure ratios as high as 30:1 and turbine inlet temperatures approaching 3000°F (1649°C).

The effect of real (nonisentropic) compressor and turbine performance can be easily seen in the



expression for the net useful work  $W_{\text{net}} = W_{3'4}$  [Huang, 1988],

$$W_{\text{net}} = \eta_T c_p T_3 \left[ 1 - \frac{1}{r^{(k-1)/k}} \right] - \frac{c_p T_1}{\eta_C} [r^{(k-1)/k} - 1] \quad (67.3)$$

where  $c_p$  is the specific heat at constant pressure, and  $\eta_T$  and  $\eta_C$  are the thermal (adiabatic) efficiencies of the turbine and compressor, respectively. This expression shows that for large  $W_{\text{net}}$  one must have high values for  $\eta_T$ ,  $\eta_C$ ,  $r$ , and  $T_3$ . For modern gas turbines,  $\eta_T$  can be as high as 0.92–0.94 and  $\eta_C$  can reach 0.88.

A gas turbine that is configured and operated to closely follow the ideal Brayton cycle (Fig. 67.4) is called a *simple cycle* gas turbine. Most aircraft gas turbines operate in a simple cycle configuration since attention must be paid to engine weight and frontal area.

However, in land or marine applications, additional equipment can be added to the simple cycle gas turbine, leading to increases in thermal efficiency and/or the net work output of a unit. Three such modifications are regeneration, intercooling, and reheating.

Regeneration involves the installation of a heat exchanger through which the turbine exhaust gases (point 4 in Figs. 67.1(b) and 67.4) pass. The compressor exit flow (point 2 in Figs. 67.1(b) and 67.4) is then heated in the exhaust gas heat exchanger before the flow enters the combustor. If the regenerator is well designed—that is, if the heat exchanger effectiveness is high and the pressure drops are small—the cycle efficiency will be increased over the simple cycle value. However, the relatively high cost of such a regenerator must also be taken into account.

Intercooling also involves use of a heat exchanger. An intercooler is a heat exchanger that cools compressor gas during the compression process. For instance, if the compressor consists of a high- and a low-pressure unit, the intercooler could be mounted between them to cool the flow and decrease the work necessary for compression in the high-pressure compressor. The cooling fluid could be atmospheric air or water (e.g., sea water in the case of a marine gas turbine). It can be shown that net work output of a given gas turbine is increased with a well-designed intercooler. Recent studies of gas turbines equipped with extensive turbine convective and film cooling show that intercooling can also allow increases in thermal efficiency by providing cooling fluid at lower temperatures, thereby allowing increased turbine inlet temperatures [ $T_3$  in Eq. (67.1)].

Reheating occurs in the turbine and is a way to increase turbine work without changing compressor work or exceeding the material temperature limits in the turbine. If a gas turbine has a high-pressure and a low-pressure turbine, a reheater (usually another combustor) can be used to "reheat" the flow between the two turbines. Reheat in a jet engine is accomplished by adding an afterburner at the turbine exhaust, thereby increasing thrust, with a greatly increased fuel consumption rate.

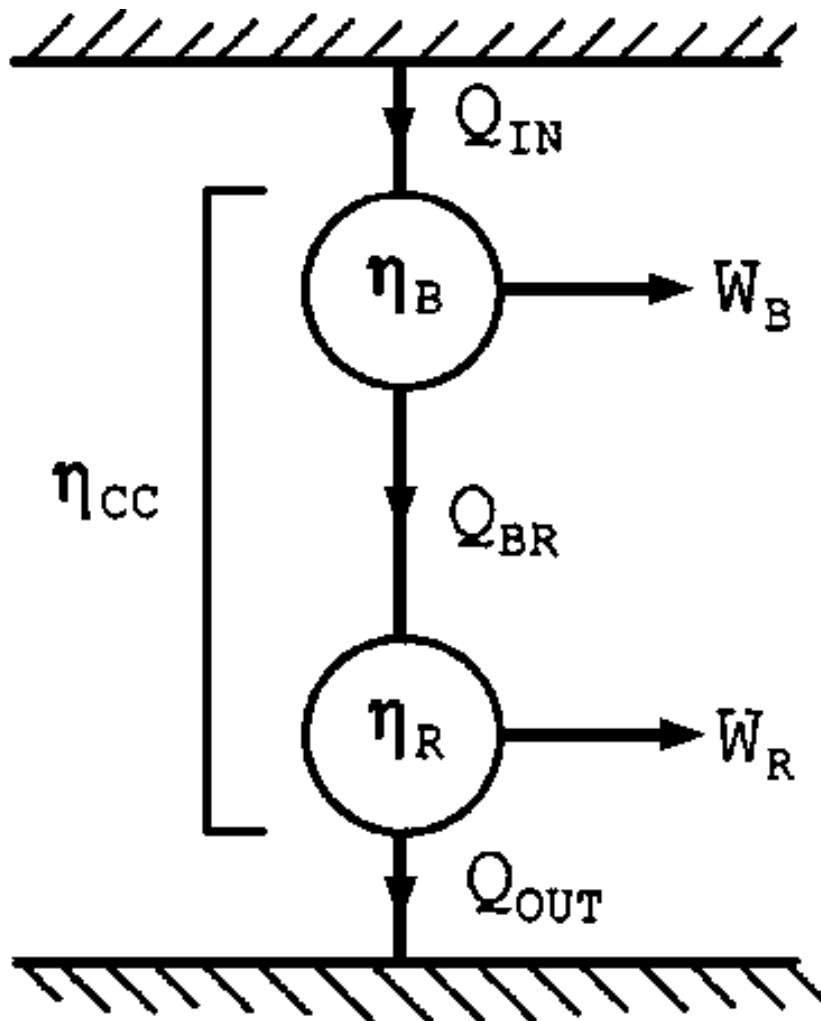
A *combined cycle* gas turbine power plant, frequently identified by the abbreviation CCGT, is essentially an electrical power plant in which a gas turbine provides useful work ( $W_B$ ) to drive an electrical generator. The gas turbine exhaust energy ( $Q_{\text{BR}}$ ) is then used to produce steam in a heat exchanger (called a *heat recovery steam generator*) to supply a steam turbine (see **Chapter 72**) whose useful work output ( $W_R$ ) provides the means to generate more electricity. If the steam is used for heat (e.g., heating buildings), the unit would be called a *cogeneration plant* (see **Chapter 73**). The sketch in Fig. 67.5 is a simplified thermodynamic representation of a CCGT and shows it

to be two heat engines (Brayton and Rankine) coupled in series. The "upper" engine is the gas turbine (represented as a Brayton cycle heat engine), whose energy input is  $Q_{IN}$ . It rejects heat ( $Q_{BR}$ ) as the input energy to the "lower" engine (the steam turbine, represented as a Rankine cycle heat engine). The Rankine heat engine then rejects unavailable energy (heat) as  $Q_{OUT}$  by means of a steam condenser. The combined thermal efficiency ( $\eta_{CC}$ ) can be derived fairly simply [Huang, 1988] and is given as

$$\eta_{CC} = \eta_B + \eta_R - \eta_B \eta_R \quad (67.4)$$

where  $\eta_B = W_B/Q_{IN}$  and  $\eta_R = W_R/Q_{BR}$  are the thermal efficiencies of the Brayton and Rankine cycles, respectively.

**Figure 67.5** Schematic of combined cycle (CC), Brayton (B), and Rankine (R) heat engines, showing work ( $W$ ) and heat ( $Q$ ) inputs and outputs. (Courtesy Global Gas Turbine News.)



Taking  $\eta_B = 40\%$  (a good value for modern gas turbines) and  $\eta_R = 30\%$  (a reasonable value at

typical CCGT conditions), the sum minus the product in Eq. (67.4) yields  $\eta_{CC} = 58\%$ , a value of combined cycle efficiency greater than either of the individual efficiencies.

This remarkable equation gives insight into why CCGTs are so efficient. The calculated value of  $\eta_{CC}$  represents an upper bound on an actual CCGT, since Eq. (67.4) doesn't account for efficiencies associated with transferring  $Q_{BR}$  (duct losses, irreversibilities due to heat transfer, etc.) Actual efficiency values as high as 52 to 54% have been attained with CCGT units during the last few years, thus coming close to values given by Eq. (67.4).

## 67.3 Gas Turbine Components

---

A greater understanding of the gas turbine and its operation can be gained by considering its three major components (Figs. 67.1, 67.2, and 67.3): the compressor, the combustor, and the turbine. The features and characteristics of turbines and compressors will be touched on here only briefly. A longer treatment of the combustor will be given, since environmental considerations make knowledge of its operation and design important to a wider audience.

### Compressors and Turbines

The turbine and compressor components are mated by a shaft, since the former powers the latter. A *single-shaft* gas turbine has but one shaft connecting the compressor and turbine components. A *twin-spool* gas turbine has two concentric shafts, a longer one connecting a low-pressure compressor to a low-pressure turbine (the low spool), which rotates inside a shorter, larger-diameter shaft. The latter connects the high-pressure turbine with the higher-pressure compressor (the high spool), which rotates at higher speeds than the low spool. A *triple-spool* engine would have a third, intermediate-pressure compressor-turbine spool.

Gas turbine compressors can be centrifugal, axial, or a combination of both. Centrifugal compressors (radial outflow) are robust, generally cost less, and are limited to pressure ratio of 6 or 7 to 1. They are found in early gas turbines or in modern, smaller gas turbines.

The more efficient, higher-capacity axial flow compressor is used on most gas turbines (e.g., Figs. 67.2 and 67.3). An axial compressor is made up of a number of stages, each stage consisting of a row of rotating blades (airfoils) and a row of stationary blades (called *stators*) configured so the gas flow is compressed (adverse or unfavorable pressure gradient) as it passes through each stage. It has been said that compressor operation can founder upon a metaphoric rock, and that rock is called *stall*. Care must be taken in compressor operation and design to avoid the conditions that lead to blade stall, or flow separation. The collective behavior of blade separation can lead to compressor stall or surge, which manifests itself as an instability of gas flow through the entire gas turbine.

Turbines are generally easier to design and operate than compressors, since the flow is expanding in an overall favorable pressure gradient. Axial flow turbines (Figs. 67.2 and 67.3) will require fewer stages than an axial compressor for the same pressure change magnitude. There are some smaller gas turbines that utilize centrifugal turbines (radial inflow) but most utilize axial turbines (e.g., Fig. 67.3).

Turbine design and manufacture is complicated by the need to ensure turbine component life in



the hot gas flow. The problem of ensuring durability is especially critical in the first turbine stage, where temperatures are highest. Special materials and elaborate cooling schemes must be used to allow metal alloy turbine airfoils that melt at 1800 to 1900°F (982 to 1038°C) to survive in gas flows with temperatures as high as  $T_3 = 3000^\circ\text{F}$  (1649°F).

## Combustors

A successful combustor design must satisfy many requirements and has been a challenge from the earliest gas turbines of Whittle and von Ohain. The relative importance of each requirement varies with the application of the gas turbine, and, of course, some requirements are conflicting, requiring design compromises to be made. The basic design requirements can be classified as follows:

1. High combustion efficiency at all operating conditions.
2. Low levels of unburned hydrocarbons and carbon monoxide, low oxides of nitrogen at high power and no visible smoke.
3. Low pressure drop. Three to four percent is common.
4. Combustion stability limits must be wider than all transient operating conditions.
5. Consistently reliable ignition must be attained at very low ambient temperatures, and at high altitudes (for aircraft).
6. Smooth combustion, with no pulsations or rough burning.
7. A small exit temperature variation for good turbine life requirements.
8. Useful life (thousands of hours), particularly for industrial use.
9. Multifuel use. Characteristically, natural gas and diesel fuel are used for industrial applications and kerosene for aircraft.
10. Length and diameter compatible with engine envelope.
11. Designed for minimum cost, repair, and maintenance.
12. Minimum weight (for aircraft applications).

A combustor consists of at least three basic parts: a casing, a flame tube, and a fuel injection system. The casing must withstand the cycle pressures and may be a part of the structure of the gas turbine. It encloses a relatively thin-walled flame tube, within which combustion takes place, and a fuel injection system. Sometimes the diffusing passage between the compressor and the combustor is considered a part of the combustor assembly as well.

The flame tube can be either tubular or annular. In early engines the pressure casings were tubular and enclosed the flame tubes. This configuration is called a *can combustor*. One or more can be used on an engine. If an annular casing encloses can-type flame tubes, this is called a *can-annular* or *cannular* combustor. An annular combustor consists of both an annular casing and an annular flame tube and is a common design used in aircraft gas turbines. Industrial engines often use can or cannular combustors because of their relative ease of removal.

Fuel injectors can be pressure-atomizing, air-atomizing, or air blast types. The specific styles vary widely and are often designed and manufactured by specialist suppliers. For industrial engines an injector may be required to supply atomized liquid fuel, natural gas, and steam or water (for nitric oxide reduction).

Compared to other prime movers, gas turbines are considered to produce very low levels of combustion pollution. The gas turbine emissions of major concern are unburned hydrocarbons, carbon monoxide, oxides of nitrogen ( $\text{NO}_x$ ), and smoke. Although the contribution of jet aircraft to atmospheric pollution is less than 1% [Koff, 1994], jet aircraft emissions injected directly into the upper troposphere have doubled between the latitudes of 40 to 60 degrees north, increasing ozone by about 20%. In the stratosphere, where supersonic aircraft fly,  $\text{NO}_x$  will deplete ozone. Both effects are harmful, so  $\text{NO}_x$  reduction in gas turbine operation is a challenge for the 21st century [Schumann, 1993].

## Defining Terms

**Aeroderivative:** An aviation propulsion gas turbine (jet engine) used in a nonaviation application (e.g., an electric power plant).

**Brayton cycle:** Ideal gas cycle that is approximated by a gas turbine simple cycle (see Fig. 67.4).

**Combined cycle:** The combination of a gas turbine power plant and a steam turbine power plant. The gas turbine exhaust is used as heat input to generate steam.

**Gas generator:** The compressor, combustor, and turbine components of a gas turbine.

**Jet engine:** An aviation gas turbine used for aircraft propulsion.

**Rankine cycle:** Ideal gas cycle that is approximated by a simple steam turbine power plant cycle.

## References

- Bathie, W. W. 1984. *Fundamentals of Gas Turbines*. John Wiley & Sons, New York.
- Huang, F. F. 1988. *Engineering Thermodynamics*, 2nd ed. Macmillan, New York.
- Koff, B. L. 1994. Aircraft gas turbine emissions challenge. *J. Eng. Gas Turbines Power*. 116:474–477.
- Schumann, U. 1993. On the effect of emissions from aircraft engines on the state of the atmosphere. *Fuels and Combustion Technology for Advanced Aircraft Engines*. AGARD Conference Proceedings 536, Sept.

## Further Information

- Cohen, H., Rogers, G. F. C., and Saravanamuttoo, H. I. H. 1987. *Gas Turbine Theory*, 3rd ed. Longman Scientific and Technical, Essex, England.
- Horlock, J. H. 1993. *Combined Power Plants*. Pergamon Press, Oxford, England.
- Kerrebrock, J. L. 1980. *Aircraft Engines and Gas Turbines*. MIT Press, Cambridge, MA.
- Lefebvre, A. H. 1983. *Gas Turbine Combustion*. McGraw-Hill, New York.
- Northern Research and Engineering Corporation. 1964. *The Design and Performance Analysis of Gas Turbine Combustion Chambers*; and *The Design and Development of Gas Turbine Combustors*. 1980. (They are popularly called the "Orange Books.")
- Transactions of the ASME*, Published quarterly by the American Society of Mechanical Engineers, New York. Two ASME gas turbine journals are *Journal of Turbomachinery* and *Journal of Engineering for Gas Turbines and Power*.

Woodall, D. M. "Nuclear Power Systems"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Nuclear Power Systems

---

## 68.1 Nuclear Power Applications

Terrestrial Nuclear Power • Space Nuclear Power

## 68.2 Nuclear Power Fundamentals

Fission • Radioactivity • Criticality • Radiation Protection • Nuclear Fuel Cycle • Reactor Types • Reactor Operations, Licensing, and Regulation • Radioactive Waste Management

## 68.3 Economics of Nuclear Power Systems

Alternative Sources of Electrical Power • Next-Generation Commercial Systems

### David M. Woodall

*University of Idaho*

Nuclear power systems use the controlled release of **nuclear energy** as a heat source for the generation of electricity or for direct thermal heat. The advantages of nuclear power include extremely high power density and large energy release per unit mass of fuel. The disadvantages include the production of radioactive materials requiring careful handling and disposal, the operation of components in a severe radiation environment, and power plant control characteristics distinct from traditional power sources.

## 68.1 Nuclear Power Applications

---

There are both terrestrial and space applications of nuclear power systems: commercial nuclear power production of electricity; nuclear propulsion of ocean-going vessels; and high-reliability electricity and heat for remote locations, including energy sources for space satellites and space science probes.

### Terrestrial Nuclear Power

#### Commercial Nuclear Electric Power

The commercial application of nuclear energy followed closely on the military release of nuclear energy in the nuclear weapons developed during World War II. The first commercial nuclear power plant was completed in the late 1950s, and there are presently (in 1995) 109 operating nuclear power plants in the U.S., generating 22% of the national electricity supply. Worldwide there are 430 operating nuclear power plants, with many countries obtaining more than 40% of their electricity from nuclear power, including France, Belgium, Sweden, Hungary, and South Korea. The principal types of commercial nuclear power systems in use include the **pressurized**

**water reactor (PWR)** and the **boiling water reactor (BWR)**. A third type of reactor being actively developed by European and Japanese utilities is the **liquid metal fast breeder reactor (LMFBR)**. The former Soviet block countries operate two Russian designs, a PWR-like reactor called the VVER and a graphite-moderated, water-cooled thermal reactor, the **pressure-tube graphite reactor (PTGR)**, also called the RBMK. The RBMK is the reactor type which experienced the Chernobyl Unit 2 reactor accident.

The design and operation of commercial nuclear power systems require the use of complex nuclear and thermal-hydraulic design codes incorporating basic engineering principles and engineering correlations. While scoping analysis principles are outlined in the following sections, the serious practitioner is encouraged to read the reference materials and the current technical literature.

### **Nuclear Propulsion**

The U.S. and Russian navies have a major commitment to nuclear propulsion for their surface and submarine fleets. Nuclear power enables a vessel to travel long distances at high speed without refueling and supports an extended submerged operation, as no oxygen is consumed for combustion. The U.S. has over 130 submarines, nine cruisers, and four aircraft carriers which are nuclear powered. The Russian fleet is similarly equipped with nuclear propulsion systems. The pressurized water reactor (PWR) commercial reactor type was demonstrated by the naval nuclear propulsion program. The storage of spent reactor cores as high-level waste (HLW) is a technical issue for naval operations.

### **Space Nuclear Power**

Power sources for space operations are limited by the high cost of launching a system to orbit. There is a high premium placed on system reliability, due to the inaccessibility of systems in use and the high cost of alternative power systems. The options for space power include the conversion of solar energy to electricity. Solar electricity has two principal disadvantages: the requirement for battery backup power during periods of unavailability of sunlight and the decreased solar radiation for deep space missions outward from the sun. Nuclear space power systems do not suffer from either of these limitations, but are limited by public acceptance of the launch of nuclear materials, due primarily to the perception of the environmental risk of launch accident-initiated dispersal of radioactive materials.

### **Radioisotope Systems**

The nuclear power system which has found the most application in U.S. space missions is the radioisotope thermoelectric generator (RTG). Electricity is produced by a thermoelectric converter, powered by the thermal energy from the decay of a long-lived, alpha-emitting isotope of plutonium,  ${}_{94}\text{Pu}^{238}$ , with a half-life of 88 years.

### **Space Reactors**

Small nuclear reactors were used in the former Soviet Union's space power program. Reactors powering thermoelectric energy conversion systems were routinely used in earth observation

satellites. The U.S. space program launched and operated in space only one nuclear reactor, SNAP-10A, during the early space program.

## 68.2 Nuclear Power Fundamentals

---

### Fission

The basic source of energy in nuclear power is the fissioning of the nucleus of an isotope of uranium or plutonium. The energy released per fission depends on the particular isotope, and is approximately  $3.2 \times 10^{-11}$  J (joules) for the thermal neutron fission of  ${}_{92}\text{U}^{235}$ , the most common nuclear fuel. Thus approximately  $3.1 \times 10^{10}$  fissions per second are required to produce 1 watt of thermal energy. Typical commercial nuclear power systems produce 1000 MW of electrical power at a thermal efficiency of about 33%.

The energy released from the fission of a nucleus is distributed among various products of the process, including nuclear fragments, neutrons, and gamma rays. An average fission releases two or three excess neutrons. These excess neutrons are available to initiate further fissions; hence nuclear fission in a suitable assemblage of materials can be a self-sustaining process. Control of the power of a nuclear power system depends on the control of the neutron population, typically through the use of neutron absorbers in movable structures called **control rods**.

In most fissions, approximately 80% of the energy released is carried by two large nuclear fragments, or fission fragments. The fission fragments have very short ranges in the fuel material and come to rest, capturing electrons to become fission product atoms. Fission products are typically radioactive, having a spectrum of radioactive decay products and half-lives. A small fraction of the fission products are neutron rich and decay by neutron emission, adding to the neutron population in the reactor. Such fission products are termed *fission product precursors*, and their radioactive decay influences the time-dependent behavior of the overall neutron population in the reactor. These **delayed neutrons** play an essential role in the control of the reactor power.

The thermal energy deposited in nuclear fuel, coolant, and structural material is dependent on the type of radiation carrying the energy. Neutrons are absorbed in stable nuclei, causing the **transmutation** to a radioactive element. Such a capture event releases an energetic photon, a capture gamma, which deposits its energy in the surrounding material. Subsequent radioactive decay of these radionuclides releases additional radiation energy in the form of decay gamma and decay beta radiation. All of the energy released in beta decay is not available to be locally deposited, due to the extremely long range of the antineutrinos which share the energy available with the beta radiation. [Table 68.1](#) delineates the distribution of energy from a typical fission, including information on the heat produced in the reactor from each source. The convenient unit of energy used is the MeV (mega-electron-volt), where 1 MeV is equal to  $1.6 \times 10^{-13}$  J.

**Table 68.1** Energy Distribution among Products of Fission

Energy Source	Energy (MeV)	Heat Produced
Fission fragments	168	168
Fast neutrons	5	5

Prompt gammas	7	7
Decay gammas	7	7
Capture gammas	—	5
Decay betas	20	8
Total	207	200

---

## Radioactivity

Elements which undergo radioactive decay are called radioisotopes. A radioisotope can be characterized by the energetic radiation which is emitted in the decay process and by the probability of decay per unit time, the decay constant,  $\lambda$ . The products from a radioactive decay include the daughter isotope and one or more radiation emissions. The three most common radiation emissions are alpha, beta, and gamma radiation. Alpha radiation consists of an energetic helium nucleus, beta radiation is an energetic electron, and gamma radiation is a photon of electromagnetic radiation released from nuclear energy level transitions. The activity of a radioactive substance,  $A$ , is the number of decays per second and the number of radiations emitted per second, assuming a single radiation per decay:

$$A = \lambda N = \lambda N_0 e^{-\lambda t} \quad (68.1)$$

where  $t$  is the time in seconds and  $N_0$  is the initial number of atoms present at  $t = 0$ . The unit of activity is the becquerel (Bq), with 1 Bq corresponding to 1 decay per second (dps). The non-SI unit in common use for activity is the curie (Ci), with  $1\text{Ci} = 3.7 \times 10^{10}$  dps. The half-life of an isotope is the time for half of the atoms of a given sample to decay,  $T_{1/2}$ . The decay constant and half-life are related through the following formula:

$$T_{1/2} = \ln 2 / \lambda \quad (68.2)$$

## Criticality

The assemblage of nuclear fuel and other material into a device which will support sustained fission is called a **nuclear reactor**. The central region of the reactor is involved in fission energy release, containing the nuclear fuel and related structural and energy transfer materials, and is called the **reactor core**. The typical nuclear fuel is an oxide of uranium which has been enriched in the fuel isotope,  ${}_{92}\text{U}^{235}$ . While a number of fuel forms are in use, the predominant form is compressed and sintered pellets of the ceramic oxide  $\text{UO}_2$ . These pellets are stacked in cladding tubes of **zircaloy**, forming fuel pins, which are held in rectangular arrays of fuel elements by structural guides. The fuel elements are arranged in a rectangular array by core guides to form the reactor core. Coolant passes axially along the fuel elements and through the core.

The neutrons born in fission are typically high-energy neutrons, with an average energy of about  $10^{-13}$  J. However, the likelihood of producing fission is higher for low-energy neutrons, those in

collisional (thermal) equilibrium with the reactor. Such **thermal neutrons** have energies of about  $4 \times 10^{-21}$  J. Neutrons in the reactor lose energy through elastic and inelastic collisions with light nuclei, called **moderator** materials. Typical materials used to moderate the energy of the neutrons are hydrogen in the form of water, deuterium in water, and carbon in graphite.

### Reactor Kinetics

Neutrons are born in the fission process at high energy. Such **fast neutrons** travel through the core scattering from nuclei and losing energy until they are absorbed, leak from the reactor core, or are thermalized. Thermal neutrons diffuse through the reactor core, gaining or losing energy in thermal equilibrium with the nuclei of core materials until they are absorbed or leak from the system. Those neutrons absorbed in nuclear fuel may cause fission, leading to additional fast neutrons. The behavior of neutrons in the reactor can be described in terms of a neutron life cycle, with the ratio of neutrons in one generation to those in the previous generation given by the multiplication constant  $k$ . The time variation of the reactor power is a function of the multiplication constant, with the reactor condition termed subcritical, critical, or supercritical, depending on the value of  $k$  for the present configuration, as shown in [Table 68.2](#).

**Table 68.2** Critical Constant versus Reactor Power Behavior

Condition of Reactor	Multiplication Constant	Reactor Power
Critical	$k = 1$	Steady
Supercritical	$k > 1$	Increasing
Subcritical	$k < 1$	Decreasing

The value of  $k$  is dependent on geometric and material properties in the reactor and can be varied by introducing or withdrawing absorbing materials, or control rods. There are six factors which make up the average value of  $k$  for the reactor:

$$k = P_{fnl} P_{tnl} \varepsilon \eta f p \quad (68.3)$$

The first two factors are geometric, relating to the leakage of neutrons out of the core of the reactor to be lost by absorption in surrounding material. The fast nonleakage probability,  $P_{fnl}$ , is the probability that a fast neutron born in fission does not leak from the reactor before it is thermalized. The thermal nonleakage probability,  $P_{tnl}$ , is the probability that a thermal neutron does not leak from the reactor prior to being absorbed. These two terms are a function of the geometry of the core and can be approximated as follows:

$$P_{fnl} P_{tnl} = 1/(1 + M^2 B^2) \quad (68.4)$$

where  $M^2$  is the neutron migration length and  $B^2$  is the geometric buckling. The neutron migration length is

$$M^2 = L^2 + \tau \quad (68.5)$$



$\tau$  is the Fermi age, the mean square distance that a fast neutron travels in the reactor prior to becoming thermalized.  $L^2$  is the mean square distance that a thermal neutron diffuses through the reactor prior to being absorbed.  $\tau$  can be calculated from scattering theory, but this is beyond the scope of this treatment.  $L^2$  is given in terms of the thermal neutron diffusion coefficient and the **mean free path** of thermal neutrons to absorption,  $\lambda_a$ :

$$L^2 = \sqrt{D\lambda_a} \quad (68.6)$$

The next four factors are based on the average properties of the materials which make up the reactor core and are independent of geometry:

$\varepsilon$  is the fast fission factor, which is the number of neutrons produced in fission from all neutron energies divided by the number of neutrons produced by fissions from thermal neutrons. For a typical thermal reactor this parameter ranges from 1.0 to 1.05.

$\eta$  is the number of fast neutrons produced per thermal neutron absorbed in the fuel. Thus  $\eta$  is a function of the mean free path of thermal neutrons to fission, the absorption mean free path, and the number of neutrons produced in an average thermal fission,  $\nu$ . For a typical PWR,  $\nu$  has a value of about 2.5.

$$\eta = \nu(\lambda_a^F / \lambda_f) \quad (68.7)$$

$f$  is the thermal utilization, which is the fraction of absorbed thermal neutrons which are absorbed in the fuel material; thus it is a ratio including the mean free path for absorption in fuel,  $\lambda_a^F$ , and the mean free path including all absorptions,  $\lambda_a^T$ :

$$f = \lambda_a^T / \lambda_a^F \quad (68.8)$$

$p$  is the resonance escape probability, the likelihood that a neutron loses energy to the thermal energy range without being captured by a nuclear resonance, which leads to loss of the neutron without producing fission. The resonance escape probability depends substantially on the heterogeneous geometry of the reactor, because neutrons leaving the fuel can thermalize in a moderator prior to reentering the fuel by diffusion. **Neutron transport theory** must be used to compute  $p$ .

In practice, the reactivity,  $\rho$ , is often used instead of the multiplication constant, where

$$\rho = (k - 1)/k \quad (68.9)$$

Changes in core configuration or material properties can be characterized as reactivity insertions, where  $\rho$  equal to zero corresponds to steady power operation, a positive reactivity insertion (withdrawing an absorber control rod, for example) leads to a power increase, and a negative reactivity insertion (inserting an absorber control rod) leads to a power decrease.

## Reactor Thermal Hydraulics

The transfer of thermal energy through a coolant to an energy conversion system is more complex for a nuclear power system than for a combustion-fueled power system. Most of the fission energy is deposited in the nuclear fuel, typically a high-temperature ceramic with poor conductivity. Transfer of heat to the coolant occurs through conduction in the fuel pellet; conduction, radiation, and convection in the gap between the pellet and the cladding; conduction through the cladding; and convection to the coolant, flowing in channels along the fuel pins. Material properties of the fuel change over time with fuel **burn-up**, due to fission product entrainment, radiation damage, and thermal cycling, which leads to pellet swelling and cracking.

The balance of plant for a nuclear power plant is similar to that of a conventional power plant; for additional details refer to **Chapter 69**, on power plants, and **Chapter 72**, on turbines. However, the presence of coolant-borne radioisotopes may place additional constraints on the operation of steam generators and primary coolant pumps in a nuclear system. Thermal energy management systems are the major source of operational and maintenance costs for a commercial nuclear power system.

## Radiation Protection

Analysis of the radiation hazard from nuclear energy systems must include the effect of such systems on materials, biological systems, and the environment. The radiation environment of the fission process requires shielding of radiation-sensitive systems from the radiation produced during reactor operation, principally neutrons and gammas. The fission products produced and the **transuranic** (TRU) isotopes require environmental isolation and long-term storage.

Radiation protection concepts include methods of dealing with both external and internal sources of radiation. National and international standards on radiation protection are set by the National and International Commissions on Radiation Protection (NCRP and ICRP). Standards are enforced through guidelines and procedures promulgated by federal agencies, such as the U.S. Nuclear Regulatory Commission (NRC) and the U.S. Environmental Protection Agency (EPA). The general public and the radiation worker are protected from exposure to radiation at levels known to cause harm.

Radiation damage occurs due to the deposition of energy in materials, which leads to ionization and atomic displacement damage. The damage to biological systems is a function of the type of radiation, and is due to the production of oxidizing radicals following radiation's ionizing effect. Each source of radiation is characterized by its relative biological effectiveness (*RBE*), which is the damage of a given amount of the particular radiation relative to the damage of the same energy deposited by 100 keV x-rays. The unit of absorbed dose is the gray, where 1 gray (Gy) = 1 J (absorbed energy) / kg (mass of material). The unit of radiation exposure is the sievert (Sv), where

$$\text{Dose (Sv)} = \text{Absorbed dose (Gy)} \times RBE \quad (68.10)$$

The radiation exposure principle which is applied in all occupational exposures is "as low as reasonably achievable" (ALARA). Thus the actual average occupational exposure and public exposure due to nuclear power operations is much less than allowed (see [Table 68.3](#)).

**Table 68.3** Annual Radiation Exposures

Occupational limit	50 mSv
General public limit	1.0 mSv
Average background exposure from nonnuclear power sources [natural and medical radiation (dental x-rays and medical treatment)]	3.5 mSv

Radiation shielding for external sources of radiation relies on three principles: time, distance, and shielding. The exposure to radiation can be minimized by: (a) minimizing the time of the exposure to the source; (b) maximizing the distance from the source; (c) interposing absorbing materials, or shielding materials, between the source and the individual; or (d) some appropriate combination of *a*, *b*, and *c*. Exposure to radiation from internal sources is controlled by minimizing the ingestion or inhalation of such sources. There are regulations on the allowable maximum permissible concentration (MPC) of each radioisotope in air and water, based on daily consumption and keeping lifetime exposures below acceptable limits.

## Nuclear Fuel Cycle

The nuclear fuel cycle encompasses the following processes from the exploration for uranium ore to the final disposition of the nuclear waste in a geological repository.

*Exploration:* Uranium and thorium are naturally occurring elements which exist in the earth's crust. Geological evaluation of the potential of a region is followed by monitoring for the low levels of natural radiation produced by the ores.

*Mining:* Both open-pit and underground mining is done for uranium ore (pitchblende). Major suppliers of uranium ore include Australia, Canada, Russia, and Africa.

*Milling:* Uranium is separated from the ore by physical and chemical means and converted to  $\text{U}_3\text{O}_8$  (yellowcake).

*Enrichment:* Natural uranium contains 0.711% (by weight) fuel isotope  ${}_{92}\text{U}^{235}$ . The remainder (> 99%) is  ${}_{92}\text{U}^{238}$ , which can be used to breed fuel, but is not itself usable as fuel in a thermal reactor. Some process must be used to increase the  ${}_{92}\text{U}^{235}$  concentration in order to make the fuel usable in light water reactors. Enrichment methods include gaseous diffusion of  $\text{UF}_6$  gas, centrifugal enrichment, and laser enrichment.

*Fabrication:* The enriched  $\text{UF}_6$  must be converted to the ceramic form ( $\text{UO}_2$ ) and encased in a cladding material. Fuel pins are bound together into fuel assemblies.

*Use in reactor:* Fuel assemblies are placed in the core of the reactor and over a period of reactor operation fission energy is extracted. A typical reactor operation time is 12 to 18 months. Fuel management often dictates installing a third of the core as unburned fuel assemblies and discharging a third of the core as spent fuel at each refueling operation.

*Interim storage:* The spent fuel removed from the core must be cooled in water-filled storage pits for six months to a year, in order for the residual heat and radioactivity to decay.

*Reprocessing:* Cooled spent fuel can either be shipped to a fuel-reprocessing facility or sent to a permanent geological repository for long-term storage.

*Waste disposal:* HLW storage includes both the storage of spent fuel elements in a geological repository and the storage of vitrified waste from the reprocessing stream in such a repository. An interim monitored retrievable storage facility may be necessary prior to the licensing and operation of a permanent storage facility.

*Transportation:* The transporting of nuclear materials poses special hazards, due to the potential radiological hazard of the material being shipped. Specially designed shipping casks have been developed with government funding that are able to withstand a highway accident and resulting fire without the release of radionuclides to the environment.

*Nuclear safety:* Nuclear power requires special consideration of criticality safety, in order to keep nuclear fuel materials from accidentally coming into a critical configuration outside the confined environment of the reactor core.

*Nuclear material safeguards:* Nuclear materials, in particular those materials which are key to the development of a nuclear weapon, must be protected from diversion into the hands of terrorists. Special safeguard procedures have been developed to control special nuclear material (SNM).

## Reactor Types

The nuclear power systems in common use throughout the world are listed in [Table 68.4](#), along with information on the fuel form, the coolant type, the moderator, and the typical size.

**Table 68.4** Common Commercial Nuclear Power Systems

Reactor Type	Output (thermal/electric)	Fuel Form	Coolant/Moderator
Pressurized water reactor (PWR)	3400 MWth/1150 MWe	UO <sub>2</sub> pellets, 2–4% enriched uranium	H <sub>2</sub> O/H <sub>2</sub> O
Boiling water reactor (BWR)	3579 MWth/1178 MWe	UO <sub>2</sub> pellets, 2–4% enriched uranium	H <sub>2</sub> O/H <sub>2</sub> O
Pressure tube graphite reactor (PTGR)	3200 MWth/950 MWe	UO <sub>2</sub> pellets, 1.8–2.4% enriched uranium	H <sub>2</sub> O/graphite
Liquid metal fast breeder reactor (LMFBR)	3000 MWth/1200 MWe	Mixed UO <sub>2</sub> and PuO <sub>2</sub> pellets, 10–20% plutonium	Liquid sodium/none (fast reactor)
Pressure tube heavy water reactor (CANDU)	2180 MWth/648 MWe	UO <sub>2</sub> pellets, natural uranium	D <sub>2</sub> O/D <sub>2</sub> O

## Reactor Operations, Licensing, and Regulation

The licensing of nuclear power systems in the U.S. is the responsibility of the NRC. State public utility commissions are involved in general oversight of electric power generation by electric utilities, including those with nuclear power plants. Industry associations are involved in the development and promulgation of best industry practices. Two such organizations are the Institute

for Nuclear Power Operations (INPO) and the Nuclear Energy Institute (NEI). Similar organizations exist to support international nuclear power operations, including the International Atomic Energy Agency (IAEA), a branch of the United Nations, and the World Association of Nuclear Operators (WANO), an international nuclear power utility industry association.

Following use in a reactor, the spent fuel elements are depleted in the useful nuclear fuel isotopes and contain substantial quantities of fission products and transuranic elements. While they are no longer useful for production of electricity, special handling is required due to the residual radioactivity. The decay of fission products continues to provide an internal heat source for the fuel element, which requires cooling during an interim storage period. The heat source is a function of the duration of reactor operation, but following steady power operation the time dependence has the following variation:

$$P(t) = 0.066P_0[t^{-1/5} - (t + t_0)^{-1/5}] \quad (68.11)$$

where  $P_0$  is the power level during operation for time  $t_0$ , and  $t$  is the time in seconds after shutdown. Due to the weak time dependence, the thermal energy generated by fuel elements requires substantial cooling long after reactor shutdown.

## Radioactive Waste Management

The radioactive wastes generated by nuclear power systems must be handled responsibly, minimizing the radiation exposure to employees and the general public, while providing long-term environmental isolation. Radioactive wastes are characterized as low-level wastes (LLW) or high-level wastes (HLW).

LLW consists of materials contaminated at low levels with radioisotopes but requiring no shielding during processing or handling. LLW is generated not only by nuclear power plant operations but also from medical procedures, university research activities, and industrial processes which use radioisotopes. LLW is judged to have minimal environmental impact, and the preferred method of disposal is shallow land burial in approved facilities. LLW is handled by individual states and state compacts, under authority granted by the U.S. Congress in the Low-Level Radioactive Waste Policy Act of 1980.

HLW includes spent nuclear fuel and the liquid residue of nuclear fuel reprocessing. There is presently (1995) no commercial reprocessing of spent nuclear fuel in the U.S., based on U.S. nonproliferation policy. However, there is a substantial amount of HLW in the U.S. which was generated by the nuclear weapons production program. This is currently stored at U.S. DOE laboratory sites and will require substantial processing prior to long-term storage. European countries and Japan have active programs for the reprocessing of spent reactor fuel. In the U.S. under the Nuclear Waste Policy Act of 1982, HLW is scheduled for long-term geological repository storage, with the U.S. Department of Energy assigned to manage both storage and disposal programs. A national geological repository is being investigated at Yucca Mountain, Nevada. It is not expected to be ready for permanent storage of HLW until 2010. In the interim, electric utilities are using on-site wet-well or dry-cask storage of their spent fuel. A national monitored retrievable storage (MRS) facility has been proposed to meet utility needs until the

national permanent storage facility is available.

## **68.3 Economics of Nuclear Power Systems**

---

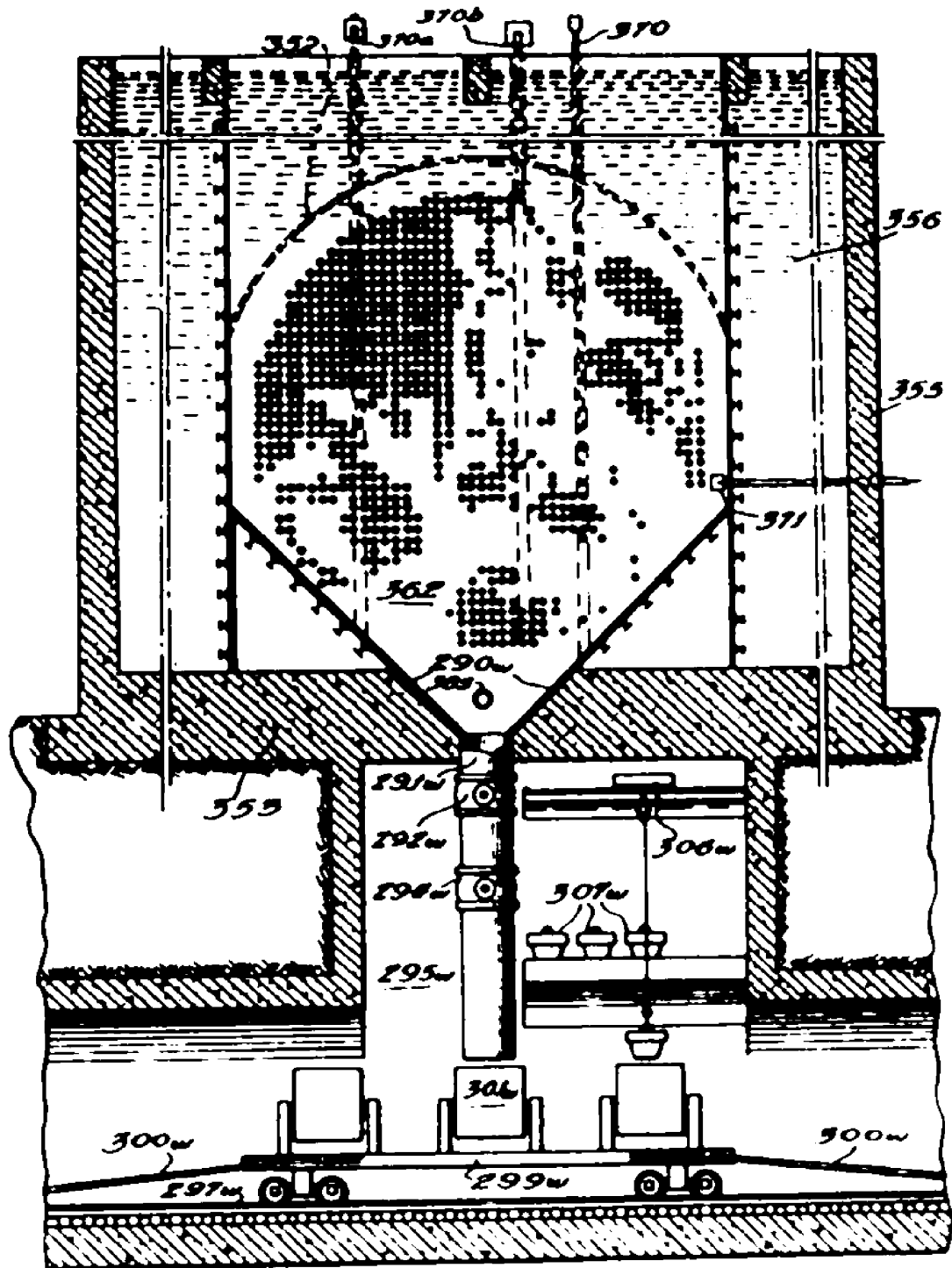
Key in the economic competitiveness of nuclear power is the ability to minimize the operational costs of such plants and to increase the plant availability. The average capacity factor for U.S. nuclear power plants in 1993 was 74.6%, the highest in history. Other factors influencing the economics are shortened construction schedules and single-step construction and operating licenses from the NRC.

### **Alternative Sources of Electrical Power**

All sources of energy produce environmental impacts. The ability of nuclear power systems to compete in a deregulated electricity supply environment will depend on the environmental impact of alternative sources. Nuclear power systems avoid the emission of CO<sub>2</sub> and other greenhouse gases into the atmosphere. In 1993, U.S. commercial nuclear power plants avoided the introduction of 133 million metric tons of carbon dioxide, 4.7 million tons of sulfur dioxide, and 2.2 million tons of nitrogen oxides which would have resulted from alternative fossil-fueled plants. The further growth of commercial nuclear power will also depend on public perception of risk, as well as the economics of central power station alternatives. Handling, storage, and transportation of nuclear waste continue to be issues of public perception, rather than technology, which is well established. The public continues to confuse the nuclear waste associated with national nuclear weapons production activities with wastes from commercial nuclear power.

### **Next-Generation Commercial Systems**

Under the Energy Policy Act of 1992 the NRC will preapprove nuclear system designs, resolve siting issues, and issue single construction and operation licenses for next-generation nuclear power systems. Each design will meet stringent U.S. nuclear safety standards. Construction cycles will be shortened to four to five years, making the economics of nuclear power among the most competitive. The NRC will approve standard plant designs under regulation 10 CFR Part 52. Two designs have received final design approval (1995), ABB Combustion Engineering's System 80+ PWR design and General Electric's Advanced Boiling Water Reactor (ABWR) design, while other vendor's designs are being evaluated. The NRC design certification will allow electric utilities to use these designs in applications to build and operate advanced nuclear power systems with assurance of the ability to operate the plant and recover their investment.



## NEUTRONIC REACTOR

Enrico Fermi and Leo Szilard

Patented May 17, 1955

#2,708,656

Fermi and Szilard conducted most of the work for what we now know as the Nuclear Reactor from 1938 to 1942. The application for this patent was filed in December 1944. They assigned their patent rights to the U.S. Atomic Energy Commission.

U.S. Patent #2,708,656 describes the nuclear fission chain reactions of uranium and construction of facilities (air- or water-cooled) for sustaining these reactions on a controlled basis. Fermi and Szilard describe many purposes for such a reactor, including generation of various subatomic particles, gamma rays, and radioactive isotopes, only briefly mentioning the creation of steam for electric power generation, the purpose we are now most familiar with in our everyday lives.

## Defining Terms

**Boiling water reactor (BWR):** Allows the liquid coolant to boil in the core, and the primary loop includes steam separators and dryers and the steam generator.

**Burn-up:** The quantity of energy released, in megawatt-days per metric ton of fuel; also the change in quantity of nuclear fuel due to fissioning.

**CANDU:** A reactor using natural uranium with deuterated water for both cooling and moderation, the preferred reactor of the Canadian nuclear power program. No fuel enrichment is required, but deuterium must be separated from water.

**Control rods:** Devices including elements which strongly absorb or scatter neutrons, used to control the neutron population in a nuclear reactor.

**Delayed neutrons:** Neutrons born from the decay of neutron-rich fission products. The emission of these neutrons is controlled by the radioactive decay of their parent fission product; hence they are delayed from the fission event.

**Fast neutrons:** Neutrons born in fission, typically with energies well above the thermal range.

**Liquid metal fast breeder reactor (LMFBR):** A reactor that operates on a fast neutron spectrum and breeds additional nuclear fuel from nonfuel  ${}_{92}\text{U}^{238}$ . Sodium metal is the liquid coolant, and a secondary sodium coolant loop transfers energy through a heat exchanger to a tertiary loop of water, which generates steam for a turbine.

**Mean free path:** The average distance which a neutron of a given energy travels through matter prior to an interaction of the type specified (i.e., scattering or absorption).

**Moderator:** A material which scatters neutrons rather than absorbing them, typically a light element which causes neutrons to lose energy efficiently during scattering.

**Nuclear energy:** The release of energy from fissioning the nucleus, as opposed to fossil energy, which releases energy from chemical combustion.

**Nuclear reactor:** An assemblage of materials designed for the controlled release of nuclear energy.



**Neutron transport theory:** The theory which includes the detailed spatial and time behavior of neutrons traveling through matter; transport analysis requires detailed understanding of the nuclear properties of materials.

**Pressure-tube graphite reactor (PTGR):** A reactor in which fuel assemblies are contained in water-filled tubes which provide flowing cooling. Those tubes penetrate a graphite block, with the graphite providing neutron moderation.

**Pressurized water reactor (PWR):** The PWR has a pressurized primary loop which transfers heat through a heat exchanger to a secondary water loop which includes the steam generator. Heat transfer in the core occurs in the subcooled phase.

**Reactor core:** The central region of a nuclear reactor, where fission energy is released.

**Thermal neutrons:** Neutrons which have kinetic energies comparable to the thermal energy of surrounding materials, and hence are in collisional equilibrium with those materials.

**Transmutation:** The change of a nucleus to a different isotope through a nuclear reaction such as the absorption of a neutron and the subsequent emission of a photon (capture gamma).

**Transuranic:** Isotopes higher in atomic number than uranium; they are created by nuclear transmutations from neutron captures.

**Zircaloy:** An alloy of zirconium which is used for cladding of nuclear fuel. Zirconium has a very small probability of absorbing neutrons, so it makes a good cladding material.

## References

- Knief, Ronald A. 1992. *Nuclear Engineering: Theory and Technology of Commercial Nuclear Power*, 2nd ed. Hemisphere, New York.
- Lamarsh, John R. 1983. *Introduction to Nuclear Engineering*, 2nd ed. Addison-Wesley, Reading, MA.
- Murray, Raymond L. 1994. *Nuclear Energy*, 4th ed. Pergamon Press, Oxford, England.
- Todreas, Neil E. and Kazimi, Mujid S. 1990. *Nuclear Systems I, Thermal Hydraulic Fundamentals*. Hemisphere, New York.

## Further Information

The research and development related to nuclear power system technologies is published in *Transactions of the ANS*, published by the American Nuclear Society (ANS), La Grange Park, IL, and in *IEEE Transactions on Nuclear Science*, published by the IEEE Nuclear and Plasma Sciences Society, Institute of Electrical and Electronics Engineers, New York.

The National Council of Examiners for Engineering and Surveying (NCEES) sets examinations for professional engineering registration in the nuclear engineering area. The major work behaviors for those working in the nuclear power systems areas are outlined in *Study Guide for Professional Registration of Nuclear Engineers*, published by the ANS.

Further details on the history and present status of space nuclear technologies can be obtained from *Space Nuclear Power*, by Joseph A. Angelo, Jr., and David Buden or from the *Space Nuclear Power Systems* conference proceedings (1984–present), edited by Mohamed S. El-Genk and Mark D. Hoover, all published by Orbit Book Co., Malabar, FL.

El-Waki, M. M. "Power Plants"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

### 69.1 The Rankine Cycle

### 69.2 The Turbine

### 69.3 The Condenser

### 69.4 The Condenser Cooling System

Once-Through Cooling System • Wet Cooling Towers • Mechanical- and Natural-Draft Cooling Towers • Dry and Wet-Dry Cooling Towers

### 69.5 The Feedwater System

### 69.6 The Steam Generator

The Fuel System • Pulverized Coal Firing • Cyclone Furnaces • Fluidized-Bed Combustion • The Boiler • Superheaters and Reheaters • The Economizer • Air Preheater • Environmental Systems

### 69.7 Cycle and Plant Efficiencies and Heat Rates

## Mohammed M. El-Wakil

*University of Wisconsin*

Power plants convert a primary source of energy to electrical energy. The primary sources are (1) *fossil fuels*, such as coal, petroleum, and gas; (2) *nuclear fuels*, such as uranium, plutonium, and thorium in fission, and deuterium and tritium in fusion; and (3) *renewable energy*, such as solar, wind, geothermal, hydro, and energy from the oceans. The latter could be due to tides, waves, or the difference in temperature between surface and bottom, called *ocean-temperature energy conversion* (OTEC).

Systems that convert these primary sources to electricity are in turn generally classified as follows:

1. The **Rankine cycle**, primarily using water and steam as a working fluid, but also other fluids, such as ammonia, a hydrocarbon, a freon, and so on. It is widely used as the conversion system for fossil and nuclear fuels, solar energy, geothermal energy, and OTEC.
2. The **Brayton cycle**, using, as a working fluid, hot air–fossil fuel combustion products or a gas such as helium that is heated by nuclear fuel.
3. The combined cycle, a combination of Rankine and Brayton cycles in series.
4. Wind or water turbines, using wind, hydropower, ocean tides, and ocean waves.
5. Direct energy devices, which convert some primary sources to electricity directly (without a working fluid), such as photovoltaic cells for solar energy and fuel cells for some gaseous

fossil fuels.

In the mid-1990s U.S. power plants generated more than 550 000 megawatts. About 20% of this capacity was generated by fission nuclear fuels using the Rankine cycle. A smaller fraction was generated by hydropower, and a meager amount by other renewable sources. The largest portion used fossil fuels and the Rankine cycle. Nuclear power plants and some of the renewables are described elsewhere in this chapter. The following section describes fossil-Rankine-type power plants.

## 69.1 The Rankine Cycle

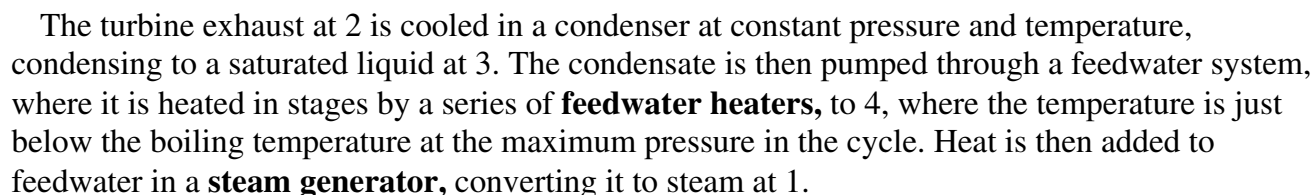
---

Rankine is a versatile cycle that can use a wide variety of heat sources. In its most common form, it uses water and steam as a working fluid. It can be built to generate large quantities of electric power, exceeding 1000 megawatts in a single power plant. It has the highest conversion efficiency (ratio of electrical energy generated to heat energy added) of all large practical conversion systems.

Figure 69.1 shows a flow diagram of a Rankine cycle. High-pressure, superheated steam is admitted to a **steam turbine** at 1, commonly at 170 bar (about 2500 psia) and  $540^{\circ}\text{C}$  (about  $1000^{\circ}\text{F}$ ), though new developments call for higher values with pressures in the supercritical range, above 221 bar (3208 psia). Steam expands through the turbine to 2, becoming a two-phase mixture of steam and water, usually 80% steam by mass, where the pressure and temperature are typically 0.07 bar and  $40^{\circ}\text{C}$  (about 1 psia and  $104^{\circ}\text{F}$ ) but vary according to the available cooling conditions in the **condenser**.

**Figure 69.1** A flow diagram of fossil-fuel Rankine-cycle power plant with one closed feedwater heater with drains pumped forward, five closed feedwater heaters with drains cascaded backward, and one open feedwater heater. HP = high-pressure turbine. IP = intermediate-pressure turbine. LP = low-pressure turbine. EG = electric generator. CO = condenser. CP = condensate pump. FP = feedwater pump. EC = economizer. DR = steam drum. BO = boiler. SU = superheaters. RE = reheaters.

## Power Plants



The energy imparted by the steam from 1 to 2 is converted to mechanical work by the turbine, which in turn drives an electric generator to produce electricity, according to the following formula:

where

© 1998 by CRC PRESS LLC

steam to the turbine

and

$$W_G = W_T \times \eta_G \quad (69.2)$$

where  $W_G$  = electrical generator power and  $\eta_G$  = electrical generator efficiency.

Modern power plant turbines are made of multiple sections, usually in *tandem* (on one axis). The first section, a high-pressure turbine, made largely of **impulse blading**, receives inlet steam and exhausts to a reheater in the steam generator. The reheated steam, at about 20% of the pressure and about the same temperature as at 1, enters an intermediate-pressure turbine made of **reaction blading**, from which it leaves in two or three parallel paths to two or three low-pressure turbines, *double-flow* and also made of reaction blading. Steam enters each in the center and exhausts at both ends, resulting in four or six paths to the condenser. This configuration divides up the large volume of the low-pressure steam—and therefore the height and speed of the turbine blades—and eliminates axial thrust on the turbine shaft. **Chapter 72** of this text describes steam turbines in greater detail.

## 69.3 The Condenser

---

The process of condensation is necessary if net power is to be generated by a power plant. (If the turbine exhaust were to be pumped back directly to the steam generator, the pumping power would be greater than the electrical power output, resulting in net negative power. Also, the second law of thermodynamics stipulates that not all heat added to a thermodynamic cycle can be converted to work; hence, some heat must be rejected.) The condenser is where heat is rejected.

To increase the cycle efficiency, the rejected heat must be minimized so that a higher percentage of the heat added is converted to work. This is done by operating the condenser at the lowest temperature, and hence the lowest pressure, possible. This is accomplished by using the lowest-temperature coolant available, usually water from a nearby large supply, such as a river, lake, or an ocean. Most power plants are situated near such bodies of water. When cooling water goes through a condenser, its temperature rises before it is readmitted to its source. To minimize this heating up and its undesirable effect on the environment, and to conserve water, cooling towers may be used. The heat rejected to the environment by the condenser  $Q_R$  is given by

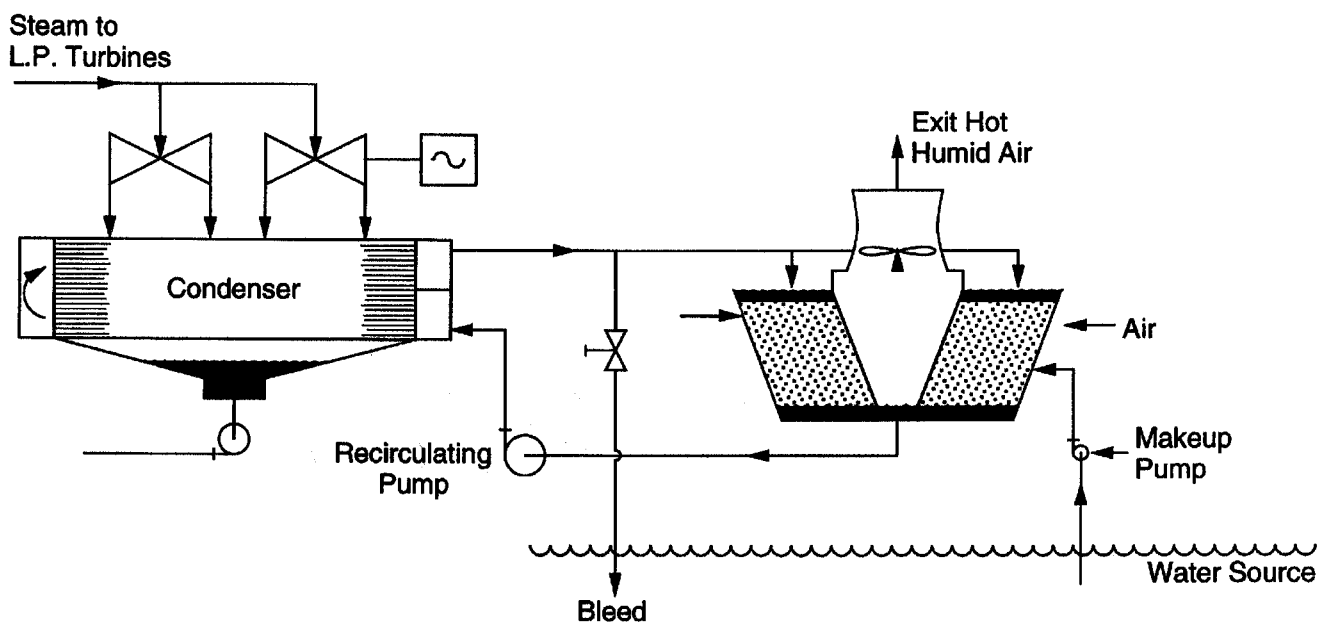
$$Q_R = m_c(h_2 - h_3) \quad (69.3)$$

where  $m_c$  = steam mass flow rate to condenser = mass flow rate of turbine inlet steam at 1 minus steam bled from the turbine for feedwater heating, as discussed later.

The most common condenser is the *surface condenser*, [Fig. 69.2](#). It is a shell-and-tube heat exchanger, composed of a steel shell with water boxes on each side connected by water tubes. Cooling water from the coolest part of the source is cleaned from debris by an intake mechanism and pumped by large circulating pumps to one of the water boxes, from which it goes through the tubes, exiting at the other box, and back to its source, at such a location to avoid reentry of the

heated water to the condenser. Such a condenser is of the *one-pass* kind. A *two-pass* condenser is one in which one box is divided into two compartments. The incoming water enters half the tubes from one compartment, reverses direction in the second box, and returns through the other half of the tubes to the second compartment of the first box. One-pass condensers require twice the quantity of cooling water as two-pass condensers, but result in lower condenser pressures and higher power plant efficiencies and are used where there are ample supplies of water. Surface condensers are large in size, often exceeding 100 000 m<sup>2</sup> (more than a million square feet) of tube surface area, and 15 to 30 m (50 to 100 ft) tube lengths in large power plants.

**Figure 69.2** A flow diagram of a power plant cooling system, with a two-pass surface condenser and a wet, mechanical-induced-draft, cross-flow cooling tower.



Another type, called the *direct-contact* or *open* condenser, is used in special applications, such as with geothermal power plants, with OTEC, and when dry cooling towers (below) are used. A direct-contact condenser is further classified as a *spray* condenser, a *barometric* condenser, or a *jet* condenser. The latter two are not widely used.

A spray direct-contact condenser is one in which demineralized cooling water is mixed with the turbine exhaust via spray nozzles. The mixture becomes a saturated liquid condensate. A fraction of it equal to the turbine flow goes to the cycle; the balance is cooled in a dry cooling tower and then recirculated to the condenser spray nozzles. The ratio of cooling water to turbine flow is large, about 20 to 25. In geothermal plants the fraction equal to the turbine flow may be returned to the ground. In OTEC it is returned to the ocean.

## 69.4 The Condenser Cooling System

A condenser cooling system may be open (or *once-through*) or partially closed, using **cooling**

**towers.** The latter are classified into *wet natural-draft cooling towers*, *wet mechanical-draft cooling towers*, and *dry cooling towers*.

## Once-Through Cooling System

Here, cooling water is taken from the source (usually at a depth where it is sufficiently cool), passed through the condenser, then returned to the source at a point that ensures against short circuiting of the warmer water back to the condenser, such as downstream of the intake.

Once-through systems are the most efficient means of cooling a condenser, but require large quantities of water and discharge warm water back to the source. Environmental regulations often prohibit the use of once-through systems, in which case cooling towers are used.

## Wet Cooling Towers

In a wet cooling tower, the warm condenser water is essentially cooled by direct contact with atmospheric air. It is sprayed over a lattice of slats or bars, called a *fill* or *packing*, which increases its surface-to-volume ratio, as in Fig. 69.2. Atmospheric air passes by the water in a *cross-flow* or *counterflow* manner. The water is cooled by exchanging heat with the cooler air and, more importantly, by partial evaporation into the heated and, hence, lower relative humidity air.

Because of evaporative losses, wet towers do not eliminate the need for water, but they appreciably reduce it, since these losses are a fraction of the total water flow. An additional loss is due to *drift*, in which unevaporated water drops escape with the air. *Drift eliminators* are added to reduce this loss. Evaporative losses depend upon the climatic conditions and could be as high as 1.5% of total water flow. Drift could be as high as 2.5% of the evaporative losses. An additional loss is *blowdown* or *bleed*. Warm water in the tower contains suspended solids and is fully aerated. Chemical additives are used to inhibit microbiological growth and scales. Thus, a certain percentage of the circulating cooling water is bled to maintain low concentrations of these contaminants. The bleed, nearly as high as the evaporative loss if high purity is to be maintained, is often returned to the source after treatment to minimize pollution. All losses must be compensated for by *makeup*; power plants using wet cooling towers are also cited near bodies of water. Other problems of wet towers are icing and fogging due to exiting saturated air in cold weather.

## Mechanical- and Natural-Draft Cooling Towers

Atmospheric air flows through cooling towers either mechanically or naturally. In the former, it is moved by one or more fans. Because of distribution problems, leaks, and possible recirculation of the hot humid exit air, most mechanical towers move the air by *induced-draft* fans. These are placed at the top and suck the hot air through the tower. Such towers are usually multicell with several fans placed in stacks atop a bank of towers, the number depending upon the size of the power plant. The fans are usually multibladed (made of aluminum, steel, or fiberglass), are driven at low speeds by electric motors through reduction gearing, and could be as large as 10 m (33 ft) in diameter. Mechanical-draft cooling towers consume power and are relatively noisy.



In the natural-draft cooling tower, air flows by a natural driving force,  $F_D$ , caused by the density differences between the cool air outside and the warm air inside the tower, given by

$$F_D = (\rho_o - \rho_i)Hg \quad (69.4)$$

where

$\rho_o$  and  $\rho_i$  = average densities of air outside and inside the tower  
 $H$  = height of the tower  
 $g$  = the gravitational acceleration

Because the difference between the densities is small,  $H$  is large—about 130 m (430 ft). The towers are imposing structures that are visible from afar, are costly to build, but consume no power. The water distribution system and fill are placed at the bottom, and most of the tower height is open space of circular cross section. The vertical profile is hyperbolic, which offers good resistance to wind pressures. Natural-draft cooling towers are usually made of reinforced concrete and sit on stilts and are mostly of the counterflow type.

A compromise between mechanical and natural draft towers is called the *hybrid* or *fan-assisted hyperbolic* cooling tower. A number of forced draft fans surround the bottom to augment the natural driving force of a shorter hyperbolic tower. The hybrid consumes less power than a mechanical and is smaller and less costly than a natural tower.

## Dry and Wet-Dry Cooling Towers

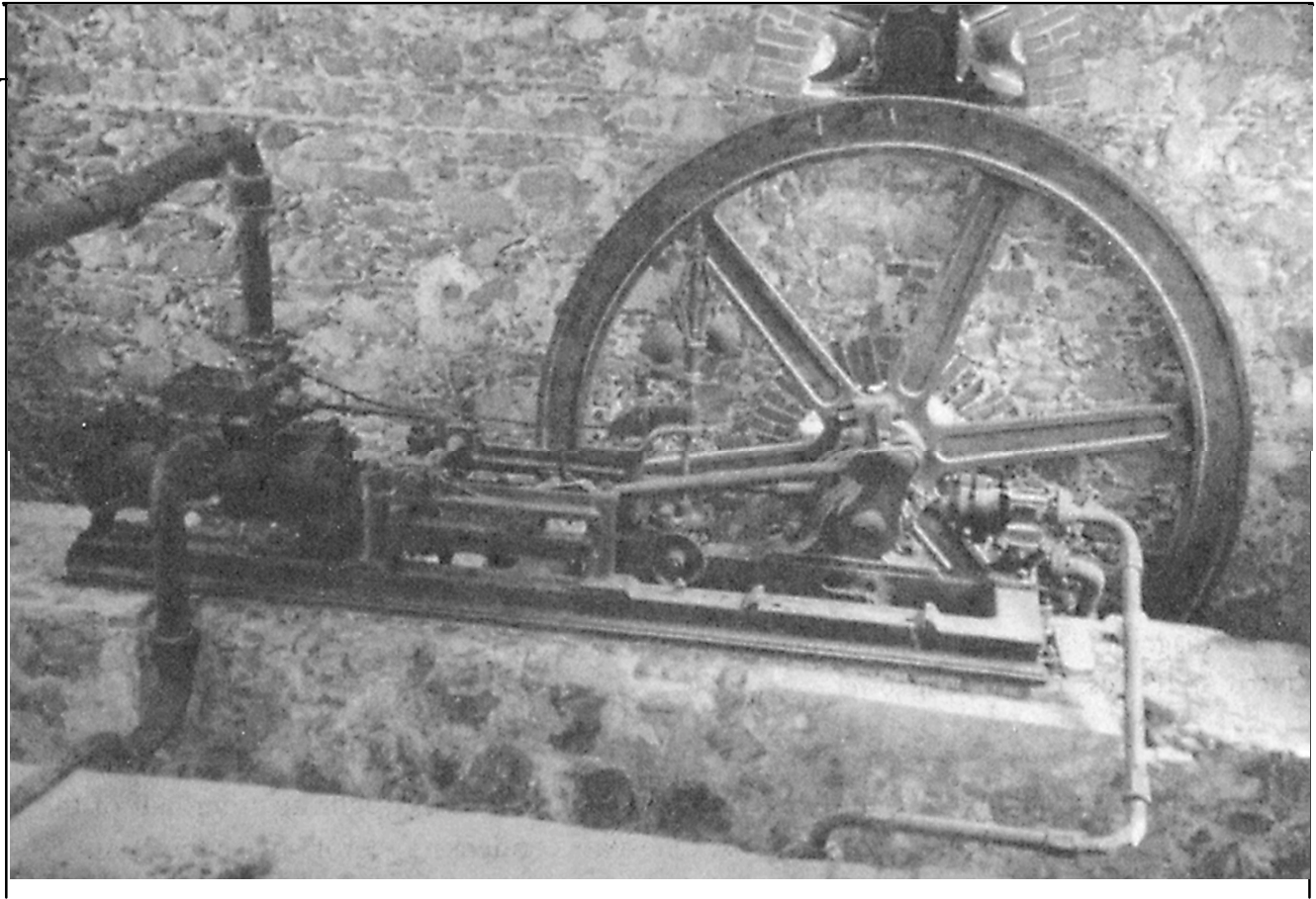
*Dry cooling towers* are used when a power plant is sited far from adequate sources of water, near coal mines or other abundant fuel to reduce transportation costs, near large power consumers to reduce electrical transmission costs, or in arid areas. They are essentially closed-type heat exchangers in which warm condenser water passes through a large number of finned tubes cooled by atmospheric air and no water is lost due to evaporation, drift, and so on. They are usually easier to maintain than wet towers and do not suffer from fogging or icing.

Dry cooling towers, however, are not as effective. Lacking evaporative cooling, they have lower heat transfer capabilities, resulting in large heat exchanger surfaces and land areas. This also results in higher condenser water temperatures and hence higher back pressures on the turbine than with wet towers, resulting in lower power plant efficiencies. The problem is aggravated further during periods of high atmospheric air temperatures.

Dry cooling towers may be mechanical or natural draft. They also may be designed to operate in *direct* or *indirect* modes. In the direct mode the turbine exhaust steam passes through large finned tubes that are cooled by the atmospheric air. Indirect dry towers, more common, use a conventional surface condenser with an intermediate coolant, such as water, or a two-phase fluid, such as ammonia. The latter, under development, improves heat transfer and results in lower penalties on the cycle efficiency.

*Wet-dry cooling towers* are combinations of the above. Warm condenser water enters a dry section of the tower—reducing its temperature partially—then goes on to a wet section. Parallel air flows to each section combine to a common exit. This reduces fogging and evaporative losses,

but at the expense of more complexity and cost.



This piston steam engine, exhibited at the present time in the partially restored Reef Bay Estate House and Sugar Factory at the National Park of St. John, U.S. Virgin Islands, was first installed sometime after 1861 by the estate owner, Mr. William H. Marsh, to replace horse power by steam power for grinding the sugar cane. It was manufactured in Glasgow, Scotland, in 1861 and is estimated to have been in operation for 50 to 60 years. Based on external combustion, the primary fuel was wood and bagasse. Its installation exemplifies a trend typical to those times, when the newly developed steam engines were found to be more economical and practical to use than human, animal, or even wind power.

The invention of the steam engine is due to Thomas Savery (1698) and Thomas Newcomen (1712) from England. The major need that prompted their invention was coal mining, which became prominent due to depletion of forests at that time, and the steam engines were used to drive pumps for draining water from coal mines. Steam engines were used afterward for many applications, including locomotion. The latter was primarily for railroad transportation, and even for cars. Because they operate on external combustion, where a variety of fuels can be used and emissions can be controlled to much lower levels than in conventional internal engines, and because modern technology allows them to operate at high efficiency, research and development of such engines for automotive use was revived by a number of major car manufacturers during

the 1970s and 1980s energy crisis era. (Photo courtesy of Noam Lior, University of Pennsylvania.)

---

## 69.5 The Feedwater System

---

The condensate at 3 (in [Fig. 69.1](#)) is returned to the cycle to be converted to steam for reentry to the turbine at 1. Called the *feedwater*, it is pumped by condensate and feedwater pumps—to overcome flow pressure losses in the feedwater system and the steam generator—and enters the turbine at the desired pressure. The feedwater is heated successively to a temperature close to the saturated temperature at the steam generator pressure. This process, called *regeneration* or *feedwater heating*, results in marked improvement in cycle efficiency and is used in all modern Rankine cycle power plants, both fossil and nuclear.

Regeneration is done in stages in feedwater heaters, which are of two types: (1) *closed* or *surface* type and (2) *open* or *direct-contact* type. The former are further classified into *drains cascaded backwards* and *drains pumped forward*. All types use steam bled from the turbine at pressures and temperatures chosen to match the temperatures of the feedwater in each feedwater heater. The amount of steam bled from the turbine is a small fraction of the total turbine flow because it essentially exchanges its latent heat of vaporization with sensible heat of the single-phase feedwater.

Closed feedwater heaters are shell-and-tube heat exchangers where the feedwater flows inside tubes and the bled steam condenses over them. Thus, they are much like condensers but are smaller and operate at higher pressures and temperatures. The steam that condenses is returned to the cycle. It is either cascaded backwards—that is, throttled to the next lower-pressure feedwater heater—or pumped forward into the feedwater line. The cascade type is most common (see [Fig. 69.1](#)).

Open feedwater heaters, on the other hand, mix the bled steam with the feedwater, resulting in saturated liquid. The mix is then pumped by a feedwater pump to the next higher-pressure feedwater heater. Most power plants use one open-type feedwater heater, which doubles as a means to rid the system of air and other noncondensable gases; this type is often referred to as a *deaerating* or *DA* heater. It is usually placed near the middle of the feedwater system, where the temperature is most conducive to deaeration.

The mass flow rate of the bled steam to the feedwater heaters is obtained from energy balances on each heater [[El-Wakil, 1984](#)]. This determines the mass flow rate in each turbine section [which is necessary to evaluate the turbine work, Eq. (69.1)] and the heat rejected by the condenser [Eq. (69.3)].

## 69.6 The Steam Generator

---

A modern fossil-fuel power plant steam generator is a complex system. Combustion gases pass successively through the boiler, superheaters, reheaters, the economizer, the air preheater, and finally leave through a stack.

## The Fuel System

Fuel is burned in a furnace with excess air (more than stoichiometric or chemically correct). This combustion air is forced through the system from the atmosphere by a **forced-draft fan**, resulting in combustion gases at about 1650°C (3000°F). At steam generator exit the air (now called *flue gases*) is drawn out by an *induced-draft fan* at about 135 to 175°C (275 to 350°F) into the stack. This seemingly high temperature represents an energy loss to the system but is necessary to prevent condensation of water vapor in the gases, which would combine with other combustion products to form acids and to facilitate flue gas dispersion into the atmosphere.

## Pulverized Coal Firing

Furnaces have undergone much evolution. With coal, the old mechanical stokers have given way to **pulverized coal** firing in most modern systems. To pulverize, *run-of-the-mill* (as shipped from the mine) coal, averaging about 20 cm (8 in.) in size, is reduced to below 2 cm (0.75 in.) by *crushers*, which are of several types, including *rings*, *Bradford breakers*, and *hammer mills*. The crushed coal is then dried by air at 345°C (650°F) or more, obtained from the air preheater. The coal is then ground by pulverizers, which are usually classified according to speed. A common one is the medium-speed (75 to 225 rpm) *ball-and-race pulverizer*, which grinds the coal between two surfaces. One surface consists of steel balls that roll on top of the other surface, similar to a large ball bearing. Hot air then carries the powdery coal in suspension to a *classifier*, which returns any escaping large particles back to the grinders.

Pulverized coal is classified as 80% passing a #200 mesh screen (0.074 mm openings) and 99.99% through a #50 mesh screen (0.297 mm). It is fed to the furnace burners via a set of controls that also regulate primary (combustion) air to suit load demands. Large steam generators have more than one pulverizer system, each feeding a number of burners for a wide range of load control. Burners may be designed to burn pulverized coal only or to be multifuel, capable of burning pulverized coal, oil, or gas.

## Cyclone Furnaces

A **cyclone furnace** burns crushed coal (about 95 percent passing a #4 mesh screen, about 5 mm). It is widely used to burn poorer grades of coal that contain high percentages of ash and volatile matter. Primary air, about 20% of the total combustion air, and the rest, secondary and tertiary air, enter the burner successively and tangentially, imparting a centrifugal motion to the coal. This good mixing results in high rates of heat release and high combustion temperatures that melt most of the ash into a molten slag. This drains to a tank at the bottom of the cyclone where it gets solidified, broken, and removed. Ash removal materially reduces erosion and fouling of steam generator surfaces and reduces the size of particulate matter—removal equipment such as electrostatic precipitators and bag houses. The disadvantages of cyclone firing are high power requirements and, because of the high temperatures, the production of more pollutants, such as oxides of nitrogen, NO<sub>x</sub>.

## Fluidized-Bed Combustion

Another type of furnace uses **fluidized-bed combustion**. Crushed coal particles, 6 to 20 mm (0.25 to 0.75 in.) in size, are injected into a bed above a bottom grid. Air from a plenum below flows upwards at high velocity so that the drag forces on the particles are at least equal to their weight, and the particles become free or fluidized with a swirling motion that improves combustion efficiency. Combustion occurs at lower temperatures than in a cyclone, reducing  $\text{NO}_x$  formation. About 90% of the sulfur dioxide that results from sulfur in the coal is largely removed by the addition of limestone (mostly calcium carbonate,  $\text{CaCO}_3$ , plus some magnesium carbonate,  $\text{MgCO}_3$ ) that reacts with  $\text{SO}_2$  and some  $\text{O}_2$  from the air to form calcium sulfate,  $\text{CaSO}_4$ , and  $\text{CO}_2$ . The former is a disposable dry waste. Technical problems, such as the handling of the calcium sulfate, are under active study.

## The Boiler

The boiler is that part of the steam generator that converts saturated water or low-quality steam from the economizer to saturated steam. Early boilers included fire-tube, scotch marine, straight-tube, and Stirling boilers. The most recent are water-tube-water-wall boilers. Water from the economizer enters a steam drum, then flows down insulated down-comers, situated outside the furnace to a header. The latter feeds vertical closely spaced water tubes that line the furnace walls. The water in the tubes receives heat from the combustion gases and boils to a two-phase mixture. The density difference between the down-comer and the tubes causes a driving force that circulates the mixture up the tubes and into the drum. The tubes also cool the furnace walls. There are several water-wall designs. A now-common one is the *membrane* design, in which 2.75 to 3 in. tubes on 3.75 to 4 in. centers are connected by welded membranes that act as fins to increase the heat-transfer surface as well as form a pressure-tight wall protecting the furnace walls. The steam drum now contains a two-phase bubbling mixture, from which dry steam is separated by gravity and mechanically with baffles, screens, and centrifugal separators.

## Superheaters and Reheaters

Dry-saturated steam from the boiler enters a primary and then a secondary superheater in series, which convert it to superheated steam. The superheaters are made of 2 to 3 in. diameter U-tube bundles made of special high-strength alloy steels of good strength and corrosion resistance, suitable for high-temperature operation. The bundles are usually hung from the top, and called *pendant tubes*, or from the side, and called *horizontal tubes*. Another type, supported from the bottom and called the *inverted tube*, is not widely used.

Superheated steam enters the high-pressure turbine and exhausts from it to return to the steam generator, where it is reheated to about the same turbine inlet temperature in a set of reheaters, downstream of and similar in design to the superheaters. The reheated steam enters the intermediate-pressure and then the low-pressure turbines, as explained above.

Superheaters and reheaters may be of the *radiant* or *convection* types. The former, in view of the luminous combustion flames in the furnace, receive heat primarily by radiation. This heat

transfer mode causes the exit steam temperature to decrease with increasing load (steam flow). The latter receive heat by convection, the main form of heat transfer in superheaters and reheaters, which causes the exit steam temperature to increase with load. To obtain fairly constant steam temperatures, *attemperators* are placed between the primary and secondary sections of superheaters and reheaters. In its most common form, an attemperator maintains the desired temperatures by spraying regulated amounts of lower-temperature water from the economizer or boiler directly into the steam.

## The Economizer

The flue gases leave the reheaters at 370 to 540°C (700 to 1000°F). Rather than reject their energy to the atmosphere, with a consequent loss of plant efficiency, flue gases now heat the feedwater leaving the last (highest pressure) feedwater heater to the inlet temperature of the steam generator. This is done in the *economizer*. At high loads the economizer exit may be a low-quality water-steam mixture. Economizers are usually made of tubes, 1.75 to 2.75 in. in diameter, arranged in vertical sections between headers and placed on 1.75 to 2 in. spacings. They may be plain-surfaced, finned, or studded to increase heat transfer. Smaller spacings and studs are usually used with clean ash-free burning fuels, such as natural gas.

## Air Preheater

Flue gases leave the economizer at 315 to 425°C (600 to 800°F). They are now used in an *air preheater* to heat the atmospheric air, leaving the forced-draft fan to about 260 to 345°C (500 to 650°F) before admitting it to the furnace, thus reducing total fuel requirements and increasing plant efficiency. Air preheaters may be recuperative or regenerative. *Recuperative* preheaters are commonly counterflow shell-and-tube heat exchangers in which the hot flue gases flow inside and the air outside vertical tubes, 1.5 to 4 in. in diameter. A hopper is placed at bottom to collect soot from inside the tubes. *Regenerative* preheaters use an intermediate medium. The most common, called *ljungstrom*, is rotary and is driven by an electric motor at 1 to 3 rpm through reduction gearing. The rotor has 12 to 24 sectors that are filled with a heat-absorbing material such as corrugated steel sheeting. About half the sectors are exposed to and are heated by the hot flue gases moving out of the system at any one instant; as the sectors rotate, they become exposed to and heat the air that is moving in the opposite direction (into the system).

## Environmental Systems

Besides cyclone and fluidized-bed combustion, there are other systems that reduce the impact of power generation on the environment. Flue gas desulfurization systems, also called **scrubbers**, use aqueous slurries of lime-limestone to absorb SO<sub>2</sub>. **Electrostatic precipitators** remove particulate matter from the flue gases. Here, wire or discharge electrodes carry a 40 to 50 kV current and are centrally located between grounded plates or collection electrodes. The resulting current charges the soot particles, which migrate to the plates, where they are periodically removed. Fabric filters or **baghouses** also remove particulate matter. They are made of a large

number of vertical hollow cylindrical elements—5 to 15 in. in diameter and up to 40 ft high, made of various porous fabrics (wool, nylon, glass fibers, etc.)—through which the flue gases flow and get cleaned in the manner of a household vacuum cleaner. The elements are also periodically cleaned.

## 69.7 Cycle and Plant Efficiencies and Heat Rates

---

The heat added to the power plant,  $Q_A$ , and the cycle,  $Q_C$ , are given by the following:

$$Q_A = m_f \times \text{HHV} \quad (69.5a)$$

$$Q_C = Q_A \times \eta_{sg} = m_1(h_1 - h_4) \quad (69.5b)$$

where  $m_f$  = the mass flow rate of fuel to the furnace, in kg/h or lb/h, and HHV = higher heating value of fuel, in kJ/kg or Btu/lb.

The plant efficiency,  $\eta_P$ , and cycle efficiency,  $\eta_C$ , are given by the following:

$$\eta_P = W_G / Q_A \quad (69.6a)$$

$$\eta_C = W_T / Q_C \quad (69.6b)$$

The value  $\eta_P$ , given above, is often referred to as the plant *gross efficiency*. Since some of the generator power,  $W_G$ , is used within the plant to power various equipment, such as fans, pumps, pulverizers, lighting, and so on, a *net efficiency* is often used, in which  $W_G$  is reduced by this auxiliary power. Another parameter that gives a measure of the economy of operation of the power plant is called the **heat rate, HR**. It is given by the ratio of the heat added in Btu/h to the plant power in kW, which may be gross or net. For example,

$$\text{Net plant HR, Btu/kWh} = (Q_A, \text{ Btu/h}) / (W_G - \text{auxiliary power, kW}) \quad (69.7)$$

The lower the value of HR is, the better. A benchmark net HR is 10 000, equivalent to a net plant efficiency of about 34%. It could be as high as 14 000 for older plants and as low as 8500 for modern plants.

### Defining Terms

**Baghouse:** Removes particulate matter from the flue gases by porous fabric filters.

**Brayton cycle:** A cycle in which a gas (most commonly air) is compressed, heated, and expanded in a gas turbine to produce mechanical work.

**Condenser:** A heat exchanger in which the exhaust vapor (steam) of the turbine in a Rankine cycle is condensed to liquid, usually by cooling water from an outside source, for return back to the steam generator.

**Cooling tower:** A heat exchanger in which the condenser cooling water is in turn cooled by

atmospheric air and returned back to the condenser.

**Cyclone furnace:** A furnace in which crushed coal is well mixed with turbulent air, resulting in good heat release and high combustion temperatures that melt the coal ash content into removable molten slag, thus reducing furnace size and the fly ash content of the flue gases and eliminating the cost of coal pulverization.

**Electrostatic precipitator:** A system that removes particulate matter from the flue gases by using one electrode at high voltage to electrically charge the particles, which migrate to the other grounded electrode, where they are periodically removed.

**Feedwater heaters:** Heat exchangers that successively heat the feedwater before entering the steam generator using steam that is bled from the turbine.

**Fluidized-bed furnace:** A furnace in which crushed coal is floated by upward air—resulting in a swirl motion that improves combustion efficiency, which in turn gives lower combustion temperatures and reduced  $\text{NO}_x$  in the flue gases—and in which limestone is added to convert much of the sulfur in the coal to a disposable dry waste.

**Forced-draft fan:** The fan that forces atmospheric air into the steam generator to be heated first by an air preheater and then by combustion in the furnace.

**Heat rate:** The rate of heat added to a power plant in Btu/h to produce one kW of power.

**Impulse blades:** Blades in the high-pressure end of a steam turbine and usually symmetrical in shape that ideally convert kinetic energy of the steam leaving a nozzle into mechanical work.

**Once-through cooling:** The exhaust vapor from the turbine of a Rankine cycle is condensed by cool water obtained from an available supply such as a river, lake, or the ocean, and then returned to that same supply.

**Pulverized coal:** A powdery coal that is prepared from crushed and dried coal and then ground, often between steel balls and a race.

**Rankine cycle:** A closed cycle that converts the energy of a high-pressure and high-temperature vapor produced in a steam generator (most commonly steam) into mechanical work via a turbine, condenser, and feedwater system.

**Reaction blades:** Blades downstream of impulse blades in a steam turbine and having an airfoil shape that convert some of both kinetic and enthalpy energies of incoming steam to mechanical work.

**Scrubbers:** A desulfurization system that uses aqueous slurries of lime-limestone to absorb  $\text{SO}_2$  in the flue gases.

**Steam generator:** A large complex system that transfers the heat of combustion of the fuel to the feedwater, converting it to steam that drives the turbine. The steam is usually superheated at subcritical or supercritical pressures (critical pressure = 3208 psia or 221 bar). A modern steam generator is composed of economizer, boiler, superheater, reheater, and air preheater.

**Steam turbine:** A machine that converts steam energy into the rotary mechanical energy that drives the electric generator. It is usually composed of multiple sections that have impulse blades at the high-pressure end, followed by reaction blades.

## References

El-Wakil, M. M. 1984. *Powerplant Technology*. McGraw-Hill, New York.



Singer, J. G. (Ed.) 1991. *Combustion, Fossil Power*. Combustion Engineering, Windsor, CT.  
Stultz, S. C., and Kitto, J. B. (Eds.) 1992. *Steam, Its Generation and Use*. Babcock and Wilcox, Barberton, OH.

## **Further Information**

*Proceedings of the American Power Conference*. American Power Conference, Illinois Institute of Technology, Chicago, IL 60616.

ASME publications (ASME, 345 E. 47th Street, New York, NY 10017):

*ASME Boiler and Pressure Vessel Code*

*Mechanical Engineering*

*Journal of Energy Resources Technology*

*Journal of Gas Turbines and Power*

*Journal of Turbomachinery*

*Combustion and Flame*. The Journal of the Combustion Institute, published monthly by Elsevier Science, 655 Avenue of the Americas, New York, NY 10010.

Department of Energy, Office of Public Information, Washington, DC 20585.

*EPRI Journal*. The Electric Power Research Institute, P.O. Box 10412, Palo Alto, CA 94303.

*Energy, International Journal*. Elsevier Science Ltd., Bampfylde Street, Exeter EX1 2AH, England.

*Construction Standards for Surface Type Condensers for Ejector Service*. The Heat Exchange Institute, Cleveland, OH.

*Standards for Closed Feedwater Heaters*. The Heat Exchange Institute, Cleveland, OH.

*Power*. McGraw-Hill, P.O. Box 521, Hightstown, NJ 08520.

*Power Engineering*. 1421 Sheridan Road, Tulsa, OK 74112.

Swift, A., Moroz, E. "Wind Turbines"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## 70.1 Fundamentals

Resource Base • Types of Wind Turbines • Basic Equations

## 70.2 Power Regulation and Control

## 70.3 Energy Capture Estimation

## 70.4 Stand-Alone Applications

## 70.5 Cost of Energy Calculations

## 70.6 Environmental and Social Cost Issues

## 70.7 Summary

### **Andrew Swift**

*University of Texas at El Paso*

### **Emil Moroz**

*University of Texas at El Paso*

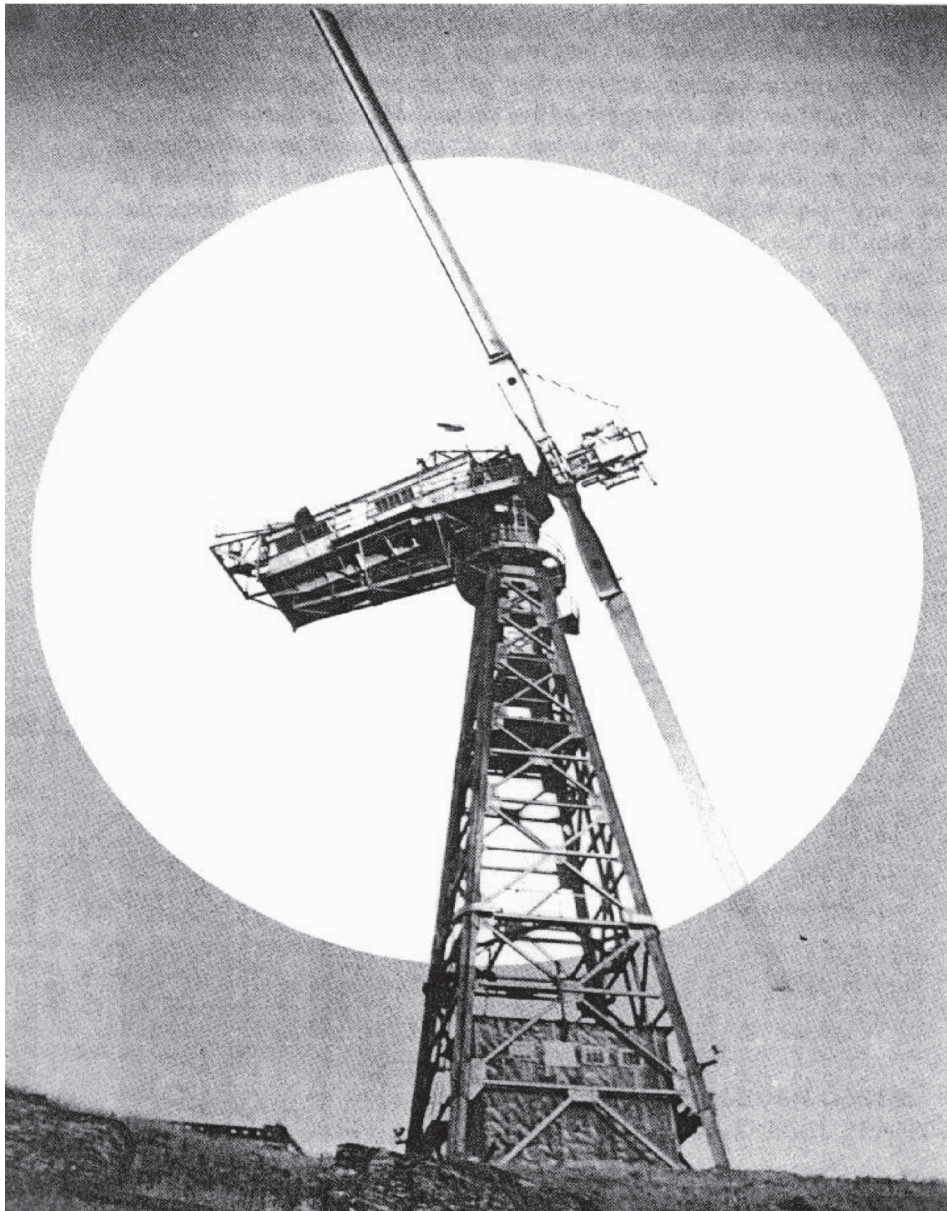
A wind turbine is a mechanical device that converts the kinetic energy from the wind into useful shaft power. The first documented evidence of wind turbine utilization comes from 10th century Persia, where wind power was used for grinding corn and pumping water. The Persian design used a vertical axis, carousel-type turbine, which was quickly adapted to use in China, India, and portions of the Muslim world. Since most of the early wind turbine designs were used to grind grain, the term *windmill*, instead of wind turbine, is often used. Following the Persian development, the European *post mill* was introduced. This configuration—named after the main post that was used to secure the mill to the ground and around which the mill turned to maintain alignment with the wind—dominated wind turbine development for several centuries. From the 12th to the 19th century the post mill was second only to the hydraulic turbine, or water wheel, in providing useful energy, and it profoundly affected European development and society during this time [Richter, 1994].

In the mid-19th century, water-pumping windmills were the key to expansion of railroads and settlement in the arid western U.S. The turbine designed for these applications was relatively small, lightweight, and self-regulating and had replaceable parts. This multibladed, slow-turning farm and ranch windmill (see Fig. 70.5) has become an icon, since millions were produced and installed throughout the U.S. and abroad.

In 1887 the talented inventor Charles Brush built the first electric wind generator at his home in Cleveland, Ohio, and used it to power the lights and motors of his laboratory. Several companies

developed and sold wind electric generators over the next 50 years, mostly to remote homesteads, until the Rural Electrification Administration connected rural customers to a central power grid and put the wind generator manufacturers out of business. While the battle for small, independent wind electric power systems was effectively lost in the U.S., the first large-scale wind turbines built for bulk electric power generation and direct connection to the utility grid were being tested in Europe. Large prototype turbines with generator ratings up to several hundred kilowatts were built in various countries, including Russia, Denmark, Britain, France, and Germany. Following this trend, the first large-scale wind turbine generator built in the U.S. was the Smith-Putnam turbine, constructed in 1940 and shown in Fig. 70.1. Rated at 1250 kilowatts, it was connected to the electric grid at Grandpa's Knob, Vermont. It operated for 16 months until bearing problems followed by a catastrophic blade failure led to the cancellation of the project in 1945 [Koeppel, 1982].

**Figure 70.1** The Smith-Putnam wind turbine, 1940. (Source: Eldridge, F. R. 1975. *Wind Machines*, p. 8. National Science Foundation, Grant Number AER-75-12937. Energy Research and Development Administration, Washington, DC.)



The worldwide oil crisis of 1973 led to a strong revival of interest in the contribution that renewable sources of energy, such as wind, could play in reducing a country's dependence on imported fuel. Furthermore, the costs of nuclear power were found to be much greater than initially anticipated, and the problems with nuclear waste storage and pollution from fossil fueled power plants were recognized. These factors contributed to the emergence in the 1980s of a worldwide development of a wind power industry—manufacturing wind turbines, installing them on **wind farms** (see [Fig. 70.2](#)), and delivering electric power to the utility grid. Although there were problems with costs and reliability with the first commercial, large-scale turbines, experience during the last 15 years has decreased costs and increased reliability dramatically. Costs have decreased from near 30 cents per kWh in 1980 to near 4 cents per kWh for a modern wind farm at a good wind site. Furthermore, availabilities over 95% are common.

**Figure 70.2** Carland Cross wind farm in Cornwall, England. (Photo courtesy of the British Wind Energy Association.)



Brown *et al.* [1994] report that presently there are approximately 20 000 turbines operating worldwide with an installed capacity of 3000 MW and an annual growth rate of 13%, making wind technology one of the fastest-growing energy sources (see [Table 70.1](#)). Most of the currently operating turbines are in the U.S. and Europe. Wind farms in California account for

approximately 65% of the present installed capacity, but projections for growth in Europe show that installed capacity there will exceed that of the U.S. shortly after the turn of the century. Also, wind power is being considered as a serious energy option by several other countries, including China, India, Mexico, New Zealand, and Ukraine.

**Table 70.1** World Wind Energy-Generating Capacity

Year	Capacity(megawatts)
1980	4
1981	16
1982	33
1983	97
1984	274
1985	694
1986	1025
1987	1401
1988	1568
1989	1579
1990	1789
1991	2208
1992	2633
1993 (est.)	2976

*Source:* Brown, L. R., 1994. *Vital Signs, 1994*, p. 51. Worldwatch Institute, Norton, New York.

## 70.1 Fundamentals

### Resource Base

Atmospheric winds are a result of the uneven heating of the earth's surface by the sun and are therefore considered a renewable resource. The power available in the wind is in the form of kinetic energy and can be calculated per unit area by

$$P = 0.5\rho AV^3 \quad (70.1)$$

where  $P$  is the power in watts,  $\rho$  is the air density ( $\text{kg}/\text{m}^3$ ),  $A$  is the swept area intercepting the wind ( $\text{m}^2$ ), and  $V$  is the wind speed ( $\text{m}/\text{s}$ ).

The total wind power available in the earth's atmosphere is huge. However, to calculate practical wind power potential, estimates must be limited to areas where wind turbines can be feasibly located to extract and deliver the power. Although there have been proposals to locate wind turbines on blimps and from helicopters to take advantage of more energetic winds at high altitudes, most of these concepts have been dismissed in favor of land-based, or offshore, tower and foundation designs. When these constraints are combined with siting limitations, such as



urban and environmentally sensitive areas, the practical wind electric generation potential is greatly reduced—but it is still enormous. In Europe, for example, wind power could theoretically satisfy all the continent's present electricity needs. In the U.S. the great plains area alone could meet the nation's electricity needs several times over [Brown *et al.*, 1993].

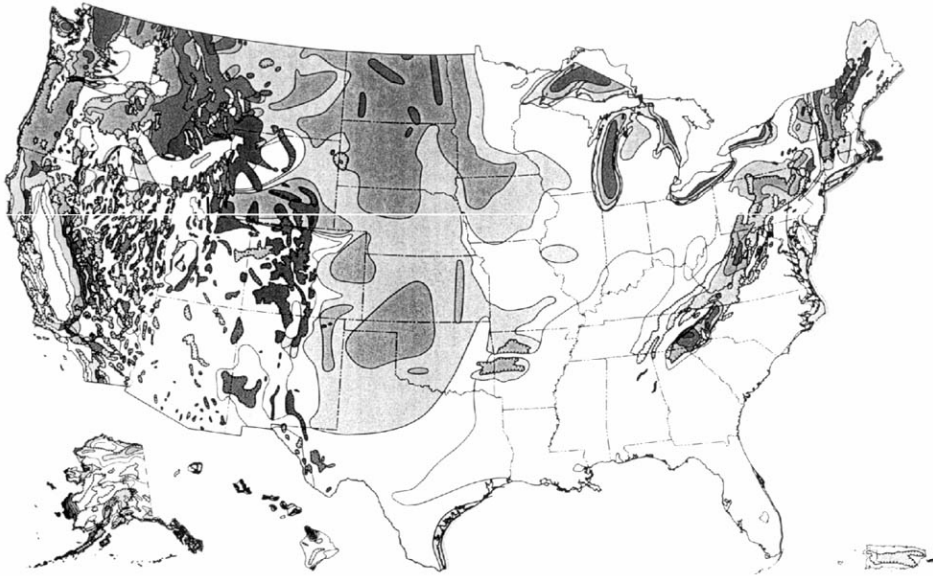
Wind power potential is based on long-term average wind speed at a site and is usually determined by power class rating, as shown in Table 70.2. Sites with power class 4 and above are considered economic for development with available technology and present conventional energy costs. Class 3 sites are considered marginal now, but will be viable as the technology improves and costs of generation are lowered. Figure 70.3 shows a wind atlas of the U.S. produced by Battelle Laboratories [Elliott *et al.*, 1986], and Fig. 70.4 shows the wind electric generating potential of the U.S. for class 4 sites and above as a fraction of the present U.S. electricity supply [Utility Wind Interest Group, 1992].

**Table 70.2** Classes of Wind Power Density

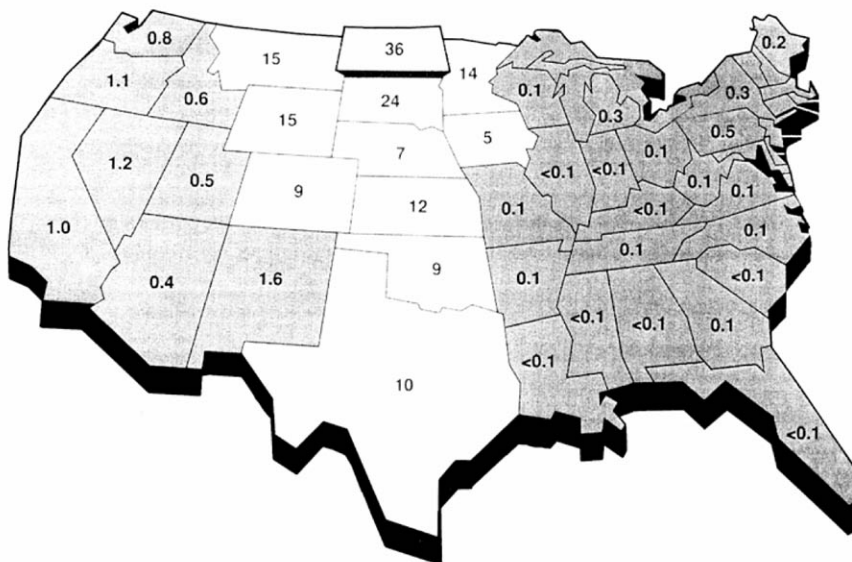
Wind Power Class	10 m (33 ft)			50 m (164 ft)		
	Wind Power W/m <sup>2</sup>	Speed		Wind Power W/m <sup>2</sup>	Speed	
		m/s	mph		m/s	mph
1	0	0	0	0	0	0
2	100	4.4	9.8	200	5.6	12.5
3	150	5.1	11.5	300	6.4	14.3
3	200	5.6	12.5	400	7.0	15.7
4	250	6.0	13.4	500	7.5	16.8
5	300	6.4	14.3	600	8.0	17.9
6	400	7.0	15.7	800	8.8	19.7
7	1000	9.4	21.1	2000	11.9	26.6
Ridge Crest estimates (local relief > 1000 ft)						

Source: Elliott, D. L. 1986. *Wind Energy Resource Atlas of the United States*, p. 12. DOE/CH 10093-4. Solar Technical Information Program, National Renewable Energy Laboratory, Golden, CO.

**Figure 70.3** U.S. wind resource map. (Source: Elliott, D. L. 1986. *Wind Energy Resource Atlas of the United States*, p. 12. DOE/CH 10093-4. Solar Technical Information Program, National Renewable Energy Laboratory, Golden, CO.)



**Figure 70.4** U.S. wind electric generating potential by state. (Source: UWIG. 1992. *America Takes Stock of a Vast Energy Resource*, p. 31. Electric Power Research Institute, Palo Alto, CA. With permission.)



## Types of Wind Turbines

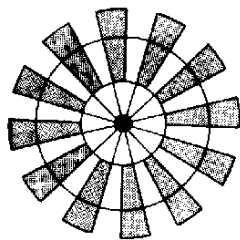
The simplest form of wind turbine can be constructed from an oil drum that has been split along its longest axis and mounted upon a suitable shaft. Named after its designer, the Savonius type relies on differential drag to provide its driving force. This type has found some limited application in water-pumping activities where high starting torque is of great benefit. See [Fig. 70.5](#).



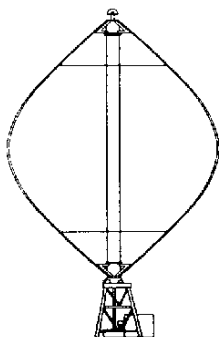
**Figure 70.5** Typical wind turbine configurations.



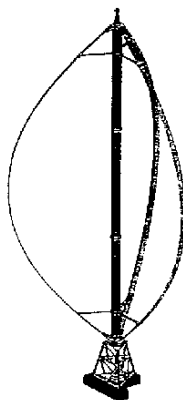
**Savonius**



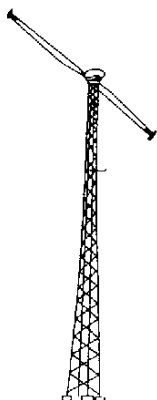
**Farm and Ranch**



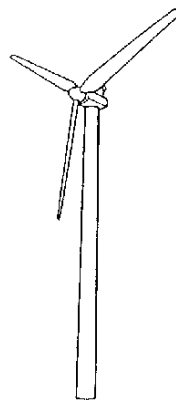
**Darrieus, 2 bladed**



**Darrieus, 3 bladed**



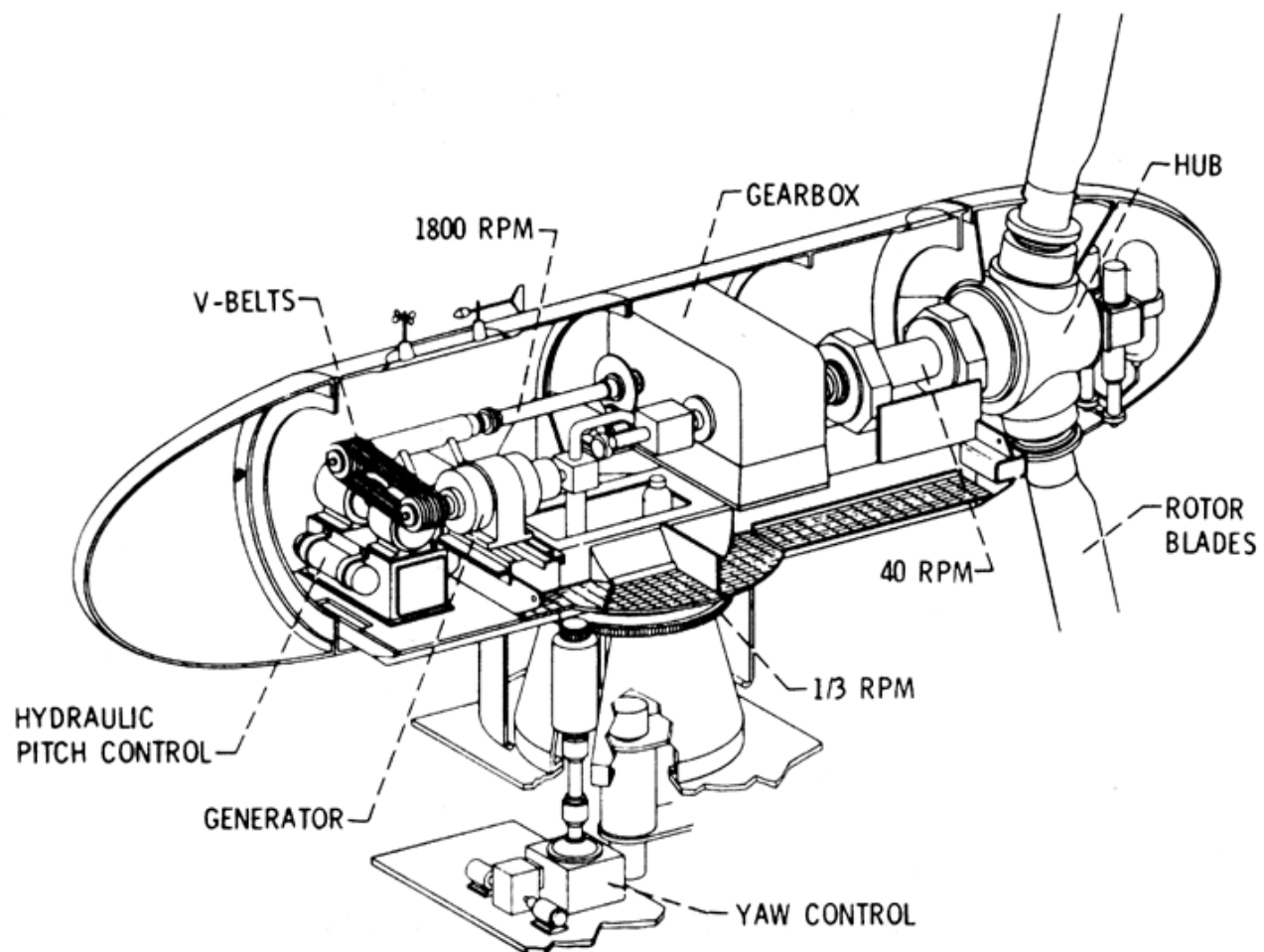
**Horizontal axis, 2 bladed  
Lattice tower**



**Horizontal axis, 3 bladed  
Tubular tower**

Unlike the Savonius type, commercial wind turbines designed to generate electricity utilize aerodynamic lift to provide their driving force. There are two broad categories of wind turbines: **horizontal axis (HAWT)** and **vertical axis (VAWT)**. Augmentation devices have been proposed for both types, but the additional energy extraction has not proven sufficient to justify the increased cost and complexity. VAWTs can receive wind from any direction and thus require no aligning devices. They also benefit from having their generator at ground level, which simplifies servicing. However, they do not exploit the typically increasing energy of the wind with height and must operate in the more turbulent regions of the earth's boundary layer. Additionally, severe alternating stresses are induced by the windstream in the blades as they rotate about the vertical shaft, making careful fatigue design important. HAWTs, on the other hand, can be placed on tall towers, which allow them to exploit the best winds. Both types typically have two or three blades, although HAWTs have been built with single blades and with four or more blades. A windform of modern HAWTs is shown in Fig. 70.2 and is depicted in Fig. 70.6 showing the major components.

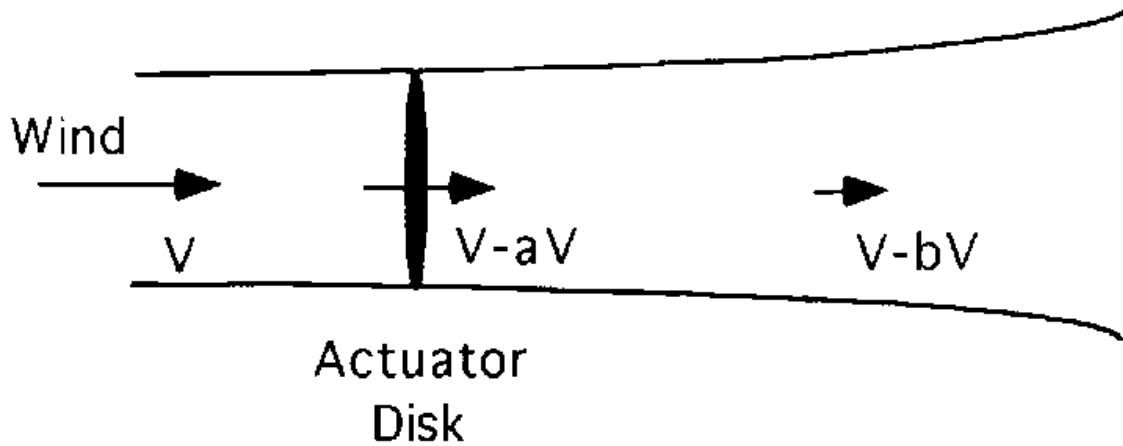
**Figure 70.6** Cut-away view of a typical horizontal axis wind turbine. (Source: Eldridge, F. R. 1975. *Wind Machines*, p. 30. National Science Foundation, Grant Number AER-75-12937. Energy Research and Development Administration, Washington, DC.)



## Basic Equations

Wind turbines extract energy from the air by reducing the freestream velocity. If one assumes incompressible flow and an ideal frictionless actuator disk placed in the airstream, as shown in Fig. 70.7, one can equate momentum and energy rates through the disk such that the flow downstream of the disk is slowed and  $b = 2a$ . This methodology is called **momentum theory** and is the theoretical basis for propeller, helicopter, and wind turbine rotor analysis.

**Figure 70.7** Actuator disk.



Since the thrust force at the disk can be computed by the total change in momentum across the disk, and power is thrust force times the air velocity at the disk, the mass flow is  $\rho AV(1 - a)$ . The power available at the disk is

$$P = 2\rho AV^3 a(1 - a)^2 \quad (70.2)$$

where  $a$  is a nondimensional fraction called the *axial induction factor*.

To calculate the efficiency of energy capture, a power coefficient is defined as

$$C_P = \frac{\text{Power extracted}}{\text{Power available}} \quad (70.3)$$

Using Eqs. (70.1)–(70.3),

$$C_P = 4a(1 - a)^2 \quad (70.4)$$

Since mass flow through the disk must be maintained for energy extraction, it is impossible to extract all of the energy available in the airstream by bringing the freestream velocity to zero. Thus there is a limit to the available energy capture, called the **Betz limit**. It can be calculated by taking the first derivative of Eq. 70.4, setting it equal to zero, and solving. Then,  $C_{P_{\max}} = 16/27 = 0.593$  when  $a = 1/3$ . Thus, optimum power extraction is limited to 59.3% of the power available in the wind stream and occurs when  $a = 1/3$  and  $b = 2/3$ . That is, the freestream wind velocity is slowed to 2/3 of its original value.

Likewise, the axial thrust force can be calculated and an axial force coefficient can be defined:

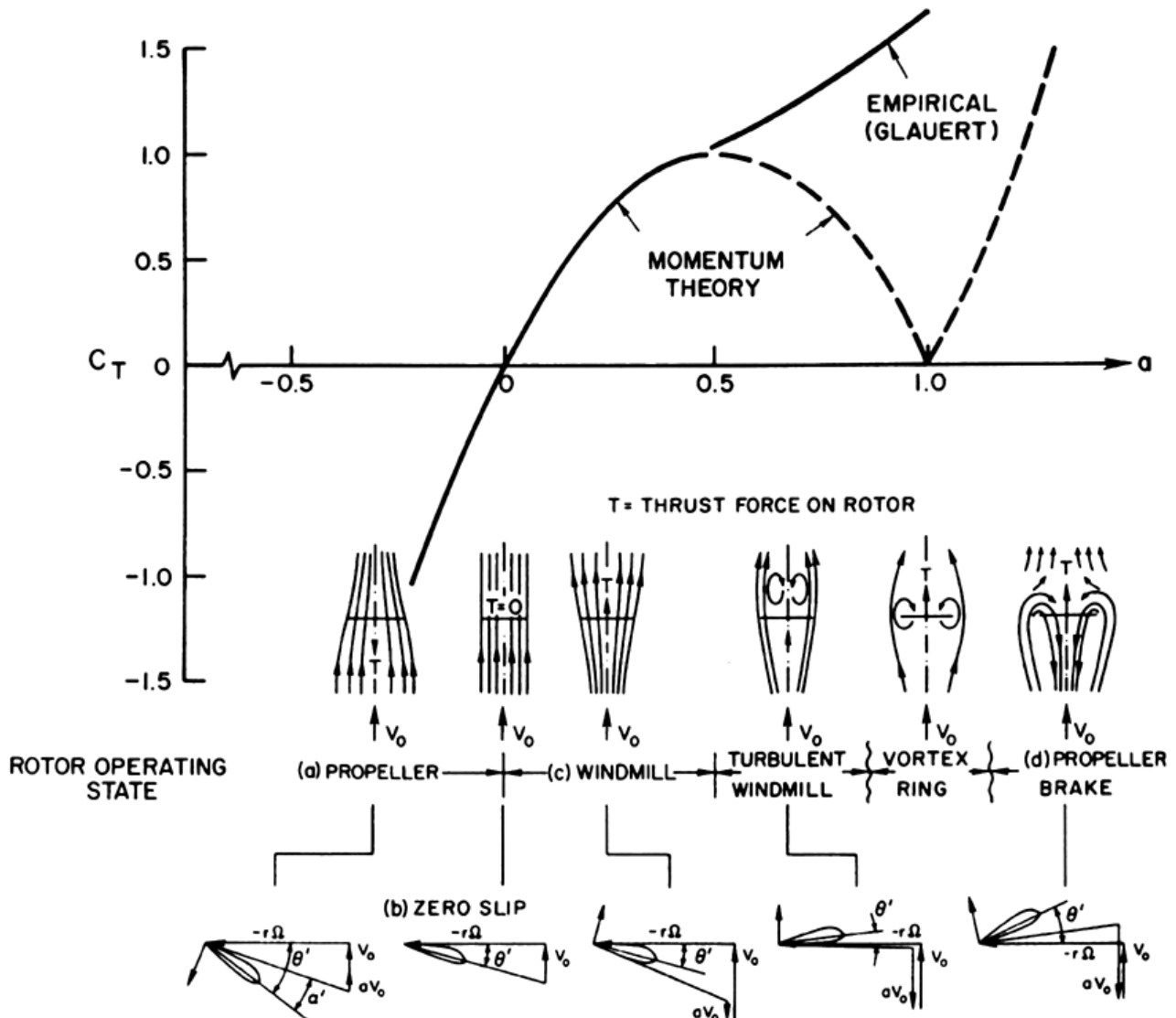
$$C_T = \frac{\text{Thrust}}{0.5\rho AV^2} \quad (70.5)$$

$$C_T = 4a(1 - a) \quad (70.6)$$

Finding the maximum for this equation shows that  $C_{T_{\max}}$  occurs when  $a = 0.5$ ,  $b = 1$ , and  $C_T$  has a value of 1.

Both of these relationships are constrained to induced flow factor  $a$  values less than or equal to 0.5, since, if  $a$  is greater than 0.5,  $b$  must be greater than 1, which would imply a flow reversal downstream of the actuator disk and momentum theory no longer applies. Figure 70.8 shows the flow states of a wind turbine rotor and the relationships of power and thrust coefficients to the axial induction factor,  $a$ . The curves show dotted lines for  $a > 0.5$ , since momentum theory does not apply, and indicate trends based on empirical data.

**Figure 70.8** Wind turbine operating states. (Source: Eggleston, D. M. and Stoddard, F. S. 1987. *Wind Turbine Engineering Design*, p. 32. Van Nostrand Reinhold, New York. With permission.)



Since momentum theory does not account for the real effects of drag and variable airfoil characteristics, **blade element theory** is often used to account for these effects. Figure 70.9 shows the cross section of an airfoil in rotation on a wind turbine rotor. Assuming axial flow, uniform inflow across the rotor disk, and rigid blade motion, the total velocity at the airfoil is the vector sum of the rotational velocity and the wind velocity at the disk. Using standard definitions of lift, drag, and angle of attack from airfoil theory, the driving force for the airfoil section can be expressed as

$$F = 0.5\rho(dA)\{[V(1 - a)]^2 + (\Omega r)^2\}(C_L \sin \phi - C_D \cos \phi) \quad (70.7)$$

whereas the thrust force for the section can be expressed as

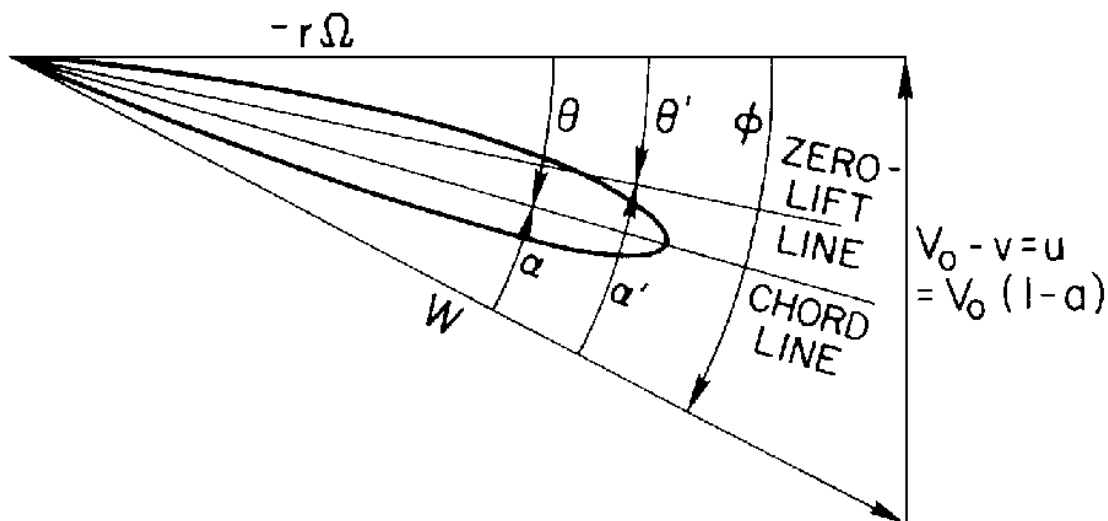
$$T = 0.5\rho(dA)\{[V(1 - a)]^2 + (\Omega r)^2\}(C_L \cos \phi + C_D \sin \phi) \quad (70.8)$$

where

- $F$  is the elemental driving force in newtons
- $dA$  is the elemental airfoil area, chord  $\times dr$  in  $\text{m}^2$
- $\Omega$  is the rotor speed in rad/s
- $r$  is the local radius of the element  $dA$  in m
- $C_L$  is the lift coefficient
- $C_D$  is the drag coefficient
- $T$  is the elemental thrust force in newtons

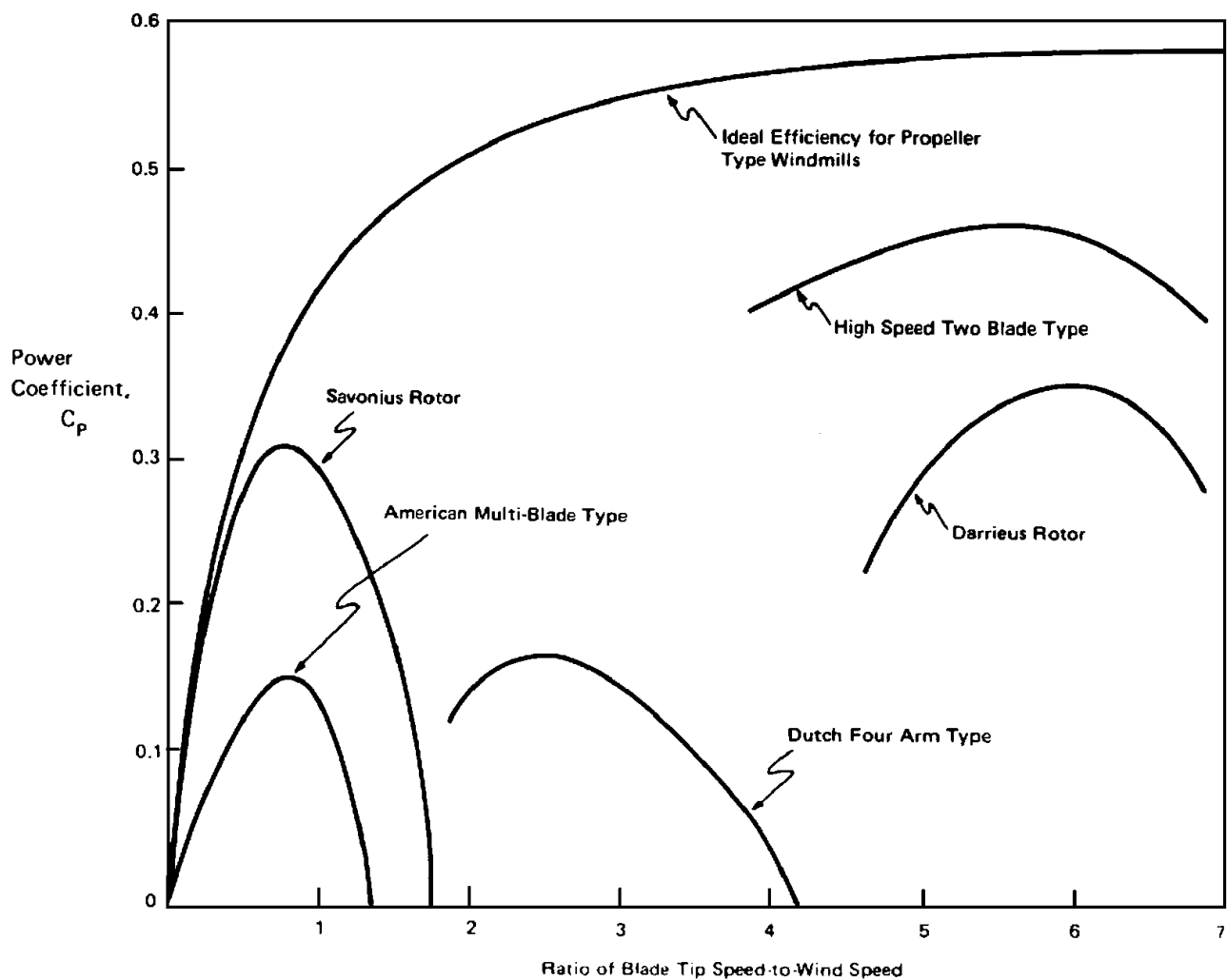
These equations show why lifting airfoil rotors, rather than drag-type rotors, are used in modern wind turbines. Drag devices are limited to the speed of the wind, whereas a lifting rotor can utilize the vector sum of the rotational speed and the wind speed, giving a power magnification on the order of a factor of 100 [Rohatgi and Nelson, 1994].

**Figure 70.9** Blade element diagram. (Source: Eggleston, D. M. and Stoddard, F. S. 1987. *Wind Turbine Engineering Design*, p. 31. Van Nostrand Reinhold, New York. With permission.)



Rotor power curves are typically plotted against the nondimensional tip speed ratio, that is, the ratio of the speed of the blade tip and the freestream wind speed. These curves are unique for a given rotor with fixed pitch and can be performance-tailored by modifying the airfoil section characteristics, pitch angle of the blades, and the blade twist over the span of the rotor. Sample power coefficient–tip speed ratio curves for various rotors are shown in Fig. 70.10. One will note that high-speed rotors with low solidity (solidity is the ratio of actual blade airfoil area, length times average chord, to rotor disk–swept area) have higher conversion efficiency, although they generally produce less torque than high-solidity rotors, especially at start-up.

**Figure 70.10** Typical performance curves for various wind turbine configurations. (Source: Eldridge, F. R. 1975. *Wind Machines*, p. 55. National Science Foundation, Grant Number AER-75-12937. Energy Research and Development Administration, Washington, DC.)



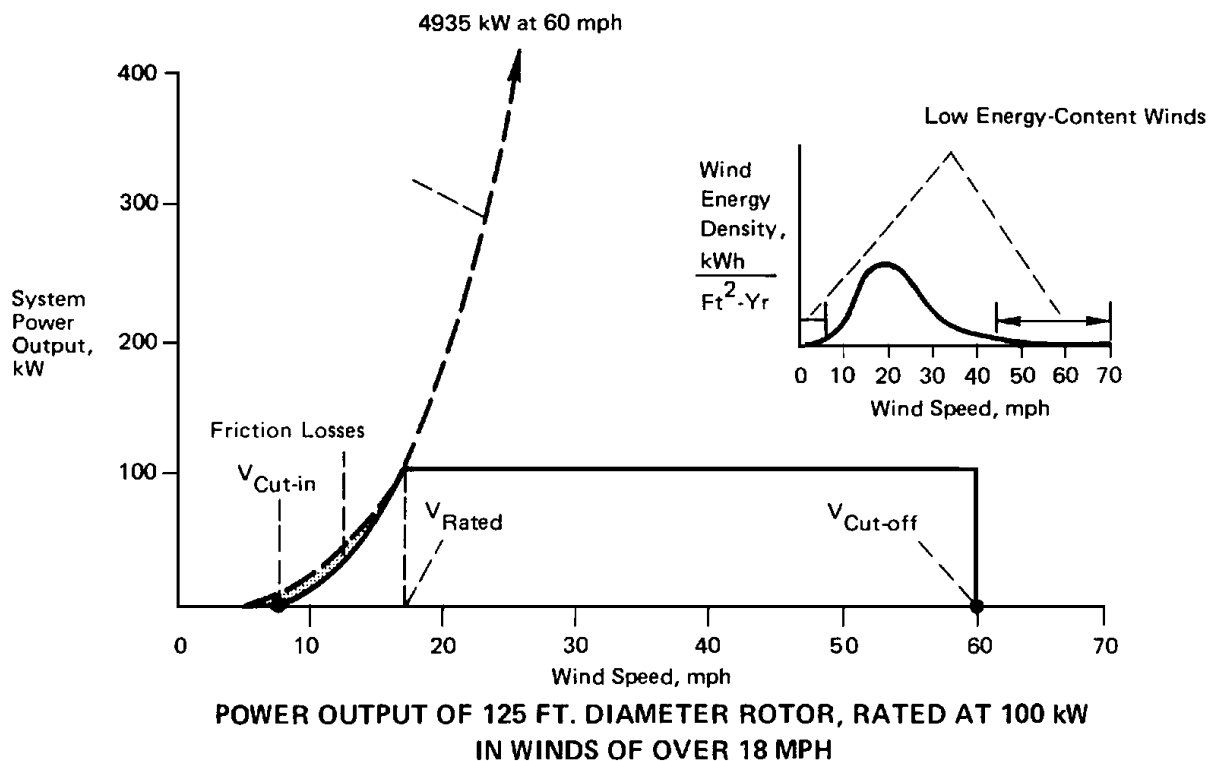
**TYPICAL PERFORMANCES OF WIND MACHINES**

If blade element theory is combined with momentum theory, an algorithm can be established to calculate the local thrust and torque values for airfoil sections of a rotor. Using this algorithm, one can estimate rotor performance for all tip speed ratios and generate power and thrust curves for rotor design. This algorithm is available in the literature for wind turbines using a computer code called PROPSH [Tangler, 1983]. A similar analysis can be completed for a VAWT but is beyond

the scope of this article, although the resulting performance curves are similar.

In order to generate a calculated wind turbine power curve that is, power versus wind speed curve, one must include the efficiency of the drive train and generator with the values of rotor performance calculated earlier. At low wind speeds and power ratings, generator efficiency is usually rather low but increases rapidly as power output increases with wind speed. Drive train and generator efficiencies near 90% are not uncommon near rated power. A typical wind turbine power curve is shown in Fig. 70.11.

**Figure 70.11** Wind turbine power curve. (Source: Eldridge, F. R. 1975. *Wind Machines*, p. 56. National Science Foundation, Grant Number AER-75-12937. Energy Research and Development Administration, Washington, DC.)



## 70.2 Power Regulation and Control

There are several power regulation and control issues that must be addressed in a modern HAWT. When the wind speed increases to a value at which the generator is producing rated power, some control action must occur so that the generator does not exceed its rated capacity and overheat. Typical methods of power regulation at **rated wind speed** are **pitch** control, **stall** control, and **yaw** control. Pitch control is accomplished by providing rotating bearings at the blade root and

allowing the blade to change its pitch angle relative to the wind, thus regulating power. Stall control is accomplished by designing the rotor so that aerodynamic stall is reached at rated wind speed and the rotor power is limited by airfoil stall. Yaw control turns the rotor out of the wind at rated wind speed and regulates power by reducing the rotor area exposed to the wind.

In addition to power regulation, loss of generator load and overspeed protection is required in case the load on the rotor is lost during operation. Loss of load will allow the rotor speed and thrust to rise rapidly and will result in turbine destruction in moderate winds if not brought under control rapidly. Pitch control rotors usually have a fail-safe pitch change to the feather position in an emergency shutdown, whereas stall control turbines use rotor tip brakes, a large mechanical brake that activates upon overspeed, or a combination of the two. Yaw control turbines simply yaw out of the wind to control overspeed.

Additional controls regulate turbine cut-in when the wind increases above starting wind speed, and cut-out in very high winds or in case of excessive vibration or other problem. Finally, upwind HAWTs require a yaw drive motor to remain aligned with the wind, whereas downwind turbines may have a yaw drive or operate in a free yaw mode to follow wind direction changes.

## 70.3 Energy Capture Estimation

---

Once a turbine power curve is established, it can be used with wind frequency information in order to estimate the annual energy production available at a given site. Wind speeds fluctuate continuously, but have been shown to generally follow a Weibull frequency distribution. One special type of Weibull distribution is the Raleigh distribution, which is often used for energy capture estimation. The Raleigh distribution is based only on average wind speed at the hub height where the turbine is to be located and can be calculated using the following equation:

$$f(V) = \frac{\pi}{2} \frac{V}{\bar{V}^2} \exp \left[ -\frac{\pi}{4} \left( \frac{V}{\bar{V}} \right)^2 \right] \quad (70.9)$$

where  $V$  is the center of a wind speed bin of width 1 m/s, ft/s, and so on in consistent units, and  $\bar{V}$  is the average wind speed, also in consistent units.

Due to the atmospheric boundary layer, wind speed varies with height above the ground. If the wind speed is known at some height other than hub height, hub height wind speed can be estimated using a power law, Eq. (70.10). An exponent of 1/7 is typically used, but the wind shear with height can vary widely, especially in complex terrain.

$$\frac{V(z)}{V(z_{\text{ref}})} = \left[ \frac{z}{z_{\text{ref}}} \right]^a \quad (70.10)$$

where, using consistent units,

$V(z)$  is the desired wind speed  
 $V(z_{\text{ref}})$  is the reference wind speed  
 $z$  is the desired height



$z_{\text{ref}}$  is the reference height  
 $a$  is the power exponent, usually 1/7

Assuming a Raleigh distribution at hub height, annual energy can be estimated by first adjusting the power curve for air density variation due to site elevation, if required, and then using the Raleigh frequency distribution and power curve to determine the number of hours per year the turbine will operate at each wind speed interval and summing the number of operating hours times the power level for each interval over all wind speeds.

## 70.4 Stand-Alone Applications

---

Small stand-alone windmills have long been a feature of the Great Plains, where they provided water and allowed settlement. These water pumpers have proven extremely robust and found application throughout the world. Their successors offer the prospect of electric power not only to the larger electric grid, but also to remote locations where the cost of utility connection has proven uneconomic. Coupled with solar electric panels, they are capable of providing reliable power for basic necessities such as water pumping, lighting, and the refrigeration of medical supplies. [Figure 70.12](#) shows a typical hybrid system, featuring a 10 kW wind turbine.

## 70.5 Cost of Energy Calculations

---

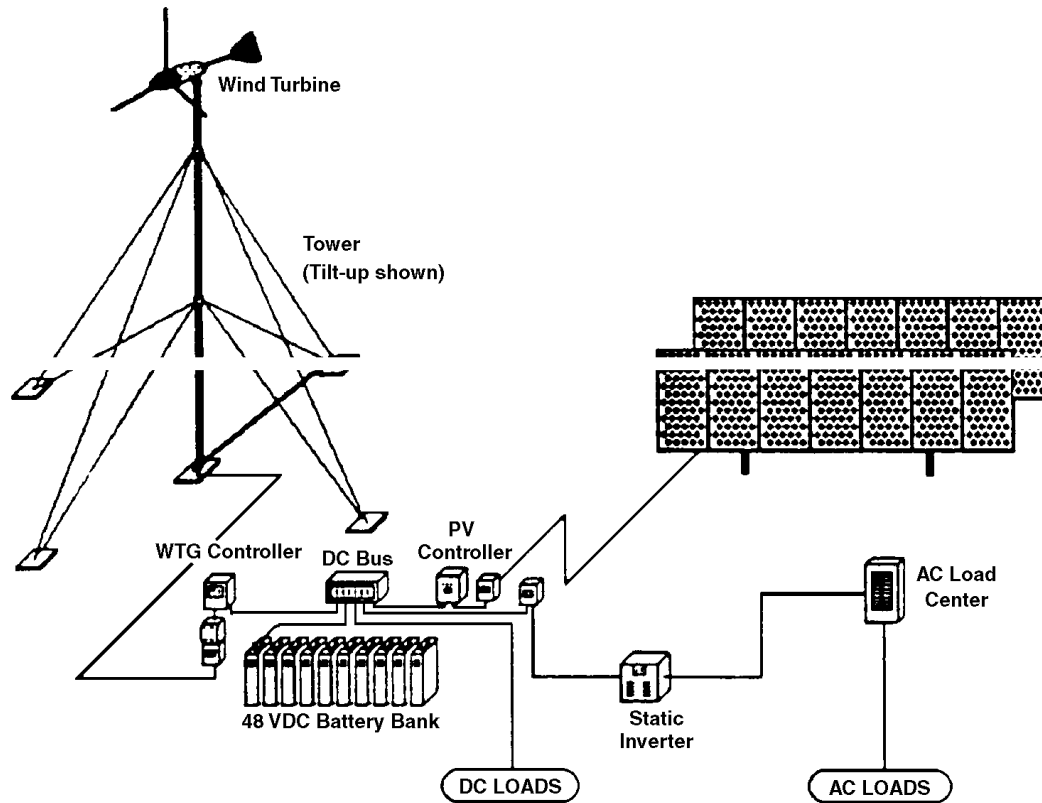
The cost of energy from a wind turbine is calculated by adding the life cycle annual cost of the capital equipment; annualized equipment replacement costs; annual operation and maintenance costs; and annual land, tax, and insurance costs. This total is then divided by the annual energy capture. Units are usually expressed in dollars per kWh. The equation is as follows:

$$\text{COE} = \frac{\text{Annual capital and operating costs}}{\text{Annual energy capture}} \quad (70.11)$$

where COE is the cost of energy.

Since wind power generation is dependent on the speed of the wind, which is variable, wind power is seen as a nondispatchable energy source and there is little value given for capacity credit by the utility. Studies have shown, however, that for large wind farms some capacity credit may

**Figure 70.12** Example of a stand-alone hybrid wind system. (Source: Bergey, M. L. S. 1989. *An Overview of Wind Power for Remote Site Telecommunications Facilities*, p. 11. Renewable Energy Power Supplies for Telecommunications Conference, British Wind Energy Association, London. With permission.)



be given since the addition of the wind farm and average predictability of the wind provide an additional generation source such that less reserve capacity is required by the utility. True dispatchability of wind power, however, requires a storage capability that is technically feasible but adds to the cost of the wind power plant and must then be added to the cost of energy.

## 70.6 Environmental and Social Cost Issues

---

Environmental and social impacts for wind energy systems are usually less severe than those for conventional power plants but are an important consideration in the selection of wind turbines for an energy project. For wind turbines these impacts include aesthetic issues, such as disruption of a landscape view, road construction, noise generation, and potential bird kills from impact with the turning rotor. These impacts must then be weighed with the social and environmental impacts of other power sources, such as carbon dioxide and acid rain precursor emissions from coal and gas fired power plants, hazardous waste storage from nuclear plants, water use by thermal plants in arid areas, and so on. **Integrated resource planning** is a relatively new planning tool used by utilities and regulators that quantifies these social and environmental costs so that the least total cost alternative can be selected.

## 70.7 Summary

---

Modern wind turbine technology is advancing rapidly and costs are decreasing. The value of a clean, non-combustion-based power generation technology is also seen as a positive factor in wind power development, as is the lack of water consumption in arid areas. The fact that wind power is not dispatchable on demand by the utility, but tied to the variances in the wind, is seen as a drawback to dominance of the technology in the power production industry. However, this circumstance is mitigated somewhat in the aggregate of many wind farms connected to the utility grid, whereby the variations of wind generation are reduced and predictability of the generation is enhanced.

### Defining Terms

**Betz limit:** Maximum energy available from the wind by a rotor.

**Blade element theory:** Assumes rotor performance can be calculated by elemental analysis of the rotor blade.

**Horizontal axis wind turbine (HAWT):** Turbine in which the rotor shaft is horizontal and the rotor disk is vertical and must be aligned with the wind.

**Integrated resource planning:** A method of examining power generation alternatives that includes all costs of generation, such as environmental and social costs, to determine the total least cost.

**Momentum theory:** Assumes that the power extracted at a rotor is determined by Newton's theory of momentum; that is, force is proportional to the change in momentum of the airstream.

**Pitch:** The angle that a rotor blade is set relative to the plane of the rotor.

**PROPSH:** A popular computer code that calculates rotor performance based on blade element and momentum theory.

**Rated wind speed:** The wind speed where rated power is reached for a given wind turbine.

**Stall:** Airfoil stall occurs when flow over an airfoil separates, lift decreases, and drag increases.

**Vertical axis wind turbine (VAWT):** Turbine in which the shaft is vertical and the rotor spins about the shaft. Wind alignment is not required.

**Wind farm:** A group of wind turbines arranged for efficient power production and delivering electric power to a grid.

**Yaw:** The angle between the normal to the rotor disk and the wind for a HAWT.

## References

- Bergey, M. L. S. 1989. *An Overview of Wind Power for Remote Site Telecommunications Facilities*. Renewable Energy Power Supplies for Telecommunications Conference, British Wind Energy Association, London.
- Brown, L. R., Kane, H., and Ayres, E. 1993. *Vital Signs, 1993*. Worldwatch Institute, Norton, New York.
- Brown, L. R., Kane, H., and Roodman, D. M. 1994. *Vital Signs, 1994*. Worldwatch Institute, Norton, New York.
- Eggleston, D. M. and Stoddard, F. S. 1987. *Wind Turbine Engineering Design*. Van Nostrand Reinhold, New York.
- Eldridge, F. R. 1975. *Wind Machines*. National Science Foundation, Grant Number AER-75-12937. Energy Research and Development Administration, Washington, DC.
- Elliott, D. L., Holladay, C. G., Barchet, W. R., Foote, H. D., and Sandusky, W. F. 1986. *Wind Energy Resource Atlas of the United States*. DOE/CH 10093-4. Solar Technical Information Program, National Renewable Energy Laboratory, Golden, CO.
- Koepl, G. W. 1982. *Putnam's Power from the Wind*, 2nd ed. Van Nostrand Reinhold, New York.
- Righter, R. 1994. *Wind Energy in America: A History*. Draft manuscript, University of Texas at El Paso, to be published in 1996 by the University of Oklahoma Press, Norman, OK.
- Rohatgi, J. S. and Nelson, V. 1994. *Wind Characteristics: An Analysis for the Generation of Wind Power*. Alternative Energy Institute, West Texas A&M University, Canyon, TX.
- Tangler, J. L. 1983. *Horizontal Axis Wind Turbine Performance Prediction Code PROPSH*. Rocky Flats Wind Research Center, National Renewable Energy Laboratory, Golden, CO.
- Utility Wind Interest Group. 1992. *America Takes Stock of a Vast Energy Resource*. Electric Power Research Institute, Palo Alto, CA.

## Further Information

For those wishing to expand their knowledge of this topic, the following sources are highly recommended:

- Freris, L. L. 1990. *Wind Energy Conversion Systems*. Prentice Hall, Hemel Hempstead, UK.
- Spera, D. A. 1994. *Wind Turbine Technology: Fundamental Concepts of Wind Turbine Engineering*. ASME, New York.

Gipe, P. 1993. *Wind Power for Home and Business*. Chelsea Green, Post Mills, VT.  
Gipe, P. 1995. *Wind Energy Comes of Age*. John Wiley & Sons, New York.  
*Windpower Monthly News Magazine*, U.S. Office, P.O. Box 496007, Suite 217, Reading, CA  
96049.

Arndt R. E. A. "Hydraulic Turbines"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Hydraulic Turbines

---

## 71.1 General Description

Typical Hydropower Installation • Turbine Classification

## 71.2 Principles of Operation

Power Available, Efficiency • Similitude and Scaling Formulas

## 71.3 Factors Involved in Selecting a Turbine

Performance Characteristics • Speed Regulation • Cavitation and Turbine Setting

### **Roger E. A. Arndt**

*St. Anthony Falls Laboratory University of Minnesota*

A hydraulic turbine is a mechanical device that converts the potential energy associated with a difference in water elevation (**head**) into useful work. Modern hydraulic turbines are the result of many years of gradual development. Economic incentives have resulted in the development of very large units (exceeding 800 megawatts in capacity) with efficiencies that are sometimes in excess of 95%.

The emphasis on the design and manufacture of very large turbines is shifting to the production of smaller units, especially in developed nations, where much of the potential for developing large base-load plants has been realized. At the same time, the escalation in the cost of energy has made many smaller sites economically feasible and has greatly expanded the market for smaller turbines. The increased value of energy also justifies the cost of refurbishment and increasing the capacity of older facilities. Thus, a new market area is developing for updating older turbines with modern replacement runners having higher efficiency and greater capacity.

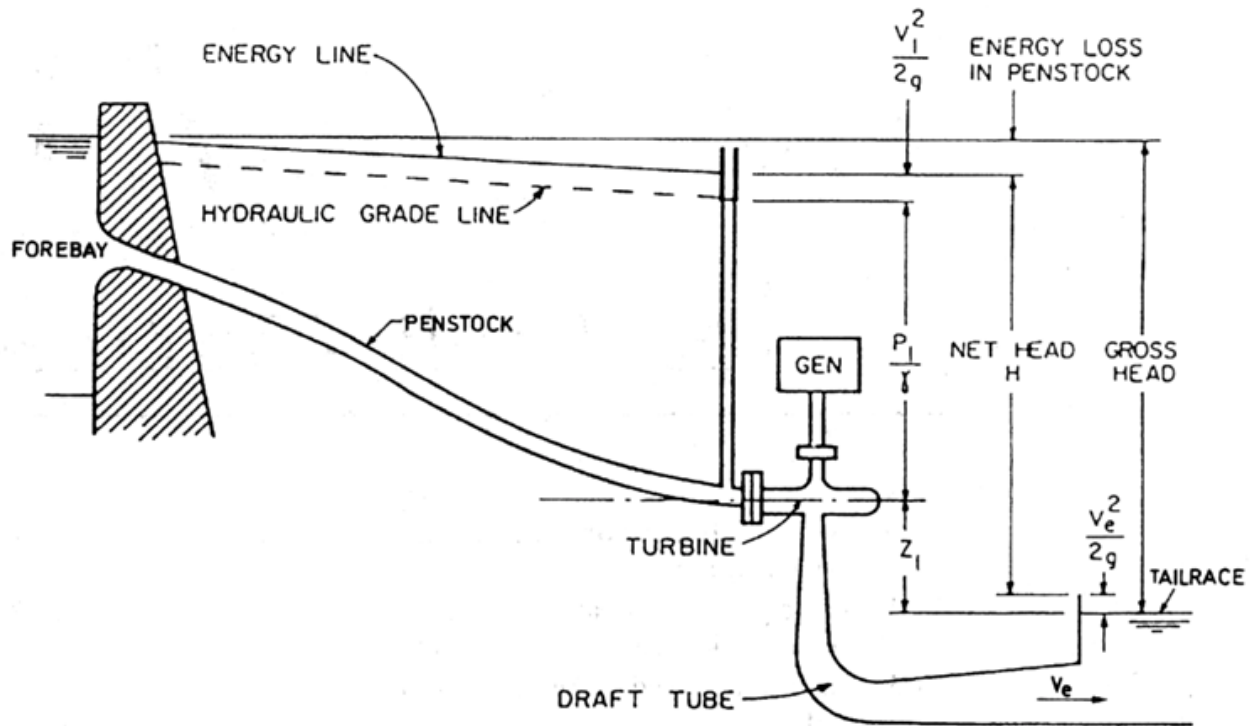
## 71.1 General Description

---

### **Typical Hydropower Installation**

As shown schematically in [Fig. 71.1](#), the hydraulic components of a hydropower installation consist of an intake, penstock, guide vanes or distributor, turbine, and draft tube. Trash racks are commonly provided to prevent ingestion of debris into the turbine. Intakes usually require some type of shape transition to match the passageway to the turbine and also incorporate a gate or some other means of stopping the flow in case of an emergency or turbine maintenance. Some types of turbines are set in an open flume; others are attached to a closed-conduit penstock.

**Figure 71.1** Schematic of a hydropower installation.



## Turbine Classification

There are two types of turbines, denoted as impulse and reaction. In an *impulse turbine* the available head is converted to kinetic energy before entering the runner, the power available being extracted from the flow at approximately atmospheric pressure. In a *reaction turbine* the runner is completely submerged and both the pressure and the velocity decrease from inlet to outlet. The velocity head in the inlet to the turbine runner is typically less than 50% of the total head available.

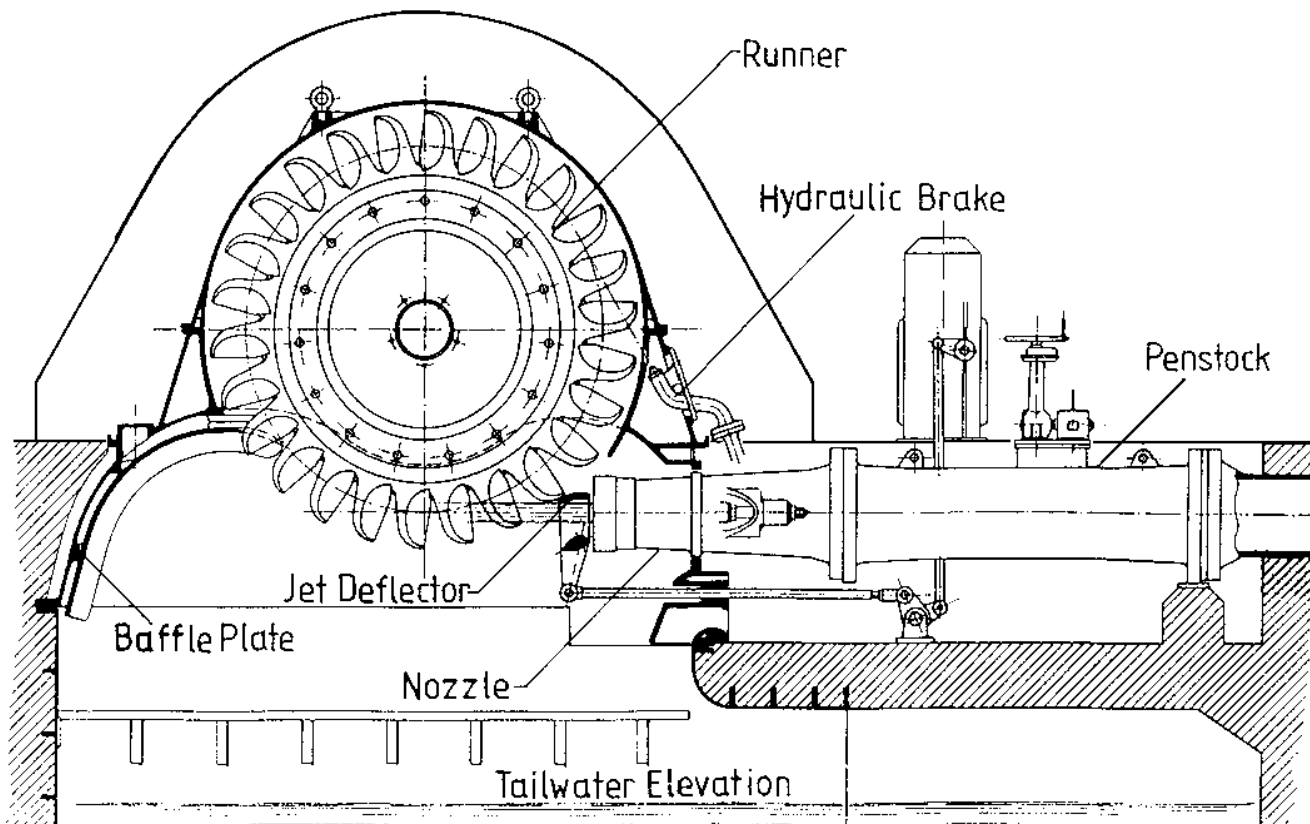
### Impulse Turbines

Modern impulse units are generally of the Pelton type and are restricted to relatively high head applications (Fig. 71.2). One or more jets of water impinge on a wheel containing many curved buckets. The jet stream is directed inwardly, sideways, and outwardly, thereby producing a force on the bucket, which in turn results in a torque on the shaft. All kinetic energy leaving the runner is "lost." A **draft tube** is generally not used since the runner operates under approximately atmospheric pressure and the head represented by the elevation of the unit above tailwater cannot be utilized. (In principle, a draft tube could be used, which requires the runner to operate in air under reduced pressure. Attempts at operating an impulse turbine with a draft tube have not met with much success.) Since this is a high-head device, this loss in available head is relatively unimportant. As will be shown later, the Pelton wheel is a low-specific speed device. Specific speed can be increased by the addition of extra nozzles, the specific speed increasing by the square



root of the number of nozzles. Specific speed can also be increased by a change in the manner of inflow and outflow. Special designs such as the Turgo or crossflow turbines are examples of relatively high specific speed impulse units [Arndt, 1991].

**Figure 71.2** Cross section of a single-wheel, single-jet Pelton turbine. This is the third highest head Pelton turbine in the world,  $H = 1447$  m,  $n = 500$  rpm,  $P = 35.2$  MW,  $N_s \sim 0.038$ . (Courtesy of Vevey Charmilles Engineering Works. Adapted from Raabe, J. 1985. *Hydro Power: The Design, Use, and Function of Hydromechanical, Hydraulic, and Electrical Equipment*. VDI Verlag, Dusseldorf, Germany.)

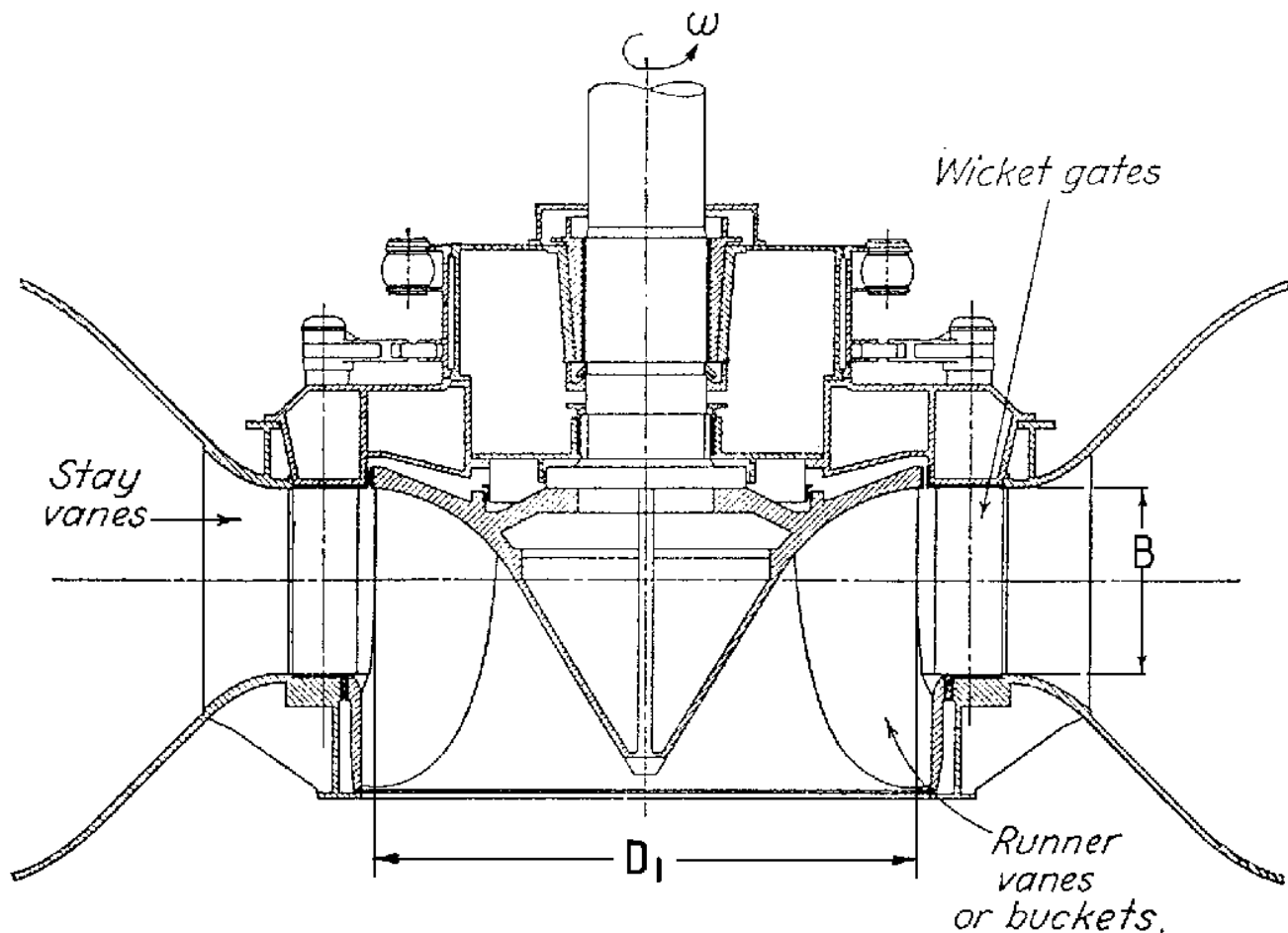


Most Pelton wheels are mounted on a horizontal axis, although newer vertical-axis units have been developed. Because of physical constraints on orderly outflow from the unit, the number of nozzles is generally limited to six or less. Whereas the power of a reaction turbine is controlled by the **wicket gates**, the power of the Pelton wheel is controlled by varying the nozzle discharge by means of an automatically adjusted needle, as illustrated in Fig. 71.2. Jet deflectors or auxiliary nozzles are provided for emergency unloading of the wheel. Additional power can be obtained by connecting two wheels to a single generator or by using multiple nozzles. Since the needle valve can throttle the flow while maintaining essentially constant jet velocity, the relative velocities at entrance and exit remain unchanged, producing nearly constant efficiency over a wide range of power output.

## Reaction Turbines

Reaction turbines are classified according to the variation in flow direction through the runner. In radial- and mixed-flow runners, the flow exits at a radius different from the radius at the inlet. If the flow enters the runner with only radial and tangential components, it is a radial-flow machine. The flow enters a mixed-flow runner with both radial and axial components. Francis turbines are of the radial- and mixed-flow type, depending on the design specific speed. A Francis turbine is illustrated in Fig. 71.3.

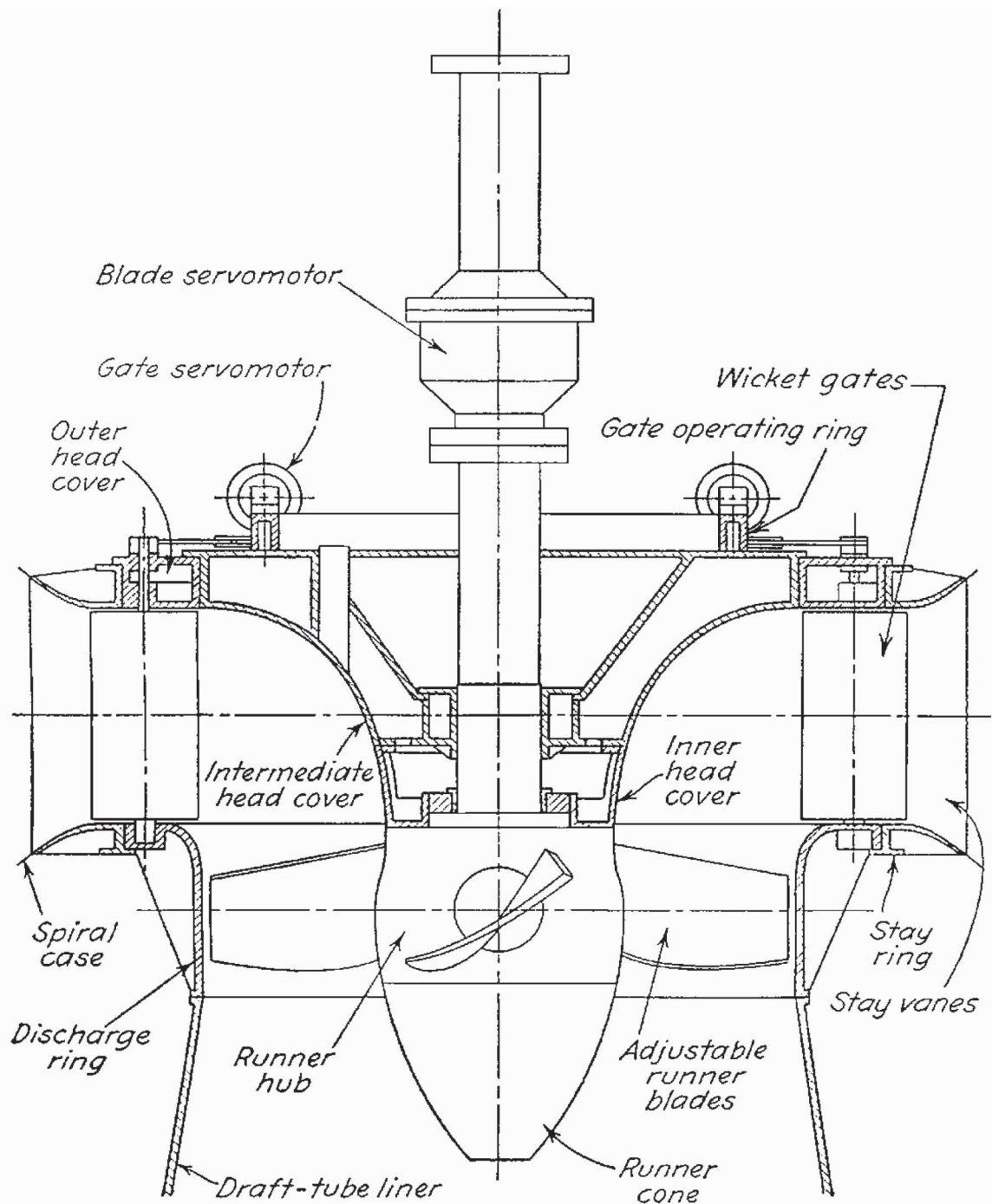
**Figure 71.3** Francis turbine,  $N_s \sim 0.66$ . (Adapted from Daily, J. W. 1950. Hydraulic machinery. In *Engineering Hydraulics*, ed. H. Rouse. Wiley, New York. Reprinted with permission.)



Axial-flow propeller turbines are generally either of the fixed-blade or Kaplan (adjustable-blade) variety. The "classical" propeller turbine, illustrated in Fig. 71.4, is a vertical-axis machine with a scroll case and a radial wicket gate configuration that is very similar to the flow inlet for a Francis turbine. The flow enters radially inward and makes a right-angle turn before entering the runner in an axial direction. The Kaplan turbine has both adjustable runner blades and adjustable wicket gates. The control system is designed so that the variation in blade angle is coupled with the wicket gate setting in a manner that achieves best overall

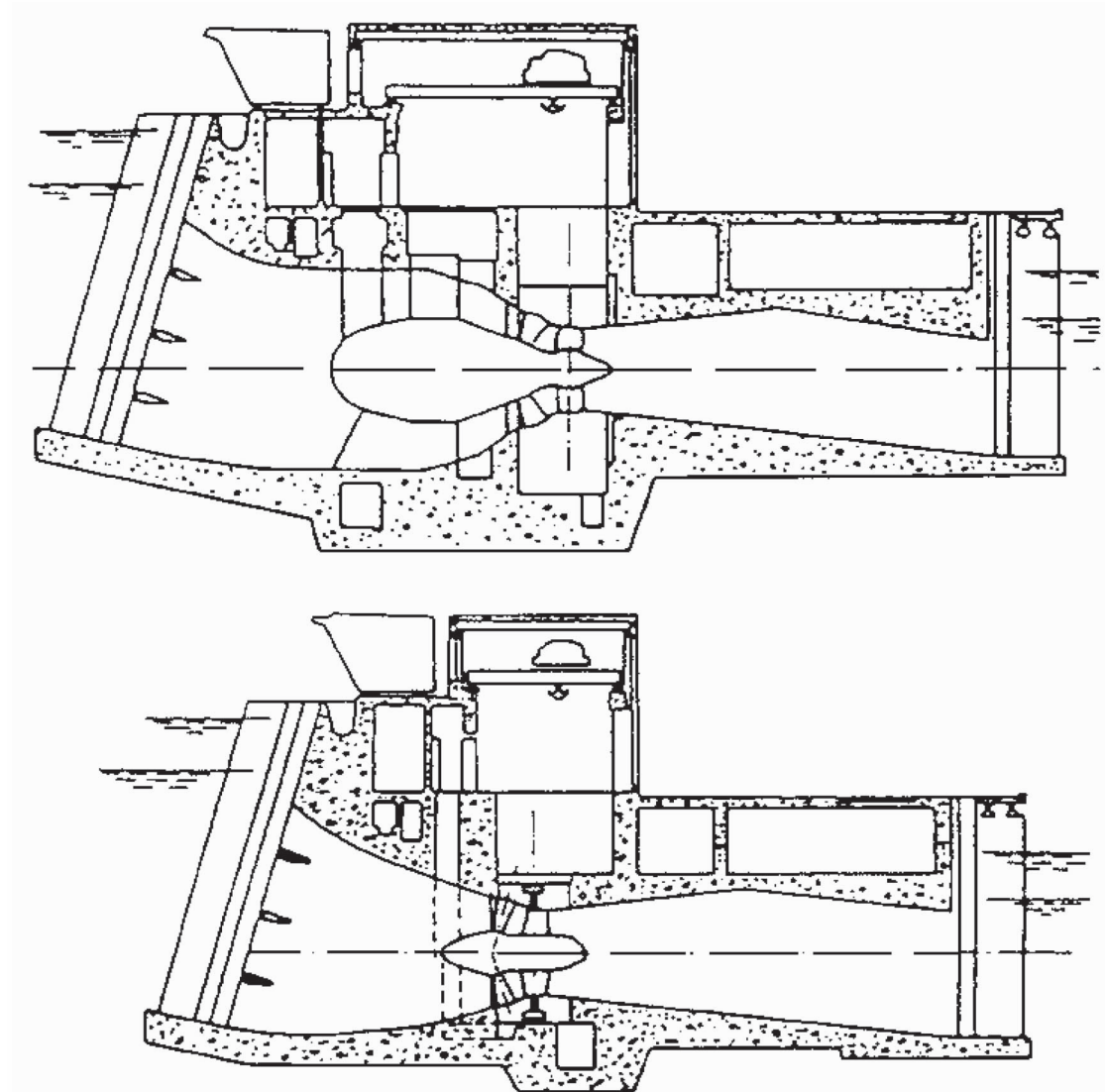
efficiency over a wide range of flow rates.

**Figure 71.4** Smith-Kaplan axial-flow turbine with adjustable-pitch runner blades,  $N_S \sim 2.0$ . (From Daily, J. W. 1950. *Hydraulic machinery*. In *Engineering Hydraulics*, ed. H. Rouse. Wiley, New York. Reprinted with permission.)



Some modern designs take full advantage of the axial-flow runner; these include the tube, bulb, and Straflo types illustrated in Fig. 71.5. The flow enters and exits the turbine with minor changes in direction. A wide variation in civil works design is also permissible. The tubular type can be fixed-propeller, semi-Kaplan, or fully adjustable. An externally mounted generator is driven by a shaft that extends through the flow passage either upstream or downstream of the runner. The bulb turbine was originally designed as a high-output, low-head unit. In large units, the generator is housed within the bulb and is driven by a variable-pitch propeller at the trailing end of the bulb. Pit turbines are similar in principle to bulb turbines, except that the generator is not enclosed in a fully submerged compartment (the bulb). Instead, the generator is in a compartment that extends above water level. This improves access to the generator for maintenance.

**Figure 71.5** Comparison between bulb (upper) and Straflo (lower) turbines. (Courtesy U.S. Dept. of Energy.)



## 71.2 Principles of Operation

---

### Power Available, Efficiency

The power that can be developed by a turbine is a function of both the head and flow available:

$$P = \eta \rho g Q H \quad (71.1)$$

where  $\eta$  is the turbine efficiency,  $\rho$  is the density of water ( $\text{kg}/\text{m}^3$ ),  $g$  is the acceleration due to gravity ( $\text{m}/\text{s}^2$ ),  $Q$  is the flow rate ( $\text{m}^3/\text{s}$ ), and  $H$  is the net head in meters. *Net head* is defined as the difference between the *total head* at the inlet to the turbine and the total head at the tailrace, as illustrated in Fig. 71.1. Various definitions of net head are used in practice, which depend on the value of the exit velocity head,  $V_e^2/2g$ , that is used in the calculation. The International Electrotechnical Test Code uses the velocity head at the draft tube exit.

The efficiency depends on the actual head and flow utilized by the turbine runner, flow losses in the draft tube, and the frictional resistance of mechanical components.

### Similitude and Scaling Formulas

Under a given head, a turbine can operate at various combinations of speed and flow depending on the inlet settings. For reaction turbines the flow into the turbine is controlled by the wicket gate angle,  $\alpha$ . The flow is typically controlled by the nozzle opening in impulse units. Turbine performance can be described in terms of nondimensional variables,

$$\psi = \frac{2gH}{\omega^2 D^2} \quad (71.2)$$

$$\phi = \frac{Q}{\sqrt{2gH} D^2} \quad (71.3)$$

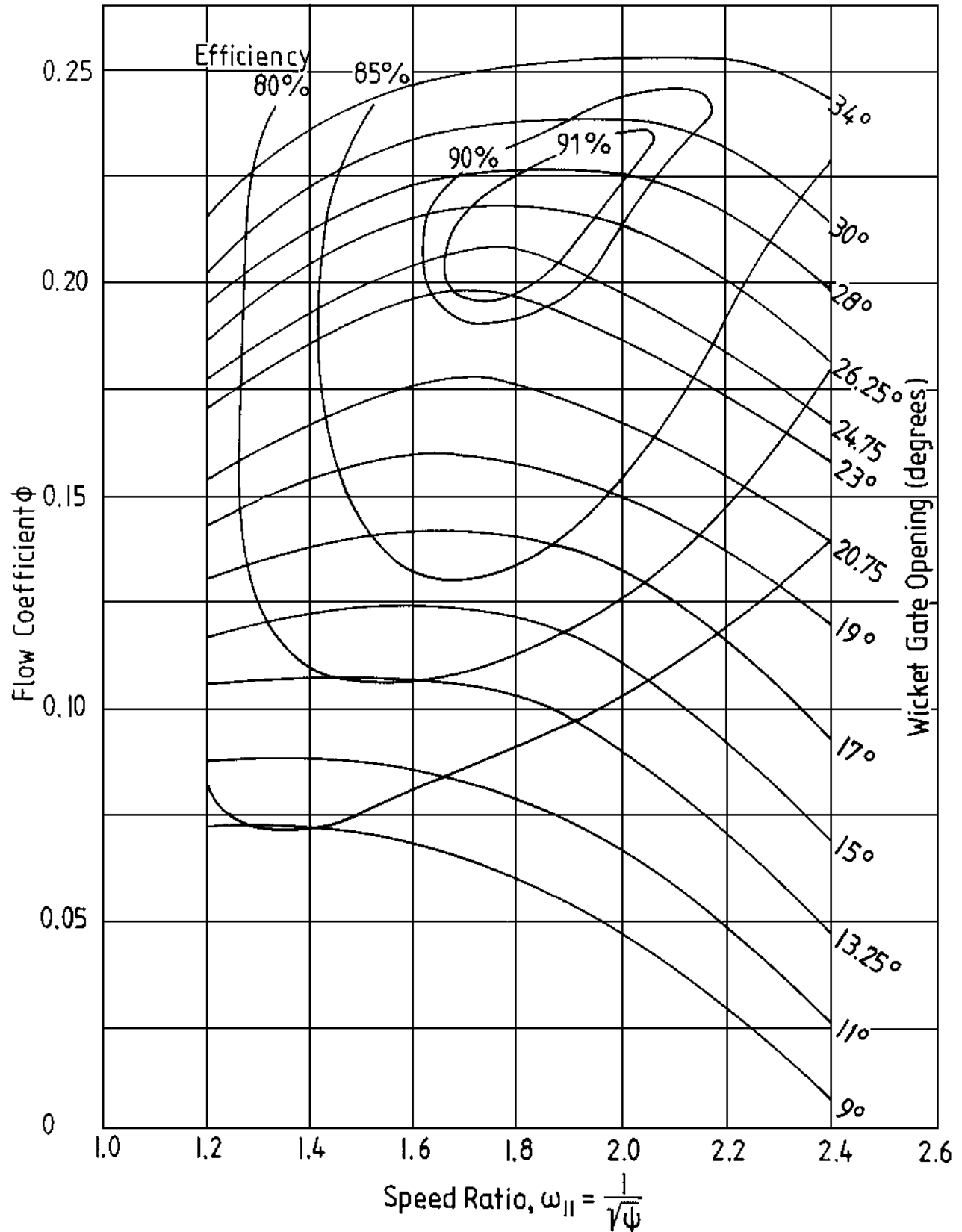
where  $\omega$  is the rotational speed of the turbine in radians per second and  $D$  is the diameter of the turbine.

The hydraulic efficiency of the runner alone is given by

$$\eta_h = \frac{\phi}{\sqrt{\psi}} (C_1 \cos \alpha_1 - C_2 \cos \alpha_2) \quad (71.4)$$

where  $C_1$  and  $C_2$  are constants that depend on the specific turbine configuration, and  $\alpha_1$  and  $\alpha_2$  are the inlet and outlet angles that the absolute velocity vectors make with the tangential direction. The value of  $\cos \alpha_2$  is approximately zero at peak efficiency. The terms  $\phi$ ,  $\psi$ ,  $\alpha_1$ , and  $\alpha_2$  are interrelated. Using model test data, isocontours of efficiency can be mapped in the  $\phi\psi$  plane. This is typically referred to as a *hill diagram*, as shown in Fig. 71.6.

**Figure 71.6** Typical hill diagram. (Adapted from Wahl, T. L. 1994. Draft tube surging times two: The twin vortex problem. *Hydro Rev.* 13(1):60–69, 1994. With permission.)



The **specific speed** is defined as

$$N_s = \frac{\omega \sqrt{Q}}{(2gH)^{3/4}} = \sqrt{\frac{\phi}{\psi}} \quad (71.5)$$

A given specific speed describes a specific combination of operating conditions that ensures similar flow patterns and the same efficiency in geometrically similar machines regardless of the size and rotational speed of the machine. It is customary to define the design specific speed in terms of the value at the design head and flow where peak efficiency occurs. The value of specific speed so defined permits a classification of different turbine types.

The specific speed defined herein is dimensionless. Many other forms of specific speed exist that are dimensional and have distinct numerical values depending on the system of units used [Arndt, 1991]. (The literature also contains two other minor variations of the dimensionless form. One differs by a factor of  $1/\pi^{1/2}$  and the other by  $2^{3/4}$ .) The similarity arguments used to arrive at the concept of specific speed indicate that a given machine of diameter  $D$  operating under a head  $H$  will discharge a flow  $Q$  and produce a torque  $T$  and power  $P$  at a rotational speed  $\omega$  given by

$$Q = \phi D^2 \sqrt{2gH} \quad (71.6)$$

$$T = T_{11} \rho D^3 2gH \quad (71.7)$$

$$P = P_{11} \rho D^2 (2gH)^{3/2} \quad (71.8)$$

$$\omega = \frac{2u_1}{D} = \omega_{11} \frac{\sqrt{2gH}}{D}, \quad \left[ \omega_{11} = \frac{1}{\sqrt{\psi}} \right] \quad (71.9)$$

with

$$P_{11} = T_{11} \omega_{11} \quad (71.10)$$

where  $T_{11}$ ,  $P_{11}$ , and  $\omega_{11}$  are also nondimensional. (The reader is cautioned that many texts, especially in the American literature, contain dimensional forms of  $T_{11}$ ,  $P_{11}$ , and  $\omega_{11}$ .) In theory, these coefficients are fixed for a machine operating at a fixed value of specific speed, independent of the size of the machine. Equations (71.6)–(71.10) can be used to predict the performance of a large machine using the measured characteristics of a smaller machine or model.



## 71.3 Factors Involved in Selecting a Turbine

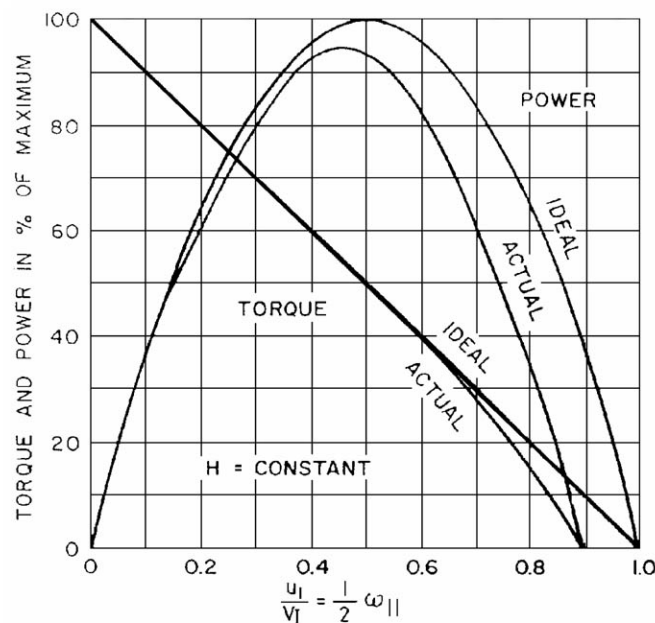
### Performance Characteristics

Impulse and reaction turbines are the two basic types of turbines. They tend to operate at peak efficiency over different ranges of specific speed. This is due to geometric and operational differences.

#### Impulse Turbines

Of the head available at the nozzle inlet, a small portion is lost to friction in the nozzle and to friction on the buckets. The rest is available to drive the wheel. The actual utilization of this head depends on the velocity head of the flow leaving the turbine and the setting above tailwater. Optimum conditions, corresponding to maximum utilization of the head available, dictate that the flow leaves at essentially zero velocity. Under ideal conditions this occurs when the peripheral speed of the wheel is one half the jet velocity. In practice, optimum power occurs at a speed coefficient,  $\omega_{11}$ , somewhat less than 1.0. This is illustrated in Fig. 71.7. Since the maximum efficiency occurs at fixed speed for fixed  $H$ ,  $V_j$  must remain constant under varying flow conditions. Thus the flow rate  $Q$  is regulated with an adjustable nozzle. However, maximum efficiency occurs at slightly lower values of  $\omega_{11}$  under partial power settings. Present nozzle technology is such that the discharge can be regulated over a wide range at high efficiency.

**Figure 71.7** Ideal and actual variable-speed performance for an impulse turbine. (Adapted from Daily, J. W. 1950. *Hydraulic machinery*. In *Engineering Hydraulics*, ed. H. Rouse. Wiley, New York. With permission.)



A given head and penstock configuration establishes the optimum jet velocity and diameter. The size of the wheel determines the speed of the machine. The design specific speed is approximately

$$N_s = 0.77 \frac{d_j}{D} \quad (\text{Pelton turbines}) \quad (71.11)$$



Practical values of  $d_j/D$  for Pelton wheels to ensure good efficiency are in the range 0.04 to 0.1, corresponding to  $N_s$  values in the range 0.03 to 0.08. Higher specific speeds are possible with multiple nozzle designs. The increase is proportional to the square root of the number of nozzles. In considering an impulse unit, one must remember that efficiency is based on net head; the net head for an impulse unit is generally less than the net head for a reaction turbine at the same gross head because of the lack of a draft tube.

## Reaction Turbines

The main difference between impulse units and reaction turbines is that a pressure drop takes place in the rotating passages of the reaction turbine. This implies that the entire flow passage from the turbine inlet to the discharge at the tailwater must be completely filled. A major factor in the overall design of modern reaction turbines is the draft tube. It is usually desirable to reduce the overall equipment and civil construction costs by using high-specific speed runners. Under these circumstances the draft tube is extremely critical from both a flow stability and an efficiency viewpoint. (This should be kept in mind when retrofitting an older, low-specific speed turbine with a new runner of higher capacity.) At higher specific speed a substantial percentage of the available total energy is in the form of kinetic energy leaving the runner. To recover this efficiently, considerable emphasis should be placed on the draft tube design.

The practical specific speed range for reaction turbines is much broader than for impulse wheels. This is due to the wider range of variables that control the basic operation of the turbine. The pivoted guide vanes allow for control of the magnitude and direction of the inlet flow. Because there is a fixed relationship between blade angle, inlet velocity, and peripheral speed for shock-free entry, this requirement cannot be completely satisfied at partial flow without the ability to vary blade angle. This is the distinction between the efficiency of fixed-propeller and Francis types at partial loads and the fully adjustable Kaplan design.

In Eq. (71.4), optimum hydraulic efficiency of the runner would occur when  $\alpha_2$  is equal to  $90^\circ$ . However, the overall efficiency of the turbine is dependent on the optimum performance of the draft tube as well, which occurs with a little swirl in the flow. Thus, the best overall efficiency occurs with  $\alpha_2 \approx 75^\circ$  for high-specific speed turbines.

The determination of optimum specific speed in a reaction turbine is more complicated than for an impulse unit since there are more variables. For a radial-flow machine, an approximate expression is

$$N_s = 1.64 \left[ C_v \sin \alpha_1 \frac{B}{D_1} \right]^{1/2} \omega_{11} \quad (\text{Francis turbines}) \quad (71.12)$$

where  $C_v$  is the fraction of net head that is in the form of inlet velocity head and  $B$  is the height of the inlet flow passage (see Fig. 71.3).  $N_s$  for Francis units is normally found to be in the range 0.3 to 2.5.

Standardized axial-flow machines are available in the smaller size range. These units are made up of standard components, such as shafts and blades. For such cases,

$$N_s \sim \frac{\sqrt{\tan \beta}}{n_B^{3/4}} \quad (\text{Propeller turbines}) \quad (71.13)$$

where  $\beta$  is the blade pitch angle and  $n_B$  is the number of blades. The advantage of controllable pitch is also obvious from this formula, the best specific speed simply being a function of pitch angle.

It should be further noted that  $\omega_{11}$  is approximately constant for Francis units and  $N_s$  is proportional to  $(B/D_1)^{1/2}$ . It can also be shown that velocity component based on the peripheral speed at the throat,  $\omega_{11e}$ , is proportional to  $N_s$ . In the case of axial-flow machinery,  $\omega_{11}$  is also proportional to  $N_s$ . For minimum cost, peripheral speed should be as high as possible—consistent with cavitation-free performance. Under these circumstances  $N_s$  would vary inversely with the square root of head ( $H$  is given in meters):

$$N_s = \frac{C}{\sqrt{H}} \quad (71.14)$$

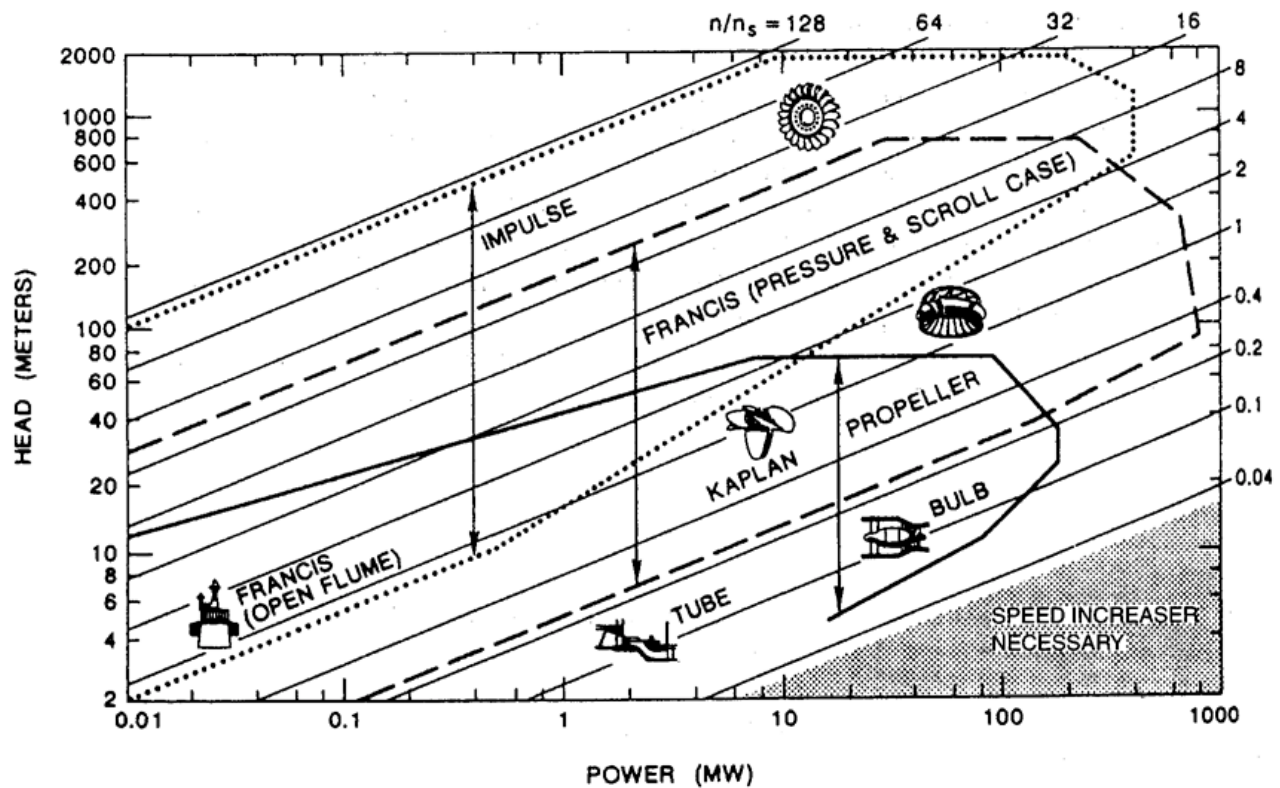
where the range of  $C$  is 21–30 for fixed-propeller units, 21–32 for Kaplan units, and 16–25 for Francis units.

### Performance Comparison

The physical characteristics of various runner configurations are summarized in [Fig. 71.8](#). It is obvious that the configuration changes with speed and head. Impulse turbines are efficient over a relatively narrow range of specific speed, whereas Francis and propeller turbines have a wider useful range. An important consideration is whether or not a turbine is required to operate over a wide range of load. Pelton wheels tend to operate efficiently over a wide range of power loading because of their nozzle design. In the case of reaction machines that have fixed geometry, such as Francis and propeller turbines, efficiency can vary widely with load. However, Kaplan and Deriaz [an adjustable-blade mixed-flow turbine (see [Arndt, 1991](#))] turbines can maintain high efficiency over a wide range of operating conditions. The decision of whether to select a simple configuration with a relatively "peaky" efficiency curve or incur the added expense of installing a more complex machine with a broad efficiency curve will depend on the expected operation of the plant and other economic factors.

**Figure 71.8** Application chart for various turbine types ( $n/n_s$  is the ratio of turbine speed in rpm,  $n$ , to specific speed defined in the metric system,  $n_s = nP^{1/2}/H^{3/4}$ , with  $P$  in kilowatts). (From Arndt, R. E. A. 1991. Hydraulic turbines. In *Hydropower Engineering Handbook*, ed. J. S. Gulliver and R. E. A. Arndt, pp. 4.1–4.67. McGraw-Hill, New York. With permission.)

**Figure 71.8**



Note in Fig. 71.8 that there is an overlap in the range of application of various types of equipment. This means that either type of unit can be designed for good efficiency in this range, but other factors, such as generator speed and cavitation, may dictate the final selection.

## Speed Regulation

The speed regulation of a turbine is an important and complicated problem. The magnitude of the problem varies with size, type of machine and installation, type of electrical load, and whether the plant is tied into an electrical grid. It should also be kept in mind that runaway or no-load speed can be higher than the design speed by factors as high as 2.6. This is an important design consideration for all rotating parts, including the generator.

The speed of a turbine has to be controlled to a value that matches the generator characteristics and the grid frequency:

$$n = \frac{120f}{N_p} \quad (71.15)$$

where  $n$  is turbine speed in rpm,  $f$  is the required grid frequency in Hz, and  $N_p$  is the number of poles in the generator. Typically,  $N_p$  is in multiples of 4. There is a tendency to select higher-speed generators to minimize weight and cost. However, consideration has to be given to speed regulation.

It is beyond the scope of this chapter to discuss the question of speed regulation in detail. Regulation of speed is normally accomplished through flow control. Adequate control requires sufficient rotational inertia of the rotating parts. When load is rejected, power is absorbed, accelerating the flywheel; when load is applied, some additional power is available from deceleration of the flywheel. Response time of the governor must be carefully selected, since rapid closing time can lead to excessive pressures in the penstock.

A Francis turbine is controlled by opening and closing the wicket gates, which vary the flow of water according to the load. The actuator components of a governor are required to overcome the hydraulic and frictional forces and to maintain the wicket gates in fixed position under steady load. For this reason, most governors have hydraulic actuators. On the other hand, impulse turbines are more easily controlled. This is due to the fact that the jet can be deflected or an auxiliary jet can bypass flow from the power-producing jet without changing the flow rate in the penstock. This permits long delay times for adjusting the flow rate to the new power conditions. The spear or needle valve controlling the flow rate can close quite slowly, say, in 30 to 60 seconds, thereby minimizing any pressure rise in the penstock.

Several types of governors are available that vary with the work capacity desired and the degree of sophistication of control. These vary from pure mechanical to mechanical-hydraulic and electrohydraulic. Electrohydraulic units are sophisticated pieces of equipment and would not be suitable for remote regions. The precision of governing necessary will depend on whether the electrical generator is synchronous or asynchronous (induction type). There are advantages to the induction type of generator. It is less complex and therefore less expensive but typically has slightly lower efficiency. Its frequency is controlled by the frequency of the grid it feeds into, thereby eliminating the need for an expensive conventional governor. It cannot operate independently but can only feed into a network and does so with lagging power factor, which may or may not be a disadvantage, depending on the nature of the load. Long transmission lines, for example, have a high capacitance, and, in this case, the lagging power factor may be an advantage.

Speed regulation is a function of the flywheel effect of the rotating components and the inertia of the water column of the system. The start-up time of the rotating system is given by

$$t_s = \frac{I\omega^2}{P} \quad (71.16)$$

where  $I$  = moment of inertia of the generator and turbine,  $\text{kg} \cdot \text{m}^2$  [Bureau of Reclamation, 1966].

The start-up time of the water column is given by

$$t_p = \frac{\sum LV}{gH} \quad (71.17)$$

where  $L$  = the length of water column and  $V$  = the velocity in each component of the water column.

For good speed regulation, it is desirable to keep  $t_s/t_p > 4$ . Lower values can also be used, although special precautions are necessary in the control equipment. It can readily be seen that

higher ratios of  $t_s/t_p$  can be obtained by increasing  $I$  or decreasing  $t_p$ . Increasing  $I$  implies a larger generator, which also results in higher costs. The start-up time of the water column can be reduced by reducing the length of the flow system, by using lower velocities, or by addition of surge tanks, which essentially reduce the effective length of the conduit. A detailed analysis should be made for each installation, since, for a given length, head, and discharge, the flow area must be increased to reduce  $t_p$ , which leads to associated higher construction costs.

## Cavitation and Turbine Setting

Another factor that must be considered prior to equipment selection is the evaluation of the turbine with respect to tailwater elevations. Hydraulic turbines are subject to pitting due to cavitation [Arndt, 1981,1991]. For a given head a smaller, lower-cost, high-speed runner must be set lower (i.e., closer to tailwater or even below tailwater) than a larger, higher-cost, low-speed turbine runner. Also, atmospheric pressure or plant elevation above sea level is a factor, as are tailwater elevation variations and operating requirements. This is a complex subject that can only be accurately resolved by model tests. Every runner design will have different cavitation characteristics. Therefore, the anticipated turbine location or setting with respect to tailwater elevations is an important consideration in turbine selection.

Cavitation is not normally a problem with impulse wheels. However, by the very nature of their operation, cavitation is an important factor in reaction turbine installations. The susceptibility for cavitation to occur is a function of the installation and the turbine design. This can be expressed conveniently in terms of Thoma's sigma, defined as

$$\sigma_T = \frac{H_a - H_v - z}{H} \quad (71.18)$$

where  $H_a$  is the atmospheric pressure head,  $H$  is the vapor pressure head (generally negligible), and  $z$  is the elevation of a turbine reference plane above the tailwater (see Fig. 71.1). Draft tube losses and the exit velocity head have been neglected.

The term  $\sigma_T$  must be above a certain value to avoid cavitation problems. The critical value of  $\sigma_T$  is a function of specific speed [Arndt, 1991]. The Bureau of Reclamation [1966] suggests that cavitation problems can be avoided when

$$\sigma_T > 0.26N_s^{1.64} \quad (71.19)$$

Equation (71.19) does not guarantee total elimination of cavitation, only that cavitation is within acceptable limits. Cavitation can be totally avoided only if the value of  $\sigma_T$  at an installation is much greater than the limiting value given in Eq. (71.19). The value of  $\sigma_T$  for a given installation is known as the plant sigma,  $\sigma_p$ . Equation (71.19) should only be considered a guide in selecting  $\sigma_p$ , which is normally determined by a model test in the manufacturer's laboratory. For a turbine operating under a given head, the only variable controlling  $\sigma_p$  is the turbine setting  $z$ . The required value of  $\sigma_p$  then controls the allowable setting above tailwater:

$$z_{\text{allow}} = H_a - H_v - \sigma_p H \quad (71.20)$$

It must be borne in mind that  $H_a$  varies with elevation. As a rule of thumb,  $H_a$  decreases from the sea-level value of 10.3 m by 1.1 m for every 1000 m above sea level.

## Defining Terms

**Draft tube:** The outlet conduit from a turbine that normally acts as a diffuser. This is normally considered an integral part of the unit.

**Forebay:** The hydraulic structure used to withdraw water from a reservoir or river. This can be positioned a considerable distance upstream from the turbine inlet.

**Head:** The specific energy per unit weight of water. *Gross head* is the difference in water surface elevation between the forebay and tailrace. *Net head* is the difference between *total head* (the sum of velocity head,  $V^2/2g$ , pressure head,  $p/\rho g$ , and elevation head,  $z$ ) at the inlet and outlet of a turbine. Some European texts use specific energy per unit mass, for example, specific kinetic energy is  $V^2/2$ .

**Runner:** The rotating component of a turbine in which energy conversion takes place.

**Specific speed:** A universal number for a given machine design.

**Spiral case:** The inlet to a reaction turbine.

**Surge tank:** A hydraulic structure used to diminish overpressures in high-head facilities due to water hammer resulting from the sudden stoppage of a turbine.

**Wicket gates:** Pivoted, streamlined guide vanes that control the flow of water to the turbine.

## References

- Arndt, R. E. A. 1981. Cavitation in fluid machinery and hydraulic structures. *Ann. Rev. Fluid Mech.* 13:273–328
- Arndt, R. E. A. 1991. Hydraulic turbines. In *Hydropower Engineering Handbook*, ed. J. S. Gulliver and R. E. A. Arndt, pp. 4.1–4.67. McGraw-Hill, New York.
- Bureau of Reclamation. 1966. *Selecting Hydraulic Reaction Turbines*. Engineering Monograph No. 20.
- Daily, J. W. 1950. Hydraulic machinery. In *Engineering Hydraulics*, ed. H. Rouse. Wiley, New York.
- IEC. 1963. *International Code for the Field Acceptance Tests of Hydraulic Turbines*. Publication 41. International Electrotechnical Commission.
- Raabe, J. 1985. *Hydro Power: The Design, Use, and Function of Hydromechanical, Hydraulic, and Electrical Equipment*. VDI Verlag, Dusseldorf, Germany.
- Wahl, T. L. 1994. Draft tube surging times two: The twin vortex problem. *Hydro Rev.* 13(1):60–69.

## Further Information

*J. Fluids Eng.* Published quarterly by the ASME.

*ASME Symposia Proc. on Fluid Machinery and Cavitation*. Published by the Fluids Eng. Div.

*Hydro Rev.* Published eight times per year by HCI Publications, Kansas City, MO.

Moody, L. F. and Zowski, T. 1992. Hydraulic machinery. In *Handbook of Applied Hydraulics*, ed. C. V. Davis and K. E. Sorenson. McGraw-Hill, New York.

*Waterpower and Dam Construction.* Published monthly by Reed Business Publishing, Surrey, UK.

Shibayama, G., Franceschinis, R. D. "Steam Turbines"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



- 72.1 Types of Steam Turbines
- 72.2 Impulse versus Reaction Turbines
- 72.3 Thermodynamics
- 72.4 Stop and Control Valves
- 72.5 Water Induction Protection
- 72.6 Generators
- 72.7 Turbine Generator Auxiliaries

Steam Seals • Turning Gear • Lube Oil

**George Shibayama**

*Doyen & Associates, Inc.*

**Robert D. Franceschinis**

*Doyen & Associates, Inc.*

A **steam turbine** is a rotary device that converts thermal energy to mechanical energy. Steam turbines are primarily used for driving electrical generators or driving mechanical equipment. This chapter will focus primarily on steam turbines used for power generation; however, the basic principals discussed also apply to turbines that are used to drive equipment. The subject of steam turbine design, performance, and operation encompasses a large amount of detailed material that is beyond the intent of this chapter. The reader is encouraged to review the texts given in the "References" and "Further Information" sections for a comprehensive discussion of steam turbine design, performance, and integration of the steam turbine in a modern power cycle.

## 72.1 Types of Steam Turbines

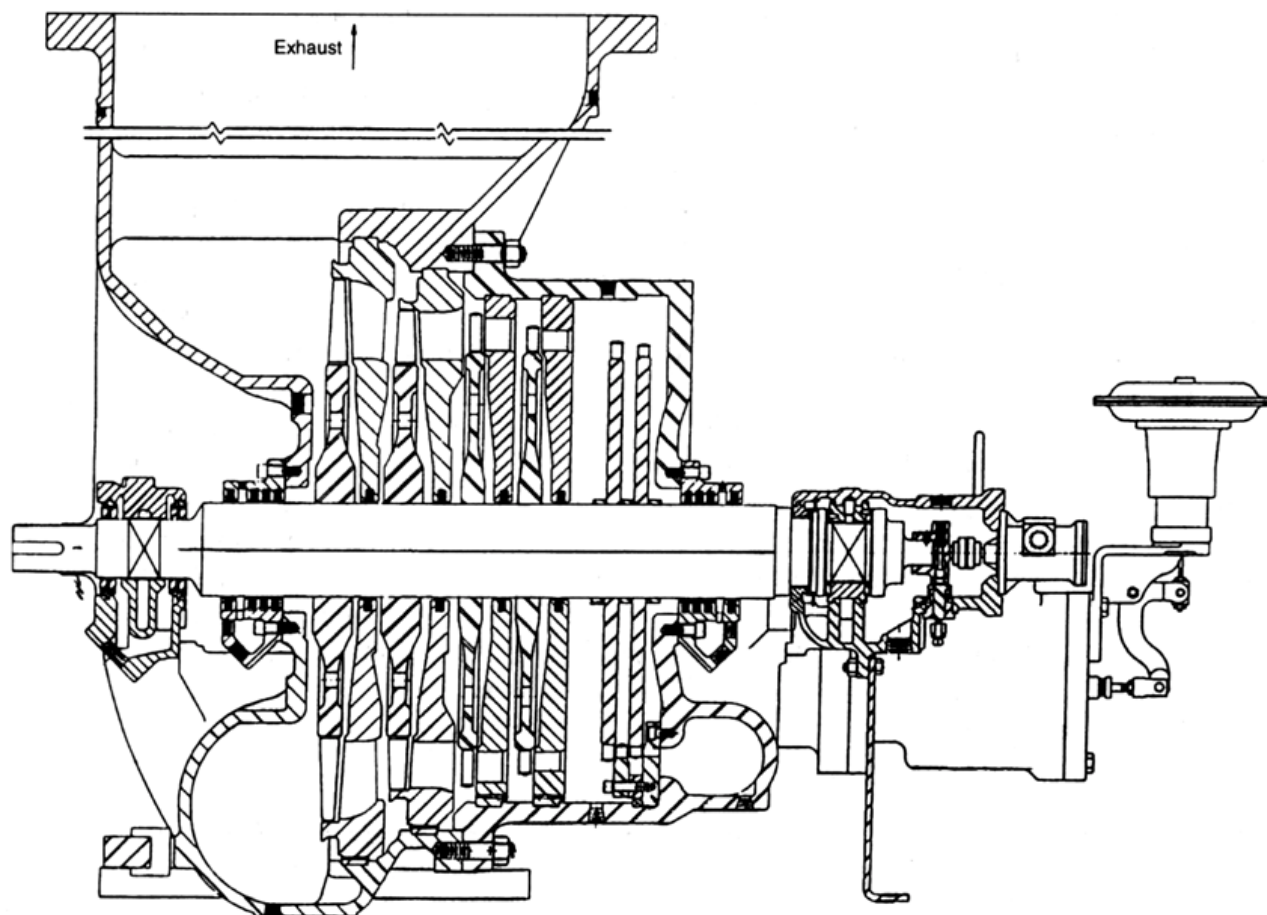
---

Steam turbines are used to either drive an electrical generator for the purposes of producing electric power or for driving mechanical equipment such as pumps, fans, chillers, and so on. Turbines used to generate power are referred to as **turbine generators**. Steam turbines used to drive equipment are referred to as **mechanical drive turbines**. Steam turbine generators range in size from less than 1 MW to as high as 1300 MW. Mechanical drive turbines are usually expressed in terms of horsepower. The more common size range for these machines is 5 hp to 15 000 hp.

A steam turbine consists essentially of several rows of blades or **buckets** arranged in series and connected on a common shaft inside an enclosure called a *casing*. As the steam expands through the turbine, the pressure and temperature decreases as the energy rotates the turbine shaft. The simplest steam turbine has one inlet and one outlet. The steam enters the machine, gives up a portion of its energy to rotate the rotor, and exhausts through the outlet.

A steam turbine that exhausts at a pressure greater than atmospheric pressure is called a **noncondensing turbine**. A steam turbine that exhausts below atmospheric pressure is called a **condensing turbine**. In the latter case the steam turbine must exhaust to a heat exchanger that condenses the steam at a subatmospheric pressure. Noncondensibles in the steam must be removed from the condenser to keep the exhaust pressure and temperature low. High back-pressures not only reduce turbine efficiency, but also can cause heating of the low-pressure section of the turbine. Most turbine generators used for power generation are the condensing type; however, in many industrial facilities, noncondensing back-pressure turbines are used to reduce pressures between processes. [Figure 72.1](#) shows a cross section of a straight condensing turbine.

**Figure 72.1** A cross-sectional view of a straight condensing turbine. (Source: Perry, R. H. and Green, D. W. 1984. *Perry's Chemical Engineers Handbook*. McGraw-Hill, New York. With permission.)



Steam can be extracted or inducted into a steam turbine at single or multiple points between the inlet and outlet. Steam extraction at an intermediate pressure can be used for a process or for the heating of feedwater to the steam boiler to improve the overall steam cycle efficiency.

The extraction of steam can either be controlled or uncontrolled. In an uncontrolled extraction the pressure of the extracted steam will decrease with an increase in extraction flow from the turbine. In a controlled extraction the extraction pressure is kept constant as extraction flow is increased or decreased. Most large utility-type steam turbines use uncontrolled extraction.










Steam can also be inducted back into a steam turbine. This approach is used in combined cycle plants where the exhaust heat from a gas turbine is recovered for the purposes of providing electrical power and steam. Inducting excess steam into the steam turbine improves the efficiency of the plant because some of the waste heat is used to generate lower-pressure steam, which generates more power.

In a large utility reheat power station, steam turbines have high-, intermediate-, and low-pressure casings or sections. As the names imply, the high-pressure steam enters the hp section; gives up a portion of its energy; becomes reheated in the boiler; enters the IP section, where more of the energy is given up; and finally enters the LP section, where some additional power is obtained prior to entering the condenser.

Steam turbines can also be classified by their configuration. A steam turbine generator for electrical power generation can be classified as a tandem compound or cross compound unit. A tandem compound unit operates with all of the steam turbines rotating on a common shaft connected to one generator. A cross compound unit has the steam turbines on two separate shafts. In this case the HP and IP turbines are on a single shaft and connected to a single generator, whereas the LP turbine is on a separate shaft connected to a separate generator. Virtually all new steam turbine installations are the tandem compound type.

As shown in [Table 72.1](#), standard designations have been adopted to indicate the types of turbines. A TC2F-33.5 is a tandem compound turbine with two exhaust flows to the condenser, and 33.5 inch last-stage blades. Similarly, the designation for a four-flow cross compound machine with 38" blades would be CC4F-38. The number of exhaust flows to the condenser is a function of the number of low-pressure turbine sections. Utility turbine generators can have two, four, or six exhausts to the condenser.

**Table 72.1** Typical Turbine-Generator Configurations

Fossil	Fossil	Nuclear
<p>TC-2F LSB 26, 30 and 33.5 in Two casings 3600 r/min</p>  <p>125-400 MW</p>	<p>TC-6F LSB 26, 30 and 33.5 in Five casings 3600 r/min</p>  <p>550-1000 MW</p>	<p>TC-4F LSB 38 and 43 in Three casings 1800 r/min</p>  <p>450-1000 MW</p>
<p>TC-4F LSB 26, 30 and 33.5 in Three casings 3600 r/min</p>  <p>250-650 MW</p>	<p>TC-6F LSB 30 and 33.5 in Five casings 3600 r/min (double reheat)</p>  <p>450-725 MW</p>	<p>TC-6F LSB 38 and 43 in Four casings 1800 r/min</p>  <p>600-1100 MW</p>
<p>TC-4F LSB 26, 30 and 33.5 in Four casings 3600 r/min</p>  <p>550-850 MW LP LP G</p>	<p>CC-4F LSB 38 and 43 in Four casings 3600/1800 r/min</p>  <p>3600 r/min</p>  <p>1800 r/min</p> <p>600-1250 MW</p>	

Data provided by the General Electric Company. TC = tandem compound. CC = cross compound. F = number of flow ducts to condenser. LSB = last-stage blade.

Source: El-Wakil, E. M. 1984. *Power Plant Technology*. McGraw-Hill, New York. With permission.

Most steam turbines for nonnuclear power generation rotate at 3600 rpm. Nuclear turbines operate at 1800 rpm since the blade rows are longer and a lower speed is necessary to reduce blade stresses. A common misconception is that the electrical power output of a steam turbine generator is related to speed. Steam turbine generators operate at a fixed speed. The power output from a steam turbine is controlled by varying the control valve position on the steam chest, thereby admitting more or less steam to the turbine. The increased torque on the steam turbine generator shaft resulting from the increased steam flow produces the increase in power output.

Mechanical drive turbines generally operate over a variable speed range. These turbines can be very useful in driving equipment that frequently operates at lower loads. As the load on a pump or fan is reduced, the turbine can slow down and reduce the capacity of the pump or fan. As the demand increases, the turbine increases in speed as required to produce the flow. This infinite type

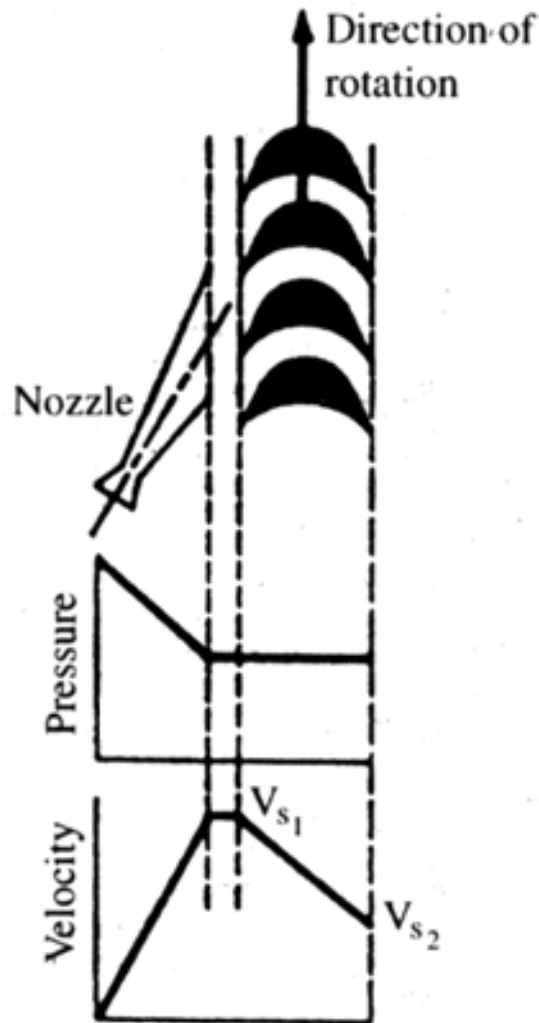
of speed adjustment can result in a large power savings compared to the constant speed operation of driven equipment.

## 72.2 Impulse versus Reaction Turbines

There are two fundamental types of steam turbines from the perspective of how energy is transferred from the steam to the turbine shaft. As discussed previously, energy is transferred by the steam expanding through a series of stationary and rotating blade sections. These blade sections can be of the **impulse** and/or **reaction** types.

In an impulse turbine the steam is directed through nozzles to impact the buckets or blades attached to the rotor shaft. The energy to rotate the turbine comes from the force of the steam impacting on the buckets. A commonly used analogue is a child's pinwheel turning in the wind. [Figure 72.2](#) illustrates the principle of impulse blading.

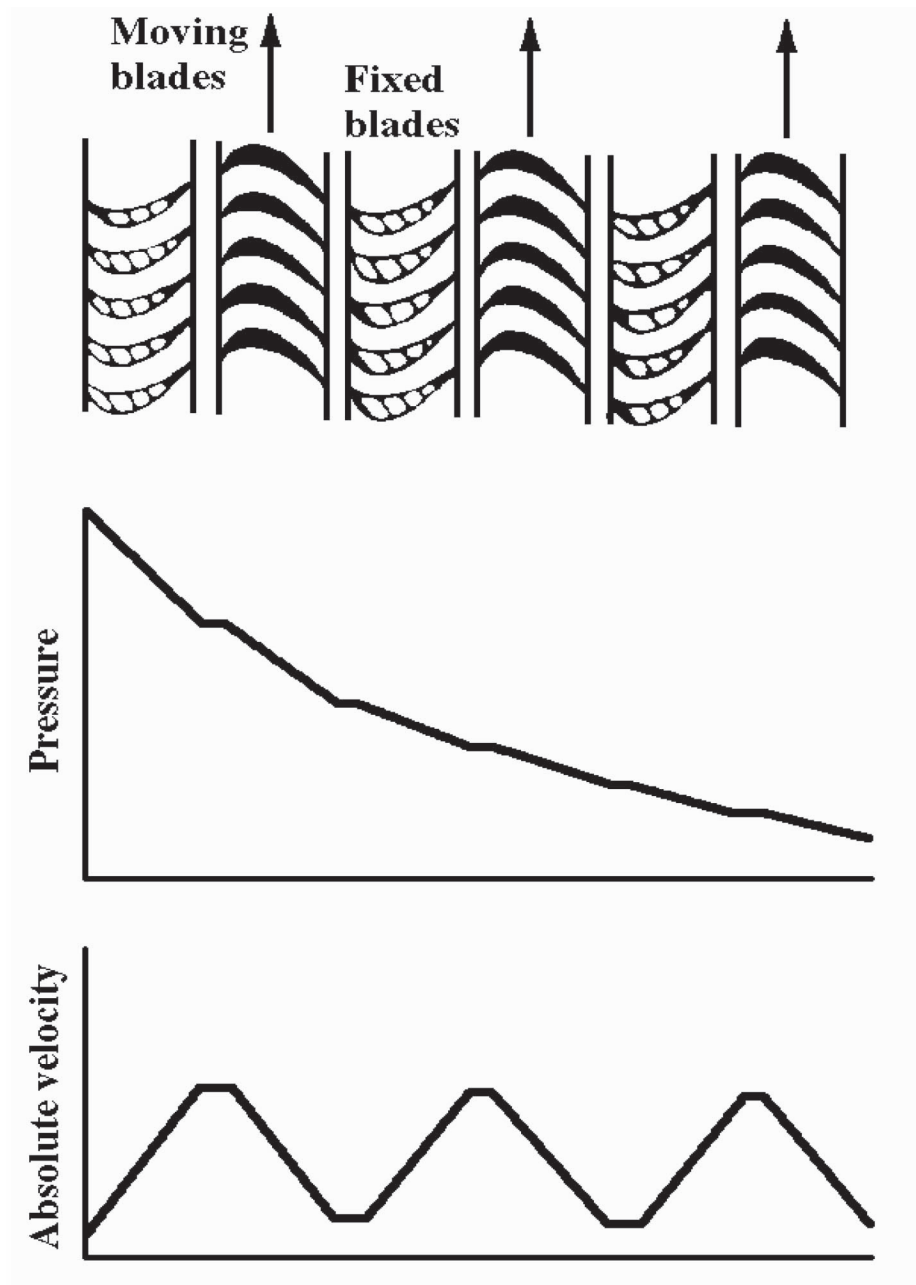
**Figure 72.2** An ideal single-stage impulse turbine. (Source: El-Wakil, E. M. 1984. *Power Plant Technology*. McGraw-Hill, New York. With permission.)



In a reaction turbine the stationary nozzles and rotating blades are the same shape. The steam

expands, increases in velocity, and loses pressure as it passes through the blade sections. The resulting force generated by this velocity turns the rotor. A common analogue is releasing an inflated toy balloon. The balloon speeds away due to the reactive force. [Figure 72.3](#) illustrates the principle of reaction blading.

**Figure 72.3** A three-stage reaction turbine. (Source: El-Wakil, E. M. 1984. *Power Plant Technology*. McGraw-Hill, New York. With permission.)



Some manufacturers use impulse blading, whereas others use a combination of both reaction and impulse blading. Both types of blading accomplish the same purpose of converting the thermal energy of the steam into useful mechanical work.

## 72.3 Thermodynamics

---

The amount of power that can be generated by a steam turbine is a function of several variables. The initial steam pressure and temperature, steam flow, exhaust pressure, and efficiency of the machine all determine how much power can be generated. Additionally, mechanical bearing and electrical generator losses in the turbine generator need to be considered, since these losses reduce the power output.

The efficiency of a steam turbine is defined as the actual work produced divided by the work produced by an isentropic expansion. An isentropic expansion is the amount of work that would be produced if no change in **entropy** occurred. An isentropic process is an idealized process that represents the amount of available energy. The second law of thermodynamics, however, states that the conversion of this thermal energy to useful work cannot be 100% efficient. In practical use, it will be less than 100% because of the second law and because of additional mechanical and electrical losses in the turbine generator itself.

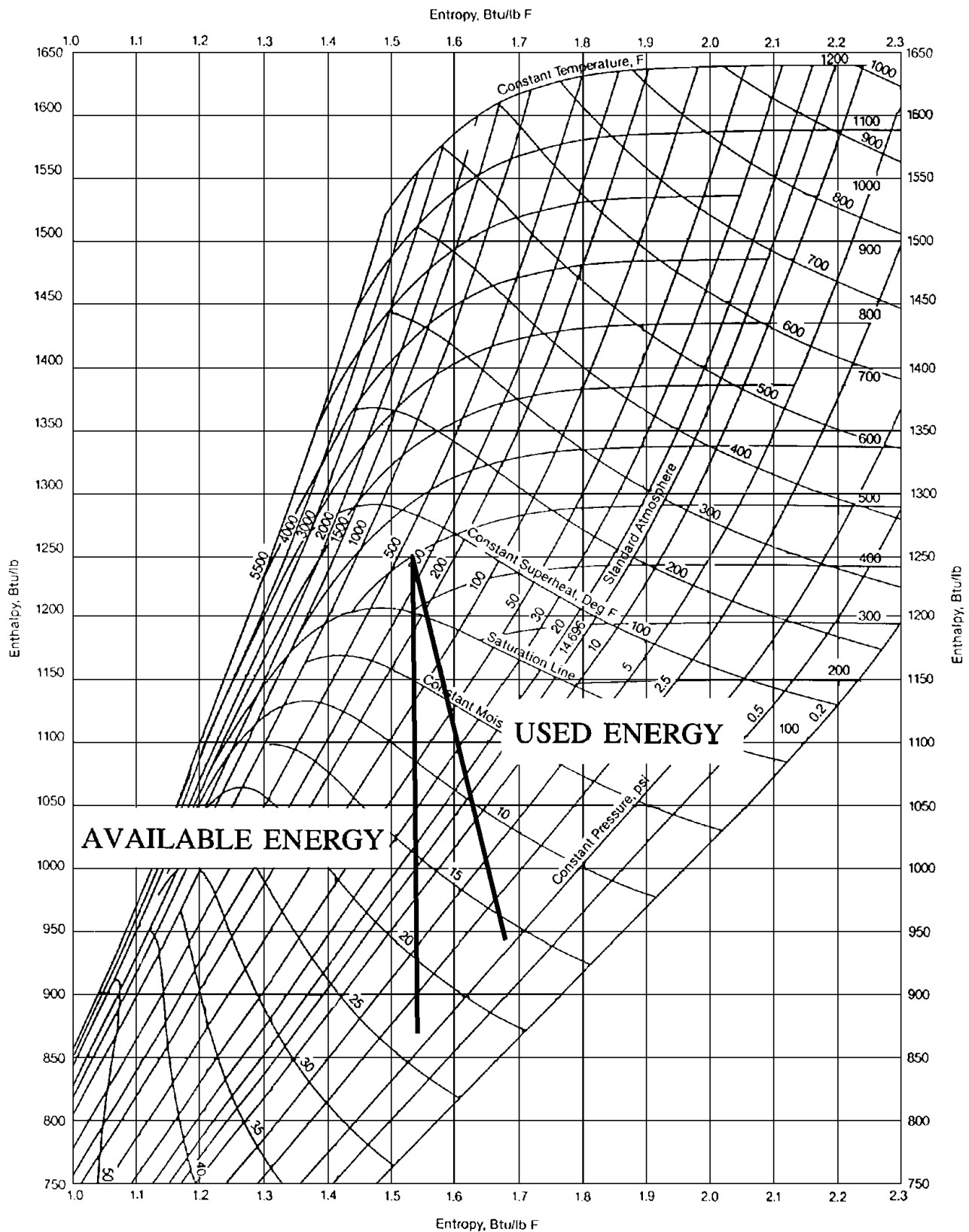
Expressed mathematically, the efficiency of a steam turbine is as follows:

$$\begin{aligned}\text{Efficiency} &= \text{Actual work} / \text{Work from isentropic expansion} \\ &= \text{Used energy} / \text{Available energy}\end{aligned}\tag{72.1}$$

Figure 72.4 shows a Mollier chart in which the expansion of a single-stage turbine is illustrated. The efficiency is a function of the machine type, size, manufacturer, and the steam inlet and exhaust conditions.



**Figure 72.4** A Mollier chart showing a turbine expansion.





The first law of thermodynamics, written for a steam turbine, states that the power output will be the mass flow rate of steam through the turbine multiplied by the difference in **enthalpy** across the turbine, with the result converted to kilowatts or horsepower. The inlet enthalpy is known from the steam conditions, whereas the exhaust enthalpy is a function of the efficiency of the expansion. Expressed mathematically, the power output for a simple nonextraction steam turbine would be as follows:

$$\text{Power output (kW)} = \text{Steam flow (lb/h)} \times (H_{\text{in}} - H_{\text{out}})/3413 \quad (72.2)$$

where  $H_{\text{in}}$  = enthalpy of steam at turbine inlet and  $H_{\text{out}}$  = enthalpy of steam at turbine outlet. The calculation of power output for a steam turbine can be illustrated by a simple example.

**Example.** An industrial, condensing, nonextraction steam turbine with a throttle pressure of 400 psig and 500 F and a steam flow of 100 000 lb/h exhausts to a condenser at 3.0" HgA. The manufacturer states that the efficiency of this machine is 80.0%. Determine the power output from this machine.

**Solution.** From steam tables:

$$\text{Enthalpy of steam at 400 psig / 500 F} = 1243.2 \text{ Btu/lb} \quad (\text{H1})$$

$$\text{Entropy of steam at 400 psig / 500 F} = 1.5225 \text{ Btu/lb} \quad (\text{S1})$$

$$\text{Enthalpy of steam at S1 and 3.0" HgA} = 868.26 \text{ Btu/lb} \quad (\text{H2S})$$

$$\text{Exhaust enthalpy} = [(H2S - H1) \times 0.80] + H1 = 943.228 \text{ Btu/lb} \quad (\text{H2})$$

$$\text{Power output} = (100\,000 \text{ lb/hr} \times (1243.2 - 943.23))/3413 = 8789 \text{ kW}$$

where 3413 is in Btu/kW-h

In this example, if testing of this turbine were contemplated, the enthalpy of the supply steam could be determined from pressure and temperature measurements. The exhaust enthalpy, however, would not be known. The determination of enthalpy in a two-phase region (where steam and moisture coexist) by pressure and temperature measurements will only yield estimates of the actual enthalpy since the actual moisture in the steam is not known. The exhaust enthalpy, however, could be estimated by back-calculating, since the power output would be measured during testing.

The above simplified example is based on a single-stage turbine with no steam extraction or induction. All of the steam entering the turbine leaves the turbine exhaust and solely produces power. In practice, steam can be extracted from or inducted to steam turbines for the purposes of exporting steam to a process or improving cycle efficiency. [Figure 72.5](#) shows a typical power cycle with four extraction points. The extraction or induction of steam will affect the power output and efficiency of the cycle and must be accounted for in the calculations.

The diagram illustrates a steam power plant cycle. The main components and their connections are as follows:

- BLR (Boiler):** The primary heat source, connected to the main steam line.
- Process Unit:** A small square box labeled "PROCESS" with an input "s" and an output "PROCESS". It is connected to the main steam line via a valve.
- TURBINE:** The central component where steam expands to produce work. It is connected to the main steam line and the condenser.
- COND (Condenser):** The component where steam is condensed. It is connected to the turbine and the HWP.
- GEN (Generator):** Connected to the turbine to produce electricity.
- HWP (High Pressure Pump):** A pump that circulates water from the condenser back to the boiler.
- GSLO CONDENSER:** A condenser that receives steam from the turbine and is connected to the HWP.
- Heating/Boiling Points:** Three points are marked with dots and numbers:  $\bullet 1$  (HDP - High Pressure Point),  $\bullet 2$  (BFP - Boiling Point), and  $\bullet 3$  (HDP - High Pressure Point).
- Flow Control:** The system includes several valves (represented by the 'X' symbol) and pumps (represented by the circle with an arrow) to manage the flow of steam and water.

Frequently, turbine performance is expressed by another term, *turbine heat rate*. The heat rate is defined as the number of Btus that must be added to the working fluid (the steam) to generate one kW-h of electrical power (Btu/kW-h). The heat rate is not only a function of the process conditions but also the type of cycle. In a power station the heat rate is affected by the number of feedwater heaters in the cycle, in addition to steam pressure, temperature, exhaust pressure, steam flow, the type of unit (reheat versus nonreheat), control valve position, and cycle losses.

Heat input to working fluid /kW output (72.3)

$$\text{Heat rate} = 3413/(\text{efficiency}) \quad (72.4)$$

where efficiency is expressed as a decimal.

Two important but frequently misunderstood parameters are the **expansion line end point (ELEP)** and the **used energy end point (UEEP)**. The expansion line end point represents the turbine exhaust enthalpy that would exist if there were no exhaust loss at the turbine exit. However, in practice, the expansion of the steam from the low-pressure turbine into the condenser results in a loss that is a function of the velocity through the turbine exhaust opening. In the calculation of power output the UEEP should be used (not the ELEP) because the exhaust loss does not contribute to the generation of power.

In the above example, if the exhaust enthalpy could be directly measured, this enthalpy would be the UEEP and not the ELEP. Expressed mathematically, the exhaust loss is defined as follows:

$$\text{Exhaust loss} = \text{UEEP} - \text{ELEP} \text{ (Btu/lb)} \quad (72.5)$$

For many complicated power cycles, a heat balance program is utilized to perform the mass and energy balances needed to calculate the performance. Simple cycles can be done manually; with the complexity of power cycles and the availability of personal computers, however, the many number of iterations required to balance the cycle can be performed more accurately and faster with a heat balance program.

## 72.4 Stop and Control Valves

---

Stop valves are provided on the turbine to admit steam during normal operation or to shut off the steam very quickly in the event of an emergency. These valves are normally fully opened or closed. The control or governing valves are located in the steam chest at the turbine inlet and control the flow of steam to the turbine. On larger units, multiple control valves are used to provide better efficiency over the load range. The turbine control valves can be operated in either a full arc admission or partial arc admission mode. In full arc admission, all of the control valves are opened simultaneously, with the stop valve used to control flow. In this mode of operation, more even heating of the turbine rotor and casing is possible. In partial arc admission, one control valve is opened at a time (sequentially). The turbine is not as evenly heated; but the efficiency of the turbine is better, since valve-throttling losses are reduced.

## 72.5 Water Induction Protection

---

The steam turbine is designed to generate power by expanding low-density steam through a series of nozzles and blade rows. It is possible, however, to induct water into the turbine from cold piping on a unit start-up, an extraction line, the boiler, or a desuperheater. **Water induction** incidents are more likely to occur on unit trips or load changes. The dense water, in comparison to the steam, can cause a considerable amount of damage to the blades and buckets on impact, and the cold water in contact with hot turbine metal can cause cracking of metal or rubbing between moving and stationary parts due to differential expansion.

The ASME has developed a standard that should be followed when designing systems associated with a steam turbine. ASME TDP-1 provides design details on how to prevent water induction. Although the standard applies mainly to steam turbine generators, the guidelines are also applicable to mechanical drive turbines. Regardless of the size of the steam turbine, support systems should be designed to prevent water induction, since the damage that can result can be considerable.

## 72.6 Generators

---

The generator is a device that converts the mechanical work of the turbine into electrical power. This conversion is accomplished by inducing an electromotive force by creating relative motion between a conductor and a magnetic field. The stationary part of a generator is called a *stator*. The moving part is called a *rotor*. In order to create the magnetic field, an exciter provides the DC electricity to magnetize the rotor. The frequency of the power generated is equal to the speed of the rotor in revolutions per unit time. A two-pole synchronous generator for a nonnuclear-type power plant must revolve at 3600 rpm to generate a 60 Hz voltage, as the following equation illustrates:

$$f = (P/2) \times (N/60) \quad (72.6)$$

where  $f$  = frequency in Hz,  $P$  = the number of magnetic poles, and  $N$  = the speed of the rotor in rpm. A nuclear turbine operating at 1800 rpm would require a four-pole generator to generate 60 Hz.

Generators must be cooled to remove the heat produced by the windings. Hydrogen gas is usually used because it has a low density and high specific heat compared to air. Hydrogen, however, is extremely flammable; precautions during filling and venting of the generator must be followed. Hydrogen mixtures are explosive when the concentration of oxygen exceeds 5.0%. Carbon dioxide is usually used as an inert gas during filling and purging of the generator to prevent the hydrogen from becoming an explosive mixture. Some smaller turbine generators use air instead of hydrogen for cooling.

## 72.7 Turbine Generator Auxiliaries

---

Many auxiliary systems that are required for operation are provided with a steam turbine. Some of the major auxiliary systems are briefly discussed below.

### Steam Seals

At locations where the steam turbine shaft penetrates the casing(s) of the turbine, a steam seal system is used to prevent steam from leaking out of the seals which are above atmospheric pressure, and air leaking into the seals which are below atmospheric pressure. A steam seal system uses steam leak-off from the high and intermediate seals during operation to seal the low-pressure seals. When the turbine is on-line, the machine is said to be "self-sealing." When the turbine is

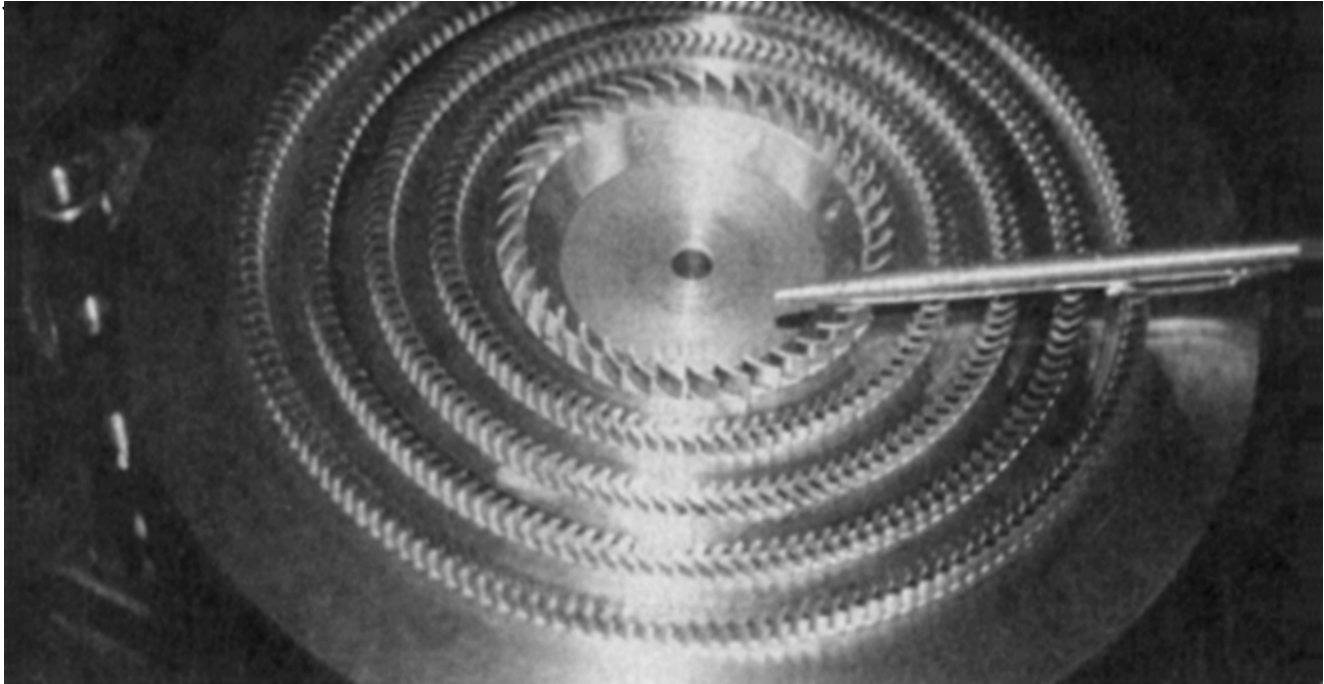
**Figure 72.6** A typical steam seal flow diagram.



## Lube Oil

© 1998 by CRC PRESS LLC

and oil coolers, which in most cases are mounted in an oil reservoir. The shaft-driven pump is used when the turbine is at 90% of rated speed or above. Below this speed, the AC or DC motor-driven pumps are used. The high-pressure control oil system, which operates the control and stop valves and turbine governor, is supplied from the lube oil system.



This turbine was developed at the University of Pennsylvania in the late 1970s under USDOE sponsorship for the purpose of producing power efficiently from low-temperature energy sources, such as solar energy at 100° C or even lower. It was employed in a novel hybrid Rankine cycle, in which steam was generated by solar energy at the  $\approx 100^\circ\text{C}$  temperature level (80% of the total energy input), and heat was added (the remaining 20%) by a higher temperature source (such as gas or solar concentrators) to superheat the steam to a top temperature of 550° C. The remarkable feature of this power cycle was that this addition of 20% high-temperature heat doubles the cycle efficiency, from about 9% for operation at the 100° C level to about 18% for operation as a hybrid.

The turbine itself is based on the well-known counterrotating (Ljungström de-Laval) principle, having two rotors rotating in opposite directions, at 15 300 rpm. Its uniqueness lies in its small size (30 hp) and high efficiency (75%) for that size range. This was an experimental prototype. Larger turbines of this type are used in both stationary and marine propulsion applications. (Photo courtesy of Noam Lior, University of Pennsylvania.)

## Defining Terms

**Buckets:** Turbine blades.

**Condensing turbine:** Any turbine with an exhaust below atmospheric pressure.

**ELEP:** Expansion line end point; enthalpy of exhaust steam if the exhaust loss at the turbine exit were neglected.

**Enthalpy:** A measure of the stored energy of a substance, expressed in Btu/lb.

**Entropy:** A measure of the ability of a substance to provide useful work from energy, expressed in Btu/lb-°F.

**Feedwater heater:** A heat exchanger that uses extraction steam to heat feedwater and condensate in a power station.

**Impulse turbine:** A turbine that uses the force of the steam impacting on the blades or buckets to rotate the turbine.

**Mechanical drive turbine:** A steam turbine connected to a pump, fan, or a similar device to provide motive power.

**Noncondensing or back-pressure turbine:** Any turbine with an exhaust above atmospheric pressure.

**Reaction turbine:** A turbine that uses the force generated by the velocity in the stages to rotate the turbine.

**Steam turbine:** A rotary engine that converts thermal energy to useful mechanical work by the impulse or reaction of steam.

**Turbine generator:** A steam turbine connected to an electrical generator to produce power.

**UEEP:** Used energy end point; enthalpy of steam at the turbine exhaust.

**Water induction:** The admission of water into a steam turbine.

## References

El-Wakil, E. M. 1984. *Power Plant Technology*. McGraw-Hill, New York.

Perry, R. H. and Green, D. W. 1984. *Perry's Chemical Engineers Handbook*. McGraw-Hill, New York.

## Further Information

The following texts are useful for their in-depth discussions on steam turbine fundamentals.

Cotton, K. C. 1993. *Evaluating And Improving Steam Turbine Performance*. Cotton Fact, Rexford, NY.

Elliott, T. C. 1989. *Standard Handbook of Powerplant Engineering*. McGraw-Hill, New York.

Potter, P. J. 1959. *Power Plant Theory And Design*. John Wiley & Sons, New York.

Salisbury, J. K. 1974. *Steam Turbines and Their Cycles*. Kreiger, Huntington, NY.

The following reference contains the guidelines recommended by the major turbine manufacturers for water induction prevention.

ASME. 1985. *Recommended Practices for the Prevention of Water Damage to Steam Turbines Used for Electric Power Generation*. ASME TDP-1-1985. American Society of Mechanical Engineers, New York.

Cooke, D. H. "Cogeneration"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



### 73.1 Cogeneration Fundamentals

Units for Power and Heat Flow • Heat Rate and Efficiency for Noncogenerative Cycles • Cogeneration Performance Criteria

### 73.2 Examples of Cogeneration

Thermal Sequence • Plants for "Difficult" Fuels • Large-Scale Power and Cogeneration

## David H. Cooke

*Power and Cogeneration Consultant*

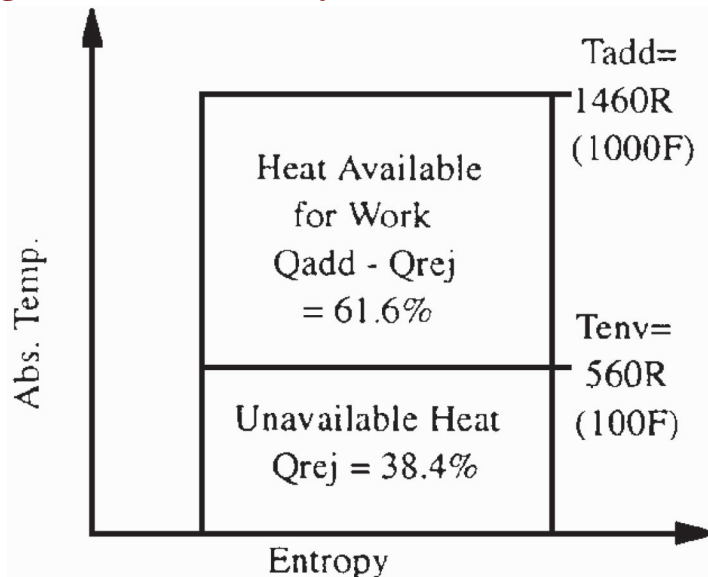
Cogeneration—or the production, from the same fuel or energy source, of electric and/or shaft *power* sequentially together with process *heat*<sup>3/4</sup> has been used in industry as an economical means of providing both forms of energy since the evolution of electric power late in the 19th century. The term *sequential*, a necessary concept in the definition according to many sources [Butler, 1984], means *thermally* sequential, as upper, then lower segments of the temperature range of energy transfer are converted to power and/or useful heat. For reasons to be made clear later, the sequential production yields significant fuel savings relative to separate production facilities for heat and power.

Prior to the oil-related world energy crisis of the 1970s, cogeneration applications were mostly limited to local, individual, "in plant" heat and power demands of industries where both needs were readily served. Although industries such as paper mills and refineries have supplied some public power since the earliest times, cogenerated power had dwindled to about 4% of the nation's demand by 1977 [Butler, 1984] because of the ever-increasing demand for electric power, lower fuel costs, and improved power generation technology. However, with the recent emphasis on energy conservation, large-scale cooperative cogeneration has been recognized as a major source of fuel savings throughout the world. If the power industry could economically cooperate on a large scale with process heat users, it could be highly beneficial to the reduction of national reliance on imported energy sources as well as the environment. The passage of the National Energy Act of 1978 and subsequent laws opened the path to greater utilization in the U.S. by promoting and making it more economical for utilities and independent power producers, including industrial heat users, to produce and sell cogenerated power for the utility grids. Progress has been slow because of economic impediments such as (1) the concentration of industry heat demand in certain regions so that cogeneration with all available process heat would result in a local excess of power, and (2) the lack of transmission lines to **wheel** cogenerated power to remote regions where it is needed. However, recent technological advances, particularly in the gas turbine, have greatly enhanced the economic potential of cogeneration, and the prognosis for continued gradual development is good.

## 73.1 Cogeneration Fundamentals

In the production of electric or shaft power, the Second Law of thermodynamics exerts a fundamental limitation in the conversion of heat into work or power. That is, only a fraction of the thermal energy released by combustion of fuel can be converted to power, depending on (1) the ratio of the absolute temperature of the environment to that of the heat supply to the working fluid in the boiler or burner, and (2) the extent to which friction and other forms of irreversibility are present in the cycle. Consider a Carnot cycle, Fig. 73.1, which is the most efficient possible means of converting heat into work. For this ideal reversible cycle, with typical "earthbound" temperatures, a fraction of the total heat added, equivalent to  $T_{\text{env}}/T_{\text{add}}$ , must be rejected—because of the impossibility of heat flowing thermally "uphill" of its own accord—giving an ideal **thermal efficiency** of only about 62%. The same principle applies to practical, well-designed, modern power systems; but with their inherent "nonrectangular" cycle characteristics and irreversibilities, in addition to the Second Law limitation, typically only 35 to 50% of the heat energy supplied can be converted to power, depending on the type and temperatures of these real cycles. The rest of the energy must remain in the form of heat and, without cogeneration, be rejected to the environment, as, for example, by condensing the exhaust steam from a power turbine with river cooling water. This, of course, results in dissipative heating of the river, which may also be an environmental detriment.

**Figure 73.1** Carnot heat rejection.

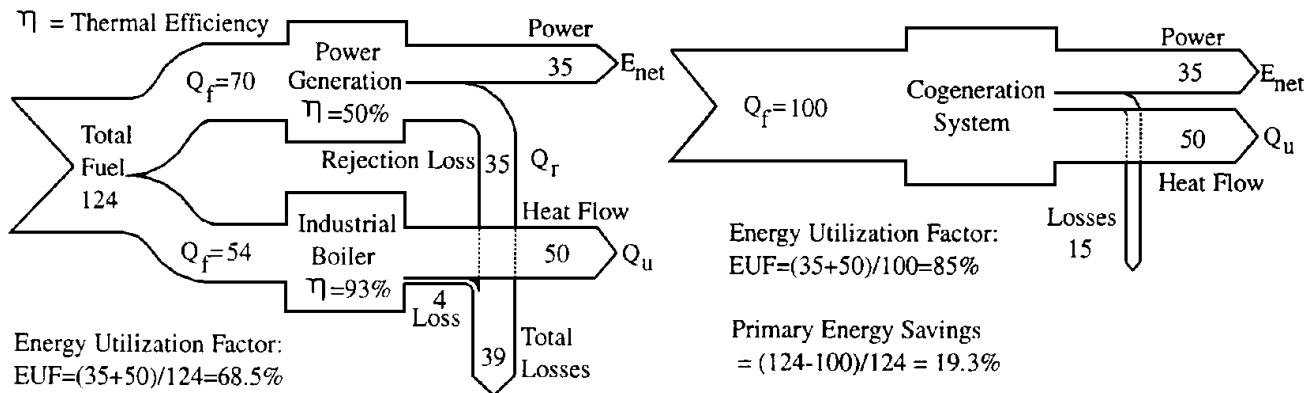


Such a thermal limitation does not exist in production of heat for process use where no conversion to work is involved. Except for relatively small leakages, all of the heat energy supplied to the transfer medium is capable of being utilized. This suggests that, if power and process heat could be produced together, some or all of the heat that otherwise would be wasted by heat rejection could be used, thus saving energy. This ideal is what actually occurs in cogeneration. During an industrial process, heat energy may be consumed by fluid heating, endothermic chemical reactions, or other needs, but any of the energy remaining in the form of low-level heat has potential value because it may ultimately reach the environment by, say, heating a building rather than being wasted.

In some cases, such as with industrial furnaces and kilns, the process need is for *high-level* (i.e., high-temperature) heat. Here, also, where effective uses of such large quantities of "leftover" heat are not available, cogeneration allows the option of converting heat energy that would otherwise be wasted to power. As will be explained later, when the power conversion is thermally below the heat utilization, it is called a "bottoming" cycle.

Figure 73.2 shows the energy savings with cogeneration by comparison of separate versus cogenerative plants accomplishing the same power generation and process heat duty. The typical power generation system is the same in either case—except, with cogeneration, energy that would have been rejected is used to supply the process heat.

**Figure 73.2** Energy savings with cogeneration. (After Boissenin, Y. 1992. *Cogeneration and energy savings. GEC Alstom Technical Review No. 10.* European Gas Turbines S.A., Belfort, France.)



## Units for Power and Heat Flow

The numbers shown on Fig. 73.2 for both power and heat flow are intended to be in megawatts, which would make the figure comparable in size to actual plants of this type. The classic units for heat (i.e., Btu, joules, calories, etc.) vary with the unit convention being used, but the use of watts as a basic unit for electric energy is common to most conventions. Further, the use of watts as the common unit for both heat flow and power is frequently done for simplicity. Accordingly, a thermal kilowatt ( $kW_t$ ) is understood to be 3412.14 Btu/h, 3600 kJ/h, or 859.9 kcal/h, according to the electrical equivalent of heat as expressed in three commonly used unit conventions. As shown in the Fig. 73.2 diagram,  $E$  will henceforth refer to electrical or shaft power, in  $kW_e$  or  $MW_e$ . For heat flow,  $Q_f$  will refer to heat release rate of fuel,  $Q_u$  to useful heat flow to process, and  $Q_r$  to heat flow wasted by rejection, in  $kW_t$  or  $MW_t$  unless otherwise noted.

The values for fuel heat release and efficiencies in Fig. 73.2 and elsewhere in this chapter are based on the **lower heating value (LHV)** of the fuel. LHV excludes the latent heat of water vapor formed from combustion of hydrogen in fuel compounds. It is excluded because this energy is almost always lost with the vapor in the flue gas at temperatures above 150°F, and using LHV makes comparative cycle analysis easier (independent of the fuel). Nevertheless, only *higher heating value (HHV)* can be directly measured and, therefore, is always the basis of fuel cost, so that in economic evaluations HHV must be determined. It is a good practice to show both LHV and HHV for fuel heat release and efficiencies where economics are involved, particularly if different fuels must be compared. This is a simple matter of multiplying LHV heat release and **heat rate**, or dividing thermal efficiency, by a constant (e.g., typically 1.11 for natural gas or 1.06

for diesel oil). For fuels without hydrogen (e.g., pure carbon or coke) the constant is 1.0 and LHV and HHV are identical.

## Heat Rate and Efficiency for Noncogenerative Cycles

It is important to recall relations for conventional power cycles so that corresponding parameters for cogeneration may be more easily explained. Heat rate is a standard parameter in both the power and industrial fields for sizing and rating power plants and is always expressed in classic heat units. It is closely related to thermal efficiency,  $\eta$ , which is a dimensionless ratio that can be expressed in percent. Considering the separate power generation system on the upper left in Fig. 73.2, we see that the fuel heat release,  $Q_f$ , to the system is 70 MW<sub>t</sub> (LHV). The net power,  $E_{\text{net}}$ , is 35 MW<sub>e</sub>. (The term *net* here refers to the power that is actually delivered to the utility grid, after subtracting cycle **auxiliary power** and any other electrical losses from the output at the generator terminals quoted by the equipment manufacturer.) With the foregoing established, heat rate and efficiency for the conventional power generation example on the upper left in Fig. 73.2 then become, with Btu/h units and HHV for natural gas:

$$\text{Net heat rate (LHV)} = \frac{Q_f \times 3412.14}{E_{\text{net}}} = \frac{238.85 \cdot 10^6}{35\,000} = 6824 \text{ Btu/kWh}$$

$$\text{Net heat rate (HHV)} = 1.11 \times 6824 = 7575 \text{ Btu/kWh}$$

$$\text{Thermal efficiency (LHV), } \eta = \frac{E_{\text{net}}}{Q_f} = \frac{35\,000}{70\,000} = 50\%$$

$$\eta = \frac{3412.14}{\text{Net heat rate}} = \frac{3412.14}{6824} = 50\%$$

$$\text{Thermal efficiency (HHV)} = \frac{50}{1.11} = 45\%$$

Note that heat rate is simply the reciprocal of thermal efficiency with unit conversion.

## Cogeneration Performance Criteria

The criteria of efficient performance of a cogeneration plant are quite different from that of a conventional power plant. In the latter, the object is usually to meet a demand for a single product, power, with the least amount of fuel. In a cogeneration plant there are *two* products, power and heat, which together require more fuel energy than either would require if produced separately (although less fuel energy than both would so require). Because the products and their required fuel are now intertwined, the heat rate and thermal efficiency definitions given above are no longer adequate. New definitions are required to account for the two products.

## Energy to Process

Figure 73.3 shows schematically an industrial boiler providing heat to a process, without power production, as in the lower left side of Fig. 73.2. The heat is provided in the form of steam, which is returned as condensate to a system reservoir, usually the deaerator, where dissolved oxygen is removed by a relatively small amount of deaerating (d/a) steam. The **net heat to process**,  $Q_u$ , is defined strictly based on the energy entering and leaving the process, except that frequently some or all of the steam delivered to the process does not return as condensate, and makeup water must be provided to keep up the system inventory. As shown in Fig. 73.3,  $Q_u$  is defined as the difference in total enthalpy, here in Btu/h, between that of the steam delivered to process and that of the sum of condensate returned plus makeup required.

**Figure 73.3** Net heat to process.

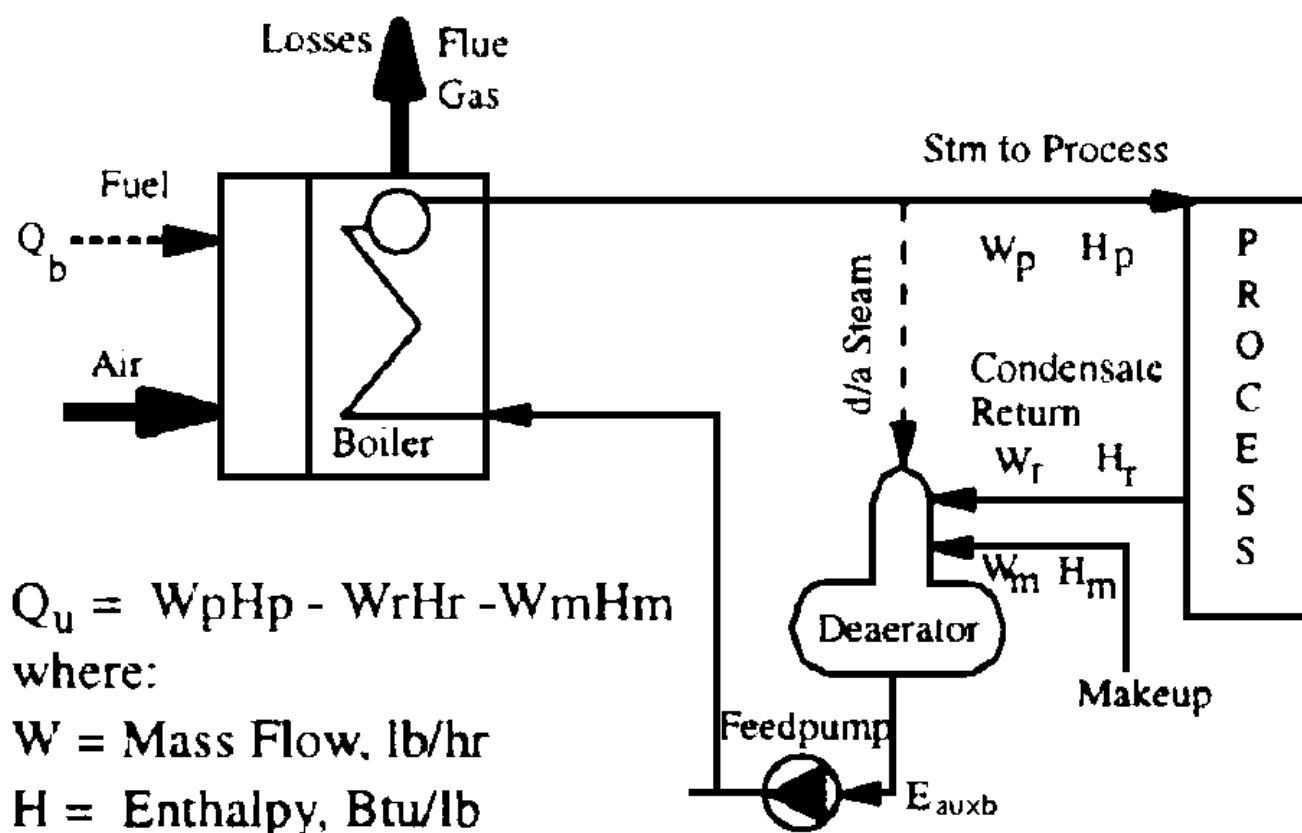


Figure 73.3 also illustrates another form of energy that is necessary in providing process heat. This is the auxiliary power necessary to run the boiler system, here represented as the electric power required to run the boiler feed pumps,  $E_{auxb}$ . This usually includes power to run boiler fans and other necessary devices and is typically less than 2% of the process heat, or about 1000 kW<sub>e</sub>, for the 50 MW<sub>t</sub> process shown in Fig. 73.2. This energy is not included in the boiler efficiency,  $\eta_b$ , which involves only the energy balance around the boiler. (Nor is any energy associated with condensing the d/a steam, since that energy does not leave the system and is present at the boiler inlet.) Further definitions for a cogenerated plant will involve  $\eta_b$  and  $E_{auxb}$  for a hypothetical heat-only plant delivering the same amount of process heat as the cogenerated plant.

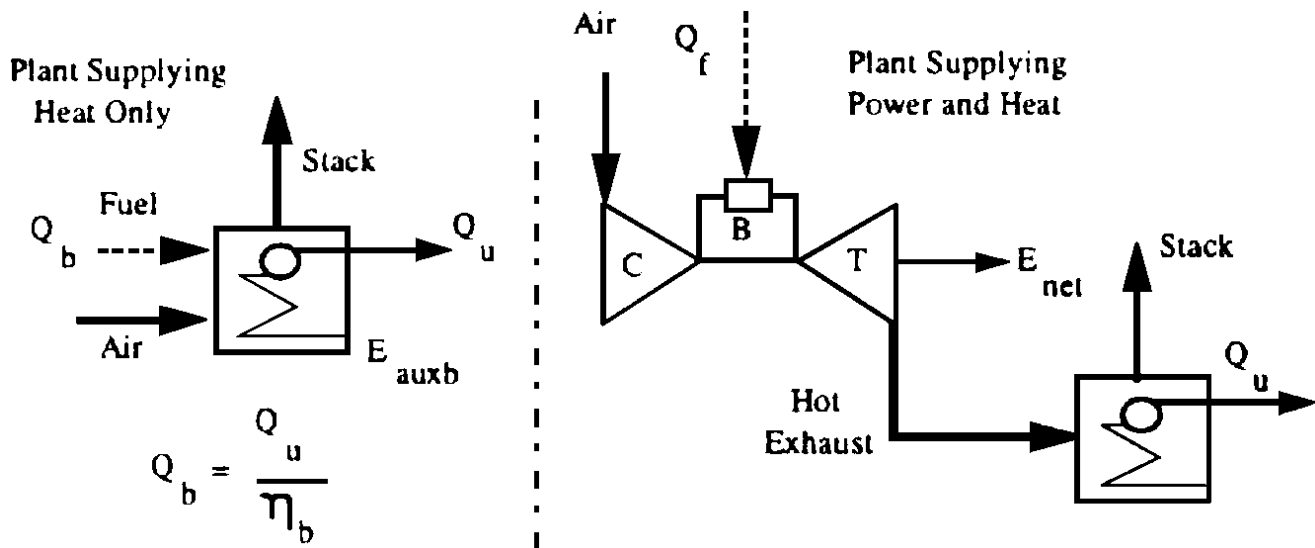
### Energy Utilization Factor (EUF) and Primary Energy Savings

These terms are defined in Fig. 73.2 as measures of the savings from cogeneration by comparing separate and combined production. However, as Horlock [1987] suggests, the EUF is not considered an efficiency, as it has been by some authors, because it mixes power and heat flow in the numerator rather than relating power to available thermal energy as in the classic thermal efficiency definition. Furthermore, it causes confusion with a more rigorous, industry-preferred, cogenerative efficiency definition that is based on noncogenerative heat rate and thermal efficiency as defined above.

### Cogenerative Efficiency

The industry-preferred cogenerative efficiency definition starts with **fuel chargeable to power (FCP)**, which is analogous to heat rate, in that it is the power related fuel required, in Btu/h, per kW<sub>e</sub> produced. Figure 73.4 shows the relations needed to define FCP. Numerical values correspond to the right side in Fig. 73.2.

**Figure 73.4** Fuel chargeable to power. (Adapted from Kovacic, J. M. 1983. *Industrial Gas Turbine Cogeneration Application Considerations*. Report No. GER-3430. Gas Turbine Reference Library, General Electric Company, Schenectady, New York.



$$\begin{aligned}
 \text{FCP} &= \frac{\text{Total fuel} - \text{Process fuel credit}}{\text{Net power} + \text{Process aux pwr. credit}} \\
 \text{FCP} &= \frac{(Q_f - Q_u/\eta_b) \times 3412.14}{E_{\text{net}} + E_{\text{auxb}}} \\
 \text{FCP} &= \frac{(100\,000 - 50\,000/0.93) \times 3412.14}{35\,000 + 1000} \\
 &= 4382.4 \text{ Btu/kWh (LHV)}
 \end{aligned}$$

By analogy with thermal efficiency, defined earlier relative to heat rate, we now can state, using Btu, **cogenerative efficiency** as

$\eta_{\text{FCP}} = (E_{\text{net}} + E_{\text{auxb}})/(Q_f - Q_u/\eta_b) = 3412.14/\text{FCP}$  [Horlock, 1987; Kovacik, 1983]. A numerical example, using Fig. 73.2 (right side), is  $3412.14 / 4382.4 = 77.9\%$ . The validity of the FCP relationship is apparent from applying it to the extremes of  $E_{\text{net}}/Q_u$ . If  $Q_u$  approaches 0,  $E_{\text{net}}/Q_u$  becomes very large and FCP reduces to the noncogenerative power heat rate. This effectively happens when  $E_{\text{net}}/Q_u$  is larger than about 5 [Boissenin, 1992]. The value  $\eta_{\text{FCP}}$  is maximized when  $Q_u$  is equal to the noncogenerative heat rejection from pure power production for the given power system. For a power system with noncogenerative thermal efficiency of 50%, this peak occurs when  $E_{\text{net}}/Q_u = 1$  and all of the erstwhile rejected heat is used to supply the process. For  $E_{\text{net}}/Q_u$  approaching zero, the system becomes "heat only," with thermal efficiency approaching that of the industrial boiler in Fig. 73.2.

### Fuel Split

The FCP does more than provide an index of energy efficiency. It also provides one means of dividing the fuel between the two products, which is necessary for allocation of fuel cost in economic evaluation. For example, again using the numbers in Fig. 73.2, the total cogenerated fuel is 100 MW<sub>t</sub> and the net power is 35 MW<sub>e</sub>. From above, the FCP is 4382.4 Btu/kWh, so that the power production share of the fuel is  $35\,000 \times 4382.4 = 153.38 \cdot 10^6$  Btu/h, or 45 MW<sub>t</sub> (dividing by 3412.14). The remainder,  $100 - 45 = 55$  MW<sub>t</sub>, would be fuel chargeable to heat, meaning chargeable to the **thermal host**.

The problem with the foregoing, as pointed out by Horlock [1987] and others, is that, strictly thermodynamically, there is no alternative to giving equal "weight" to a unit of heat and a unit of power, and, in allocating the fuel using FCP as the basis of cogenerative efficiency, *all* of the fuel savings from cogeneration are credited to the power producer. Horlock suggests a "value-weighted" efficiency and fuel allocation method that usually charges less of the fuel to the thermal host according to the ratio of the "sale value" of a kWh of heat to a kWh of power. There are several more sophisticated approaches to economic optimization beyond the scope of this article, that are given in Horlock and other references cited later.



## 73.2 Examples of Cogeneration

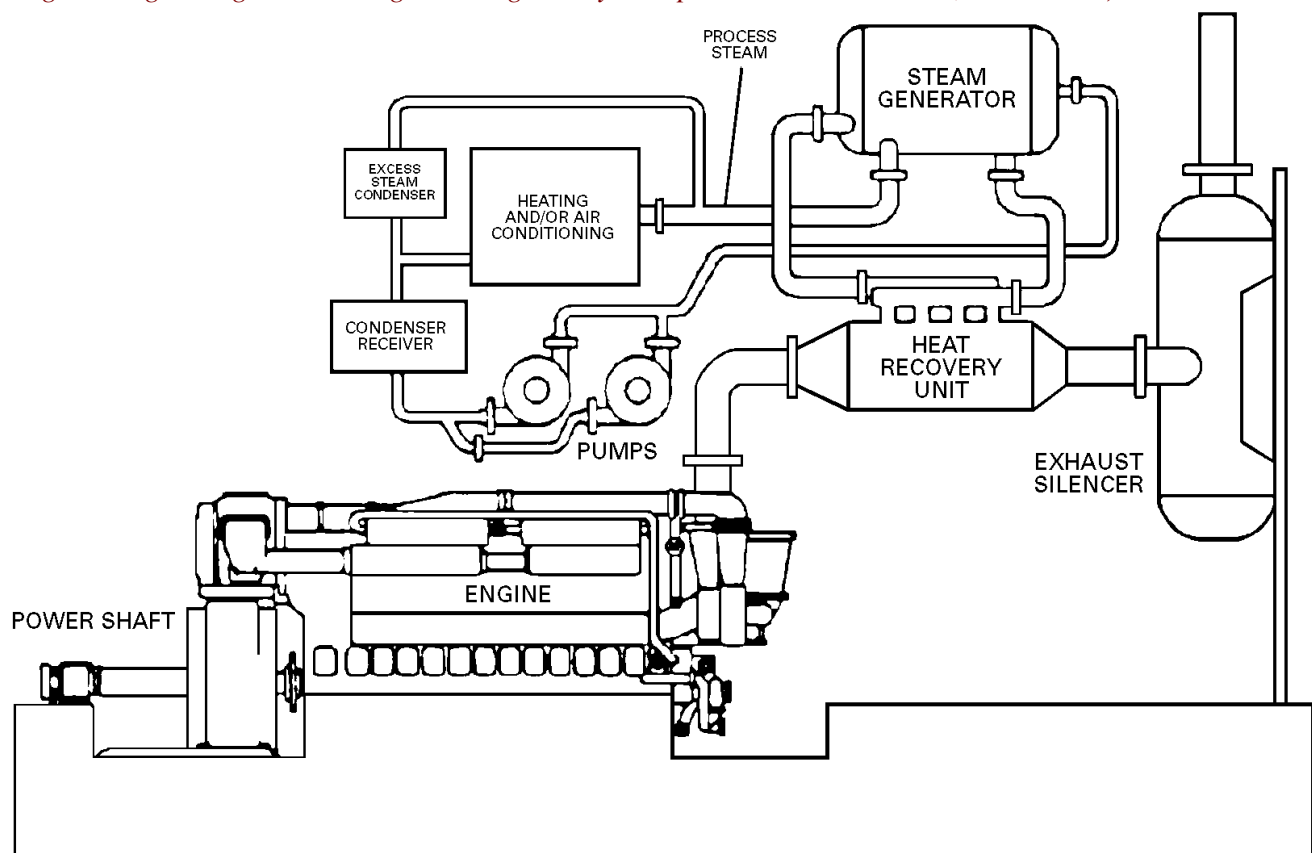
Cogeneration plants may be classified according to the *thermal sequence* of power and heat production, which relates to the type of industries served. Plants so classified are said to involve "topping" or "bottoming" cycles. The *fuel* is another characteristic that affects the design of the plant. The magnitude of the energy output will determine the *physical size* of the plant, which affects the type of machines used. Brief descriptions of typical plants as affected by each of these characteristics are given in the following paragraphs.

### Thermal Sequence

The topping cycle—in which the major conversion to power is thermally *above* the delivery of heat to process—is by far the most effective and economically attractive form of cogeneration. This is because, in power production, the heat most readily available for cogeneration is that which would otherwise be rejected at low temperatures. Further, conversion to power, for which demand and potential revenue are highest, is most efficient at high temperatures. Nevertheless, the temperature level of heat rejection can be high enough for many processes, and/or the temperature range of effective power and heat utilization can be shared.

Figure 73.5 shows a typical topping cycle example involving a diesel engine. For small to medium power systems up to 10 MW<sub>e</sub>, the diesel can be very economical with regard to capital and operating cost. The exhaust temperature, ranging from 1000 to 1400°F, can be used to generate hot water and/or steam through a heat recovery steam generator (HRSG).

**Figure 73.5** Small topping cycle cogeneration plant. (Source: Butler, C. H. 1984. *Cogeneration: Engineering, Design, Financing, and Regulatory Compliance*. McGraw-Hill, New York.)





In a bottoming cycle the primary energy source is first applied to a useful heating process at high temperature, and the rejected heat emerging from the process is then used for power production. It is not as effective as the topping cycle, because the Carnot/second-law principle of low-temperature heat rejection during conversion of heat into work is on the wrong side, working against effective heat recovery rather than with it. Obviously, the bottoming cycle is necessary only when the process need is for relatively large amounts of high-temperature heat, such as in industrial furnaces and kilns. [Figure 73.6](#) shows a high-temperature furnace process for production of ethylene,  $C_2H_4$ , the raw material for manufacturing plastics. Ethylene is produced by "cracking" heavier molecules such as ethane,  $C_2H_6$ , by raising the temperature of the ethane feedstock to 1600°F in a furnace maintained at 2100°F by burners firing natural gas. The feedstock flows inside a tube in a flat "cracking coil" (S-shaped dashed line in [Fig. 73.6](#)). As the molecules of ethane break into ethylene (and other compounds), it is important to quench, or rapidly cool the "cracked gas" to optimize the "yield" and prevent the cracking process from proceeding too far, forming carbon (coke). A steam generation loop (quench loop) accomplishes this by generating high-pressure steam, thus cooling the gas. The steam thus formed is superheated in the hot convective section of the furnace outlet where the stack gases are cooled, recovering the leftover heat from the furnace. The superheated steam is used in steam turbines that drive compressors for the downstream ethylene separation processes. The quench loop and the steam turbocompressors will be recognized as a bottoming cycle. Note, also, that a condenser is used, which indicates that it was impractical to recover all of the furnace heat due to the down-side location of the power cycle.

**Figure 73.6** Ethylene furnace bottoming cycle (and topping cycle). (*Source: Cooke, D. H. 1987. Combined Cycle Thermodynamic Inquiries and Options for Cogeneration Facilities in the Process Industry. ASME Paper No. 87-JPGC-Pwr-61. ASME, New York.*)

**Figure 73.6**

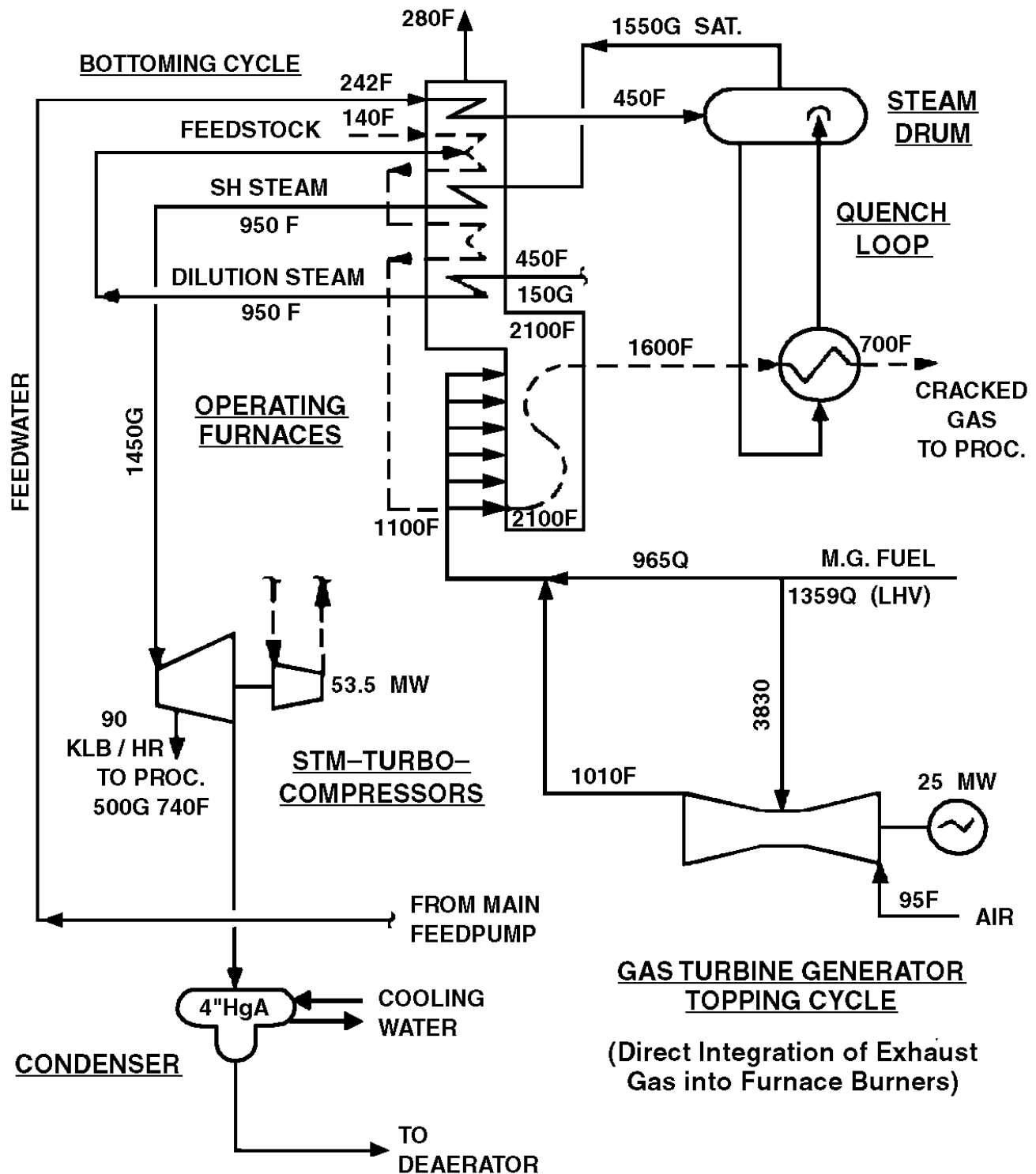


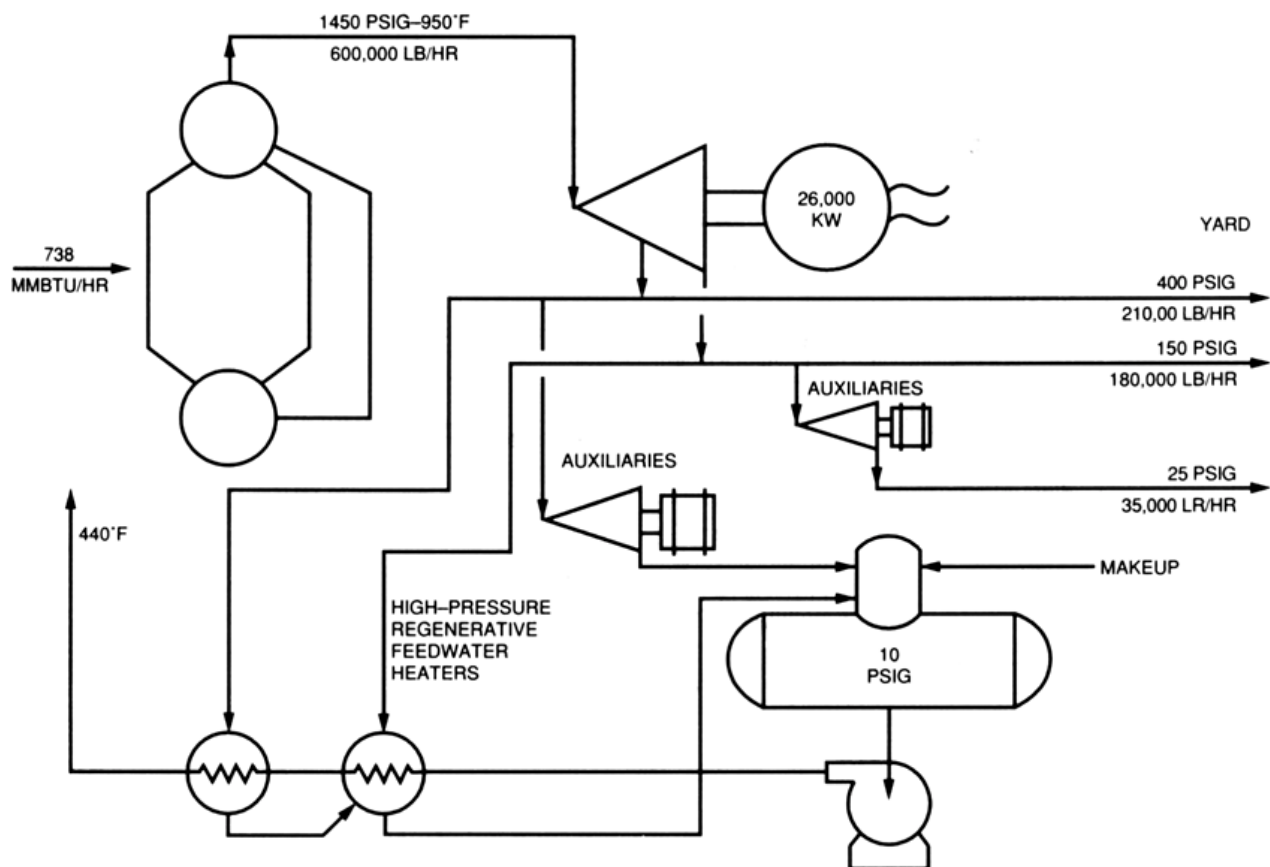
Figure 73.6 also illustrates a *topping* cycle consisting of several gas turbines generating 28 MW<sub>e</sub> of electric power each and exhausting directly into the furnace burners. There is enough residual oxygen in the gas turbine exhausts that the 1010°F exhaust gas acts as heated combustion

air, reducing the fuel required to maintain the furnaces "at temperature."

## Plants for "Difficult" Fuels

Over the years, the workhorse of utility and industrial power generation has been the steam power, or Rankine, cycle. This type of plant features a furnace type, near-stoichiometric-combustion boiler, producing steam at high pressure and temperature in a separate closed loop free of erosive particles for expansion through the turbines. In recent years, technological advances with the gas turbine have produced plants with considerably higher efficiency and lower capital cost than the pure Rankine. However, the pure Rankine still enjoys one advantage in that cheaper, and possibly more abundant, solid fuels such as coal, lignite, and coke—more difficult in terms of handling, burning, and waste disposal than gaseous or liquid fuels—can be used. Figure 73.7 shows a typical pure Rankine cycle cogeneration plant supplying steam for process heat at several pressure levels in a refinery.

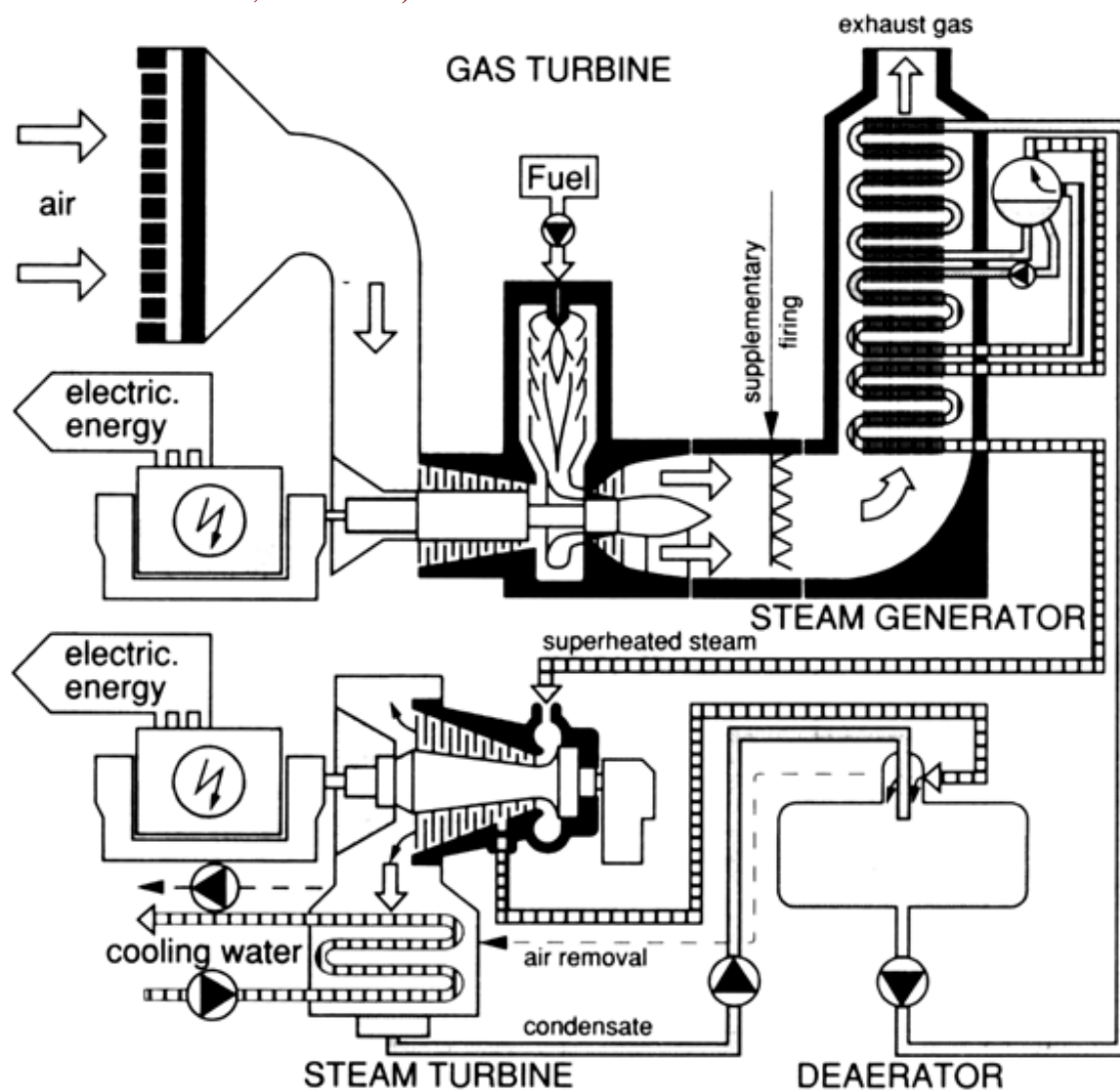
**Figure 73.7** Rankine cycle cogeneration plant. (Source: Cooke, D. H. 1987. *Combined Cycle Thermodynamic Inquiries and Options for Cogeneration Facilities in the Process Industry*. ASME Paper No. 87-JPGC-Pwr-61. ASME, New York.)



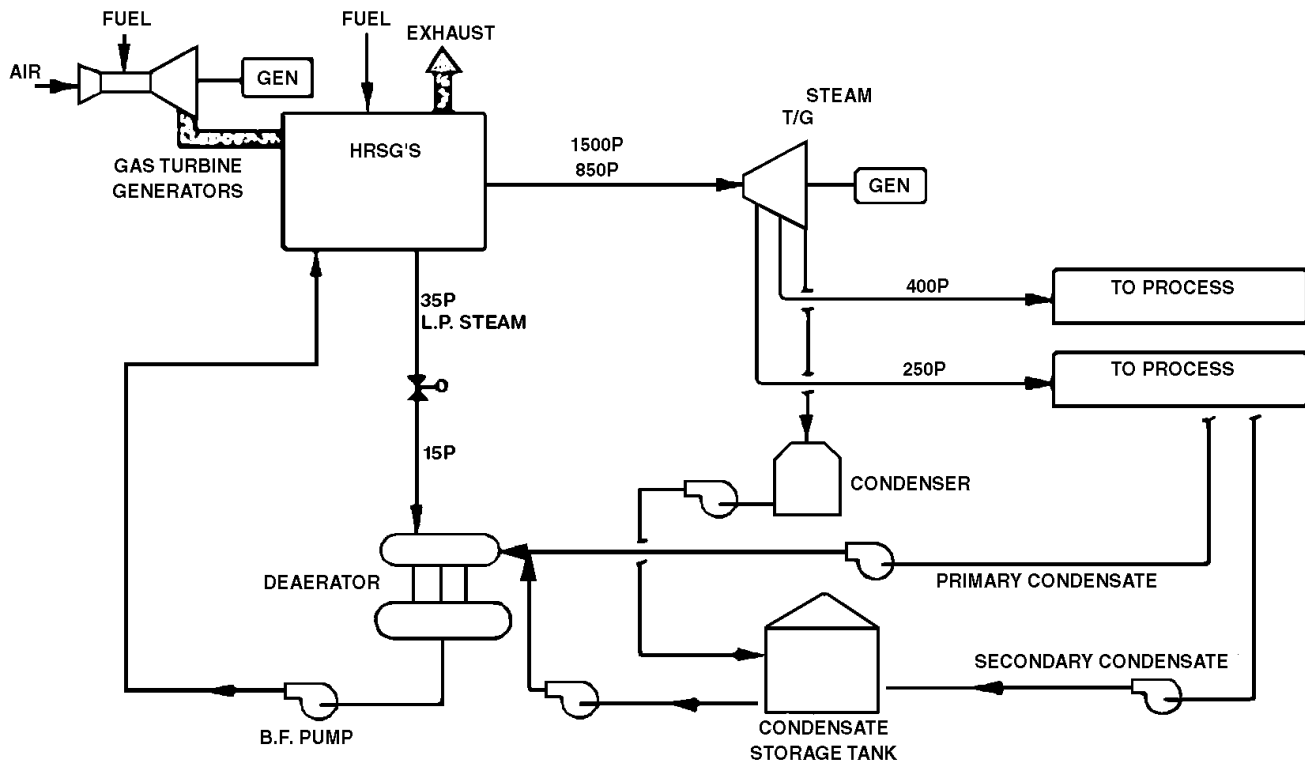
## Large-Scale Power and Cogeneration

The gas turbine, or Brayton/Joule, cycle has come into its own as the cycle of choice for large-scale power (typically up to 230 MW<sub>e</sub> with a single engine in *simple cycle*) and cogeneration systems. Using technology developed for aircraft, the gas turbine produces power very efficiently at extremely high temperatures (typically 2200°F, versus 1000°F maximum for steam), exhausting at about 1000°F, so that a complete Rankine cycle steam plant, with the gas turbine exhaust replacing the furnace boiler, can *double* the simple cycle plant output in **combined cycle**. Safe, reliable, combined cycle power systems of this type typically reach 50% net thermal efficiency, compared to 35% for the best pure Rankine plants. The combined cycle offers great flexibility for cogeneration. For example, instead of converting the high-level exhaust heat to power, the entire steam output can be used for process as shown in Fig. 73.4. A complete combined cycle power system is shown conceptually in Fig. 73.8, including provision for supplementary firing in the gas turbine exhaust duct, which gives excellent control over a wide range of increased steam production. The use of such a system in a large multilevel steam cogeneration plant is illustrated in Fig. 73.9.

**Figure 73.8** Gas turbine combined cycle. (Source: Cooke, D. H. 1987. *Combined Cycle Thermodynamic Inquiries and Options for Cogeneration Facilities in the Process Industry*. ASME Paper No. 87-JPGC-Pwr-61. ASME, New York.)



**Figure 73.9** Combined cycle cogeneration plant. (Source: Butler, C. H. 1984. *Cogeneration: Engineering, Design, Financing, and Regulatory Compliance*. McGraw-Hill, New York.)



## Defining Terms

**Auxiliary power:** Parasitic electric power necessary to operate cycle equipment that must be accounted for in computing net output to the grid from gross generated power.

**Cogeneration:** The production, from the same fuel or energy source, of electric and/or shaft power sequentially together with process heat.

**Cogenerative efficiency:** The apparent thermal efficiency of cogenerated power production, taking credit for all of the fuel energy saved by cogeneration. It is analogous to thermal efficiency in a plant producing power only and is the reciprocal of FCP with units adjusted.

**Combined cycle:** A gas turbine, separately known in *simple cycle* as the Brayton/Joule cycle, combined with a Rankine cycle, consisting of a heat recovery steam generator (HRSG) in the downstream exhaust duct and a steam turbine converting some or all of the steam energy to power.

**Fuel chargeable to power (FCP):** Cogenerative heat rate, for example, in Btu/kWh. The total fuel heat release less the process fuel credit (gross heat release required for "heat only" production of net heat to process) divided by the cogenerated net power output plus the uncogenerated process auxiliary power. It is the reciprocal of cogenerative efficiency with adjustment of units.

**Heat rate:** In a plant producing power only, the fuel heat release per unit of power, for example,

in Btu/kWh. It is the reciprocal of thermal efficiency with adjustment of units.

**Lower heating value (LHV):** The energy released by combustion of a unit mass or standard volume of a given fuel, excluding the latent heat of water vapor formed from combustion of hydrogen in fuel compounds. Must be calculated from *higher heating value (HHV)*, which includes this, as determined by calorimeter. Important for definition of comparative efficiencies and heat rates.

**Net heat to process:** The difference in total enthalpy, for example, in Btu/h, between that of the steam delivered to process and that of the sum of condensate returned plus makeup required.

**Thermal efficiency:** In plants producing power only or heat only, the dimensionless ratio of power or thermal output to fuel input. Usually expressed in percent. For power-only plants it is the reciprocal of heat rate with adjustment of units.

**Thermal host:** In a cogeneration venture, the industrial purchaser of the heat to process.

**Wheeling:** The transmission of electric power from its source in one utility's grid system to consumers in another's, sometimes passing through systems of intermediate utilities. Important for distributing cogenerated power from concentrated industrial areas.

## References

- Boissenin, Y. 1992. Cogeneration and energy savings. *GEC Alsthom Technical Review No. 10*. European Gas Turbines S.A., Belfort, France.
- Butler, C. H. 1984. *Cogeneration: Engineering, Design, Financing, and Regulatory Compliance*. McGraw-Hill, New York.
- Cooke, D. H. 1987. *Combined Cycle Thermodynamic Inquiries and Options for Cogeneration Facilities in the Process Industry*. ASME Paper No. 87-JPGC-Pwr-61. ASME, New York.
- Horlock, J. H. 1987. *Cogeneration: Combined Heat and Power, Thermodynamics and Economics*. Pergamon Press, Oxford, U.K.
- Kovacik, J. M. 1983. *Industrial Gas Turbine Cogeneration Application Considerations*. Report No. GER-3430. Gas Turbine Reference Library, General Electric Company, Schenectady, NY.

## Further Information

The following organizations sponsor annual conferences and seminars on cogeneration with published material. Schedules of forthcoming conferences may be obtained by contacting them.

The International Gas Turbine Institute of the American Society of Mechanical Engineers (ASME), 5801 Peachtree Dunwoody Road, Suite 100, Atlanta, GA 30342-1503. Phone (407) 847-0072. Fax (407) 847-0151 or (407) 843-2517.

The Cogeneration and Competitive Power Institute of the Association of Energy Engineers, 4025 Pleasantdale Road, Suite 420, Atlanta, GA 30340. Phone (404) 925-9633. Fax (404) 381-9865.

Stanek, E. K. "Electric Machines"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## 74.1 Induction Machines

## 74.2 Synchronous Machines

## 74.3 DC Machines

### E. Keith Stanek

*University of Missouri, Rolla*

The study of electrical machines is important because a very significant percentage of the world's energy usage is accomplished by conversion from chemical, mechanical, or nuclear form into electrical form. The motivation for doing this is the ease of transmitting the energy and the exceptional control we have over the flow of this energy when it is in electrical form. One major disadvantage is that it is not possible to store significant energy in electrical form. Therefore, it is necessary to match electrical energy generation and consumption within a system on an instantaneous basis. When this is not accomplished, the system frequency tends to rise for an excess of generation and sag for a deficiency in generation.

The goal of this chapter will be to present equivalent circuits that can be used to perform steady state analysis of the three major types of rotating machines: three-phase induction, three-phase synchronous AC machines, and direct current machines. These models or equivalent circuits allow the calculation of performance indices such as efficiency, losses, output torque, output power, voltage or speed regulation, and so on. Space limitations will preclude transient or dynamic analysis of these devices. It is important to realize that the equivalent circuits presented are *not* sufficiently general for use in transient situations with the exception of calculating some motor-starting characteristics.

## 74.1 Induction Machines

---

Like all rotating electric machines, induction machines consist primarily of iron for magnetic paths, copper for electrical current flow, and insulation to separate the windings from the iron of the stator and rotor. The stator has a distributed three-phase winding embedded in slots. The rotor may have either a distributed three-phase winding or a series of copper or aluminum bars embedded in slots with shorting straps at each end of the rotor. In the former case the machine is a wound rotor type and in the latter case the machine is a **squirrel cage** type. Induction machines are usually used as motors. They are only occasionally used as generators. Therefore, this chapter will be limited to a discussion of motor operation.

The three-phase induction machine model is shown in [Fig. 74.1](#). All parameters in [Fig. 74.1](#) are

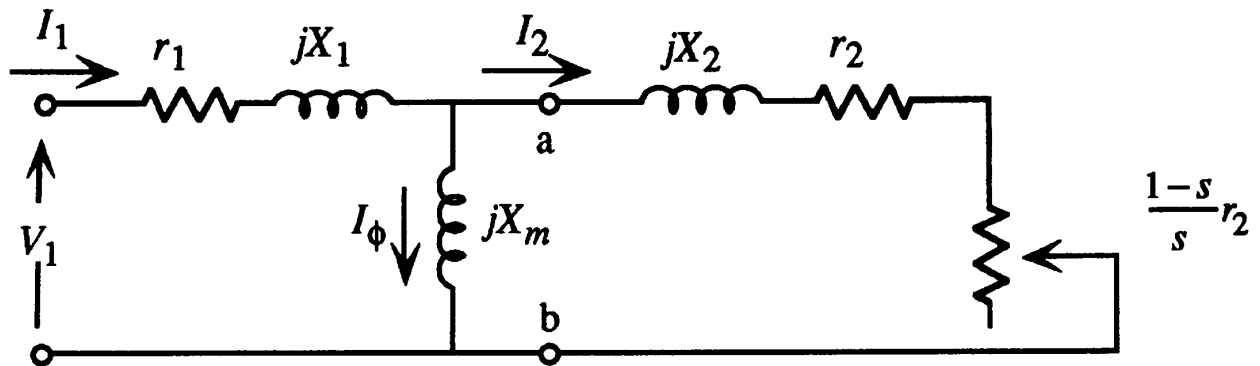


referred to the stator and are defined as follows:

- $r_1$  = Stator resistance per phase in ohms
- $r_2$  = Rotor resistance per phase in ohms
- $X_1$  = Stator leakage reactance per phase in ohms
- $X_2$  = Rotor leakage reactance per phase (ohms)
- $X_m$  = Magnetizing reactance per phase (ohms)
- $V_1$  = Line-to-neutral stator voltage (volts)
- $I_1$  = Stator current (amperes)
- $I_2$  = Rotor current (amperes)
- $s$  = Slip (per unit)  

$$= \frac{\omega_s - \omega}{\omega_s} = \frac{n_s - n}{n_s}$$
- $\omega_s$  = Synchronous rotor speed (radians per second =  $2\pi f/(P/2)$  )
- $n_s$  = Synchronous rotor speed (revolutions per minute =  $60f/(P/2)$  )
- $f$  = Stator frequency (hertz)
- $P$  = Number of stator or rotor poles
- $\omega$  = Actual rotor speed in radians per second
- $n$  = Actual rotor speed in revolutions per minute

**Figure 74.1** Basic equivalent circuit of a three-phase induction motor.



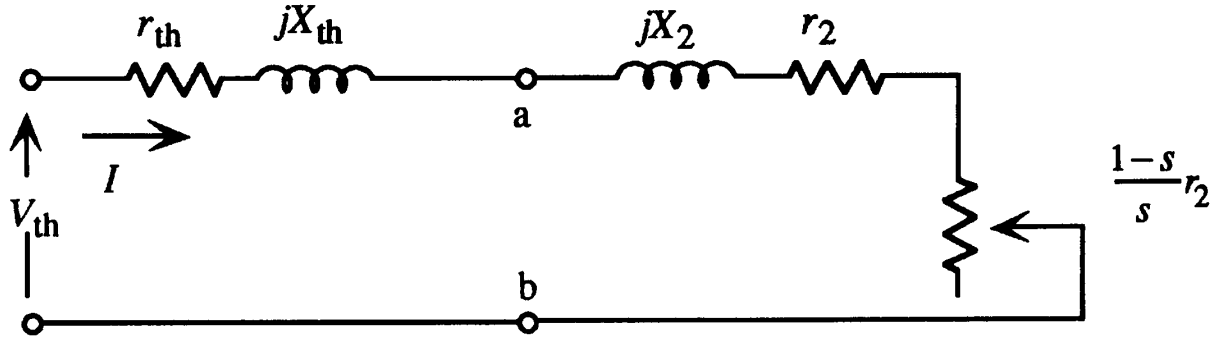
The resistor with ohmic value  $[(1 - s)/s]r_2$  is of special significance. The energy consumed in this resistor represents the energy transferred across the air gap from the stator to the rotor on a per phase basis.

This equivalent circuit is best analyzed by use of Thevenin's theorem. If this theorem is applied at terminals a-b looking toward the left, the resulting circuit is shown in Fig. 74.2. The new parameters are

$$V_{th} = \frac{jX_m}{r_1 + j(X_1 + X_m)} V_1 \quad (74.1)$$

$$r_{th} + jX_{th} = \frac{r_1 X_m (X_1 + X_m)}{r_1^2 + (X_1 + X_m)^2} + \frac{jX_1 X_m (X_1 + X_m)}{r_1^2 + (X_1 + X_m)^2} \quad (74.2)$$

**Figure 74.2** Revised equivalent circuit of a three-phase induction motor.



The current flowing in this circuit can be easily calculated if a value of slip,  $s$ , is assumed:

$$I = \frac{V_{th}}{\left( r_{th} + r_2 + \frac{1-s}{s} r_2 \right) + j(X_{th} + X_2)} \quad (74.3)$$

The power delivered across the air gap is simply

$$P_{air\ gap} = 3I^2 \frac{(1-s)}{s} r_2 \quad (74.4)$$

In this equation, the factor of 3 results from the fact that this is a per phase model and there are three phases.

Not all of the power calculated in these equations results in useful mechanical energy. The only losses accounted for in the model are stator and rotor copper losses. Other losses include core losses, friction and windage losses, and stray load losses. Generally, these losses can be approximated as being fixed or constant with varying load. This power loss,  $P_{fixed}$ , can be deducted to yield the mechanical power out,  $P_{mech}$ .

$$P_{mech} = P_{air\ gap} - P_{fixed} \quad (74.5)$$

Since power is torque times radial velocity,

$$P_{\text{mech}} = T_{\text{mech}} \omega \quad (74.6)$$

$$T_{\text{mech}} = P_{\text{mech}} / \omega \quad (74.7)$$

The stator current,  $I_1$ , can be found by using the current divider equation.

$$I_2 = I = \frac{jX_m}{r_2 + [(1-s)/s]r_2 + jX_2} I_1 = \frac{jX_m}{r_2/s + jX_2} I_1 \quad (74.8)$$

$$I_1 = \frac{r_2/s + jX_2}{jX_m} I \quad (74.9)$$

Then, it is easy to find the input power,

$$P_{\text{in}} = 3V_1 I_1 \cos \theta \quad (74.10)$$

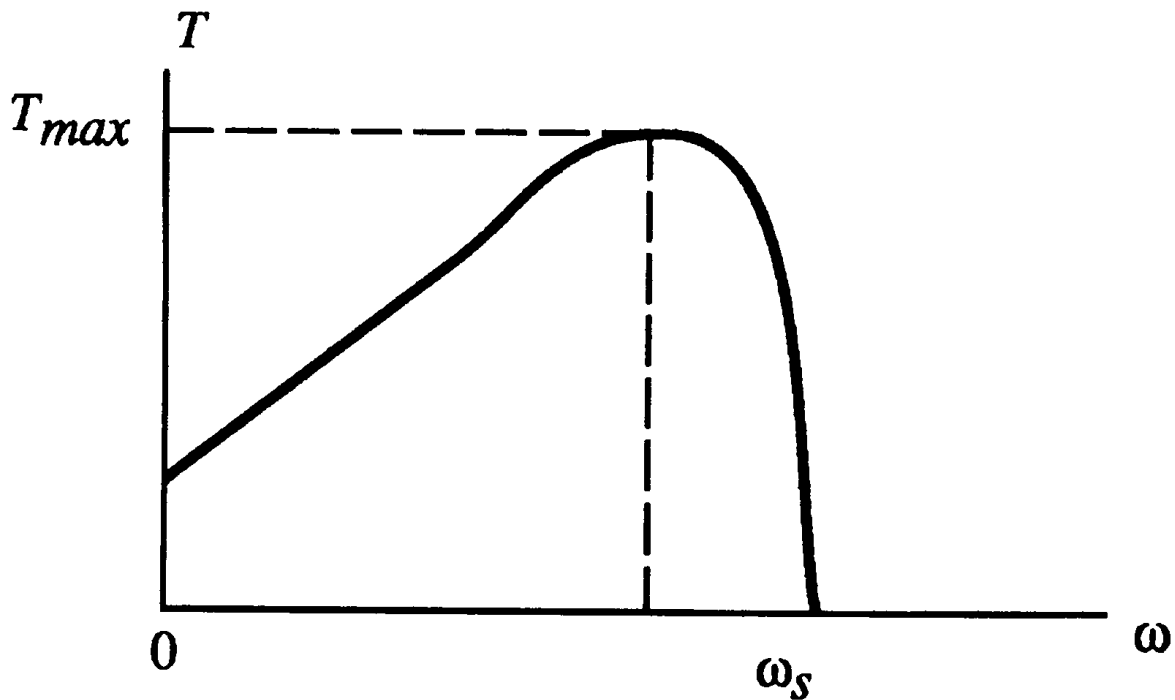
where  $\theta$  is the angle by which  $V_1$  leads  $I_1$ .

It is now possible to analyze all aspects of induction motor performance by assuming a value of slip and calculating, in order, the rotor speed, the Thevenin voltage, Thevenin impedance, rotor current, stator current, rotor copper loss, stator copper loss, power in, air gap power, mechanical power, output torque, total losses, and efficiency.

If a number of different values of slip are assumed between 0 and 1, the performance of the induction motor from standstill to synchronous speed can be found. The torque plotted as a function of rotor speed is shown in [Fig. 74.3](#). A couple of other interesting things can be found from the equivalent circuit. One of these is the starting current. At start-up, rotor speed is zero and slip is unity (1.0). Using a value of 1.0 for slip and neglecting the small current through the exciting inductance yields the starting current in amperes

$$I_{\text{start}} = \frac{V_1}{r_1 + r_2 + j(X_1 + X_2)} \quad (74.11)$$

**Figure 74.3** Typical torque-speed curve of a three-phase induction motor.



Another quantity of interest is the maximum or pull-out torque. This quantity can be found by applying the maximum power transfer theorem to the total rotor resistance  $r_2/s$ . The slip corresponding to maximum torque is

$$s_m = \frac{r_2}{\sqrt{r_1^2 + (X_1 + X_2)^2}} \quad (74.12)$$

and the torque at this slip is (in N·m)

$$T_{\max} = \frac{3V_{\text{th}}^2}{\omega_s^2 \left( r_1 + \sqrt{r_1^2 + (X_1 + X_2)^2} \right)} \quad (74.13)$$

## 74.2 Synchronous Machines

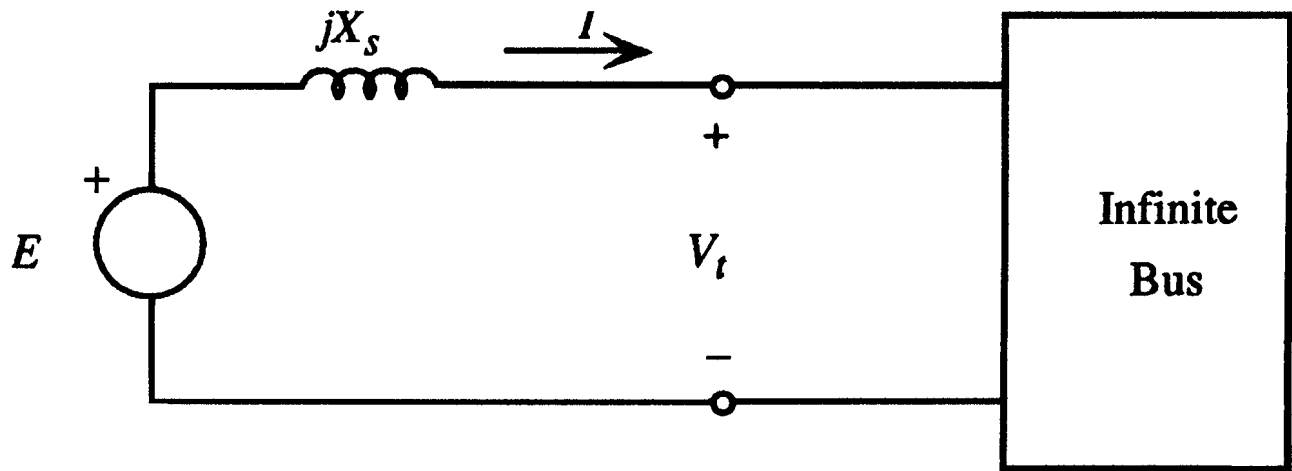
A three-phase synchronous machine has a stator winding very similar to a three-phase induction motor. The rotor is quite different. The rotor has a single winding consisting of many turns of fine

wire that is excited by a DC voltage supply connected to the rotor via brushes and slip rings.

The equivalent circuit (see Fig. 74.4) of a synchronous machine is one of the simplest for all machine types, but there are some important subtleties involved in the analysis of this circuit that need to be discussed. Variables in Fig. 74.4 are defined as follows:

- $E$  = Internal EMF in V
- $V_t$  = Terminal voltage in V
- $I$  = Armature current in A
- $X_s$  = Synchronous reactance in ohms.

**Figure 74.4** Equivalent circuit of a synchronous generator.



Consider the operation of the synchronous machine as a generator while it is connected to a large system. In the limit the "large system" would have unlimited capacity and it is referred to as an **infinite bus**. The characteristics of an infinite bus are that it has perfectly stable voltage and frequency and the phase angle is constant. If a synchronous machine is connected to an infinite bus, the terminal voltage of the machine is forced to be equal to the infinite bus voltage in both magnitude and angle. The real power output of the generator will not cause the infinite bus frequency to either sag or rise.

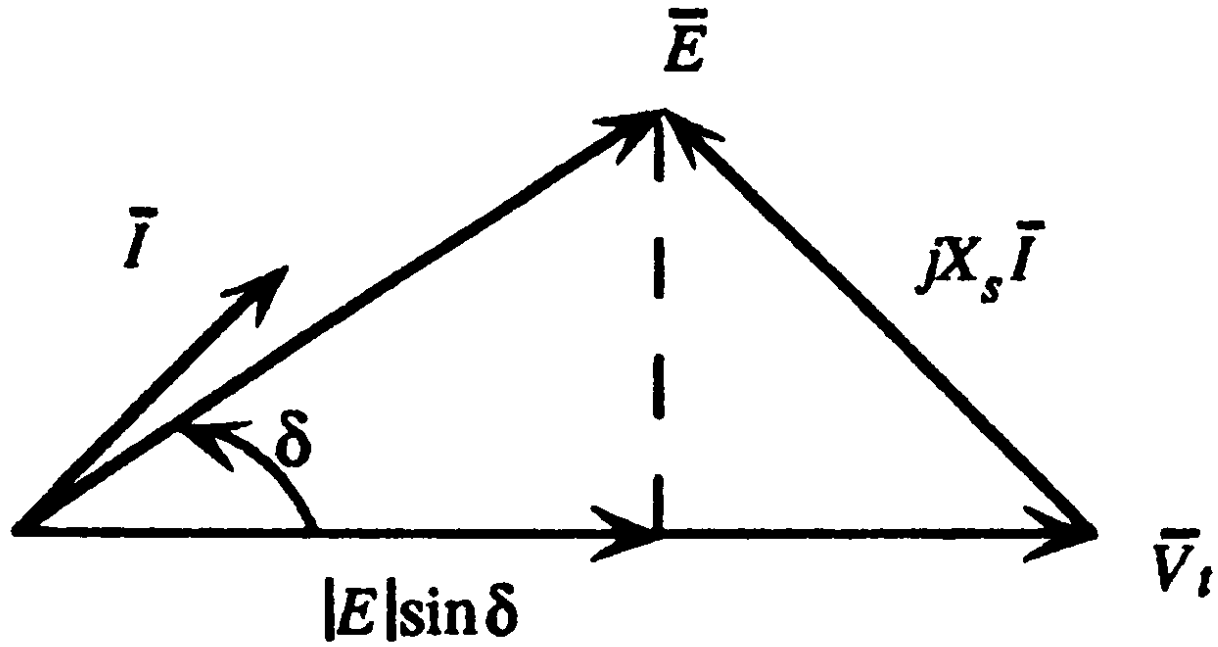
With these facts in mind, consider what happens when one of the two controllable inputs to the synchronous generator—real power in and field excitation level—is varied. Assume the real power in is varied. In order to visualize what happens, assume generator resistance is negligible and the synchronous reactance is  $X_s$  in ohms. Let the generator internal voltage be  $|E|\angle\delta = \overline{E}$ . Kirchhoff's voltage law (KVL) states that

$$\overline{E} = jX_s\overline{I} + \overline{V}_t \quad (74.14)$$

This equation is represented by the phasor diagram in Fig. 74.5. Therefore,

$$\bar{I} = \frac{\bar{E} - \bar{V}_t}{jX_s} \quad (74.15)$$

**Figure 74.5** Phasor diagram of a synchronous generator.



Assume the infinite bus voltage is  $\bar{V}_t = |\bar{V}_t| \angle 0^\circ$ . Then,

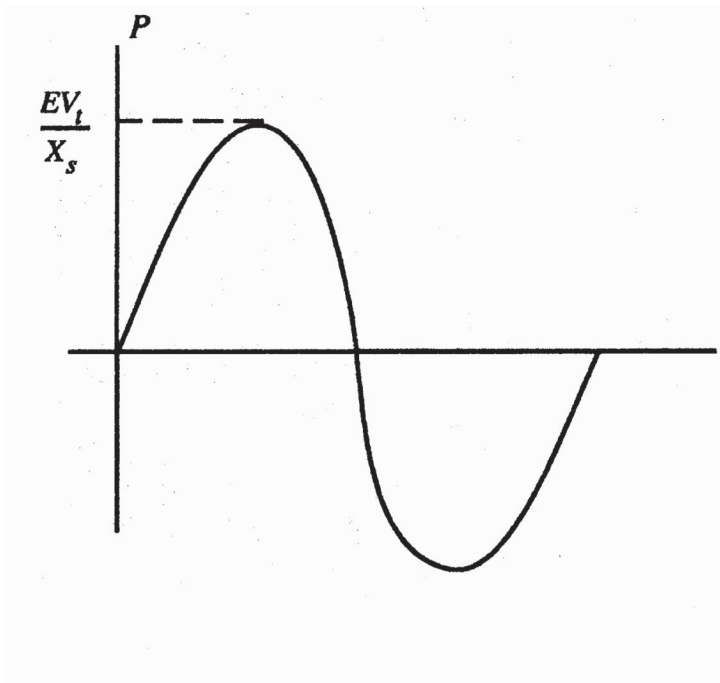
$$\bar{I} = \frac{E \angle \delta - V_t \angle 0^\circ}{jX_s} = \frac{E \angle \delta - 90^\circ}{X_s} + \frac{V_t}{X_s} \angle 90^\circ \quad (74.16)$$

The real power delivered by the internal EMF is

$$P = \text{Re}(\overline{EI}^*) = \frac{EV_t}{X_s} \sin \delta \quad (74.17)$$

Therefore, the real power generated or transformed from mechanical to electrical form is dependent on  $\delta$ , the angle by which  $\overline{E}$  leads  $\overline{V}_t$ . With fixed field current, when the mechanical power is increased, the angle  $\delta$  increases to allow the electrical power out to balance mechanical power in. Clearly, there is a limit. When  $\delta = 90^\circ$  or  $\pi/2$ , the maximum power is reached. If mechanical power is increased further, synchronism will be lost. The dependence of power on  $\delta$  is illustrated by Fig. 74.6.

**Figure 74.6** Power-angle curve of a synchronous generator.



A separate analysis can be done assuming the mechanical power in is fixed and the field current (and hence the internal EMF,  $|\overline{E}|$ ) is variable. In this case it can be shown that the reactive power out of the generator is

$$Q = \frac{V_t}{X_s} (E \cos \delta - V_t) \quad (74.18)$$

In most cases, the angle  $\delta$  is small ( $15^\circ$  or less) and  $\cos \delta$  is approximately 1.0. Thus,  $E \cos \delta - V_t \approx E - V_t$ . Note that if  $E > V_t$ ,  $Q > 0$ , and if  $E < V_t$ ,  $Q < 0$ . Therefore, an

overexcited generator supplies reactive power, whereas an underexcited generator absorbs reactive power.

In summary, if a synchronous generator is supplying power to a large system, the amount of real power supplied is controlled by the power from the prime mover. The amount of reactive power supplied is controlled by the field excitation, which controls the magnitude of the internal voltage.

The equivalent circuit of Fig. 74.4 can be used for motor analysis by reversing the polarity of the current  $\bar{I}$ . Space limitations will not permit a discussion of the operation of a synchronous machine as a motor, but the analysis is quite similar to that for generator operation.

## 74.3 DC Machines

---

The construction of a DC machine is somewhat more complicated than either of the AC machine types discussed. Suffice it to say that a single winding of many turns of fine wire is located on the stator. This is the field winding. A single but more robust winding exists on the rotor. This is the armature or high-power winding. The path of currents through this winding is constantly changing with time due to the connection of this winding to the outside world via a series of brushes sliding over a **commutator**. The commutator is basically a mechanical rectifier that rectifies the AC voltages generated in the **armature winding** to produce a DC terminal EMF.

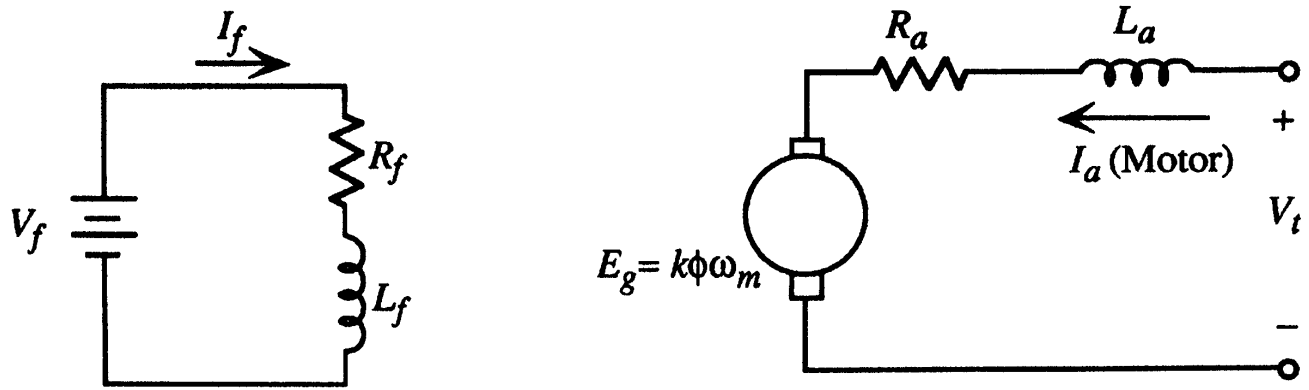
In this chapter only DC machines excited with shunt or separate fields will be discussed. The equivalent circuit for DC machines of this type is shown in Fig. 74.7. This circuit can be used for either motor or generator operation, but it is vital that the proper polarity for the armature current be used. Since DC machines are used primarily as motors, that mode of operation will be emphasized.

Terms used in Fig. 74.7 are defined as follows:

- $V_f$  = Field voltage in V
- $R_f$  = Field resistance in ohms
- $L_f$  = Field inductance in H
- $I_f$  = Field current in A
- $V_t$  = Terminal voltage in V
- $E_g$  = Generated internal voltage in V
- $R_a$  = Armature resistance in ohms
- $L_a$  = Armature inductance in H
- $I_a$  = Armature current in A
- $k$  = A constant relating generated voltage to flux and rotor speed
- $\phi$  = Air gap flux per pole in webers
- $\omega_m$  = Rotor speed in radians/second
- $T$  = Torque produced in N·m



**Figure 74.7** Equivalent circuit of a DC motor.



Note that the field copper loss is  $I_f^2 R_f$  and the armature copper loss is  $I_a^2 R_a$ . The machine also has a number of other losses due to friction, windage, and core losses. These are generally taken as fixed losses.

The product of the internal generated voltage and the armature current is equal to the power converted in the machine, electrical to mechanical in a motor or mechanical to electrical in a generator. By energy conservation,

$$E_g I_a = T \omega_m \quad (74.19)$$

or

$$k\phi\omega_m I_a = T \omega_m \quad (74.20)$$

so

$$T = k\phi I_a \quad (74.21)$$

The steady state DC equations governing motor operation are as follows:

$$V_f = R_f I_f \quad (74.22)$$

$$V_t = E_g + R_a I_a \quad (74.23)$$

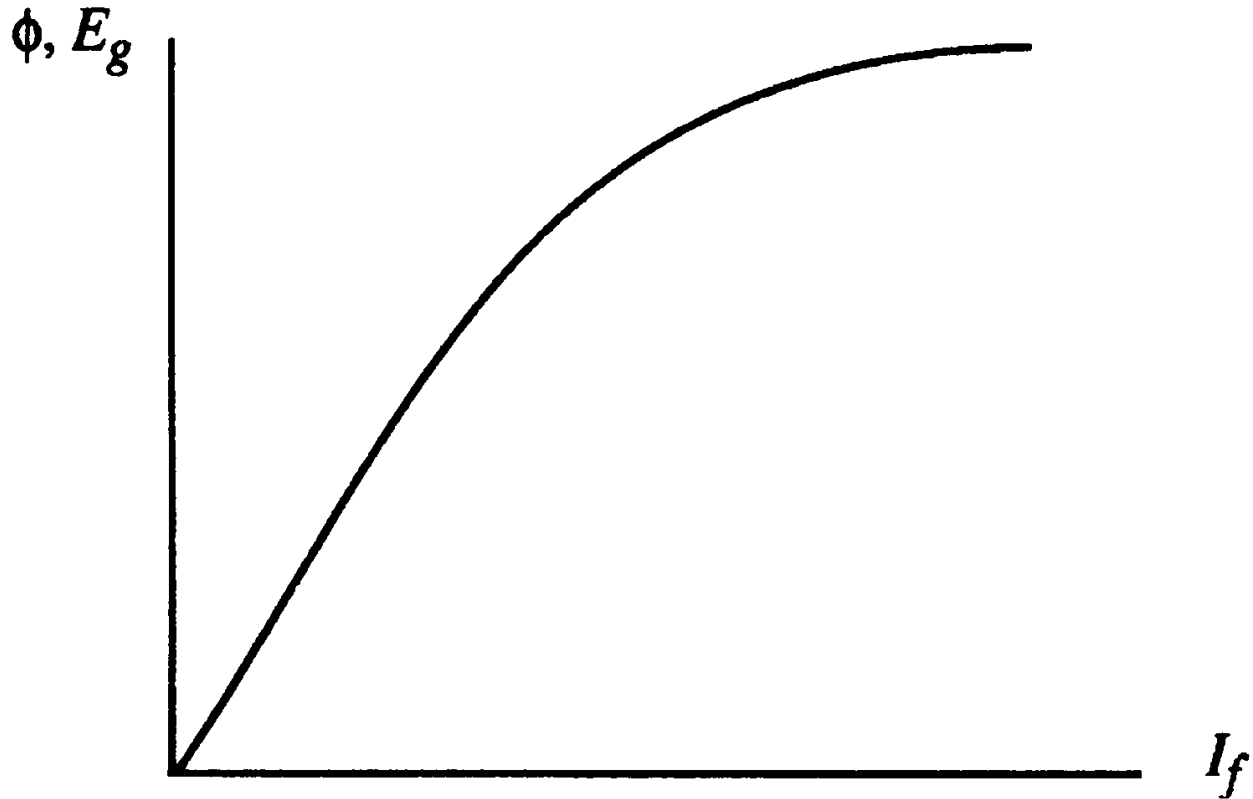
$$E_g = k\phi\omega_m \quad (74.24)$$

$$T = k\phi I_a \quad (74.25)$$

The flux per pole,  $\phi$ , is a function of the construction of the machine and the field current,  $I_f$ . A typical relationship is shown in Fig. 74.8. Note that, if  $\omega_m$  is constant,  $E_g$  is proportional to  $\phi$  and the vertical axis can be relabeled so that Fig. 74.8 becomes a plot of  $E_g$  versus  $I_f$  and is then called the machine's *open-circuit characteristic* at the given speed,  $\omega_m$ . In the linear region,

$$\phi = k_f I_f .$$

**Figure 74.8** Open-circuit characteristic of a shunt-wound DC motor.

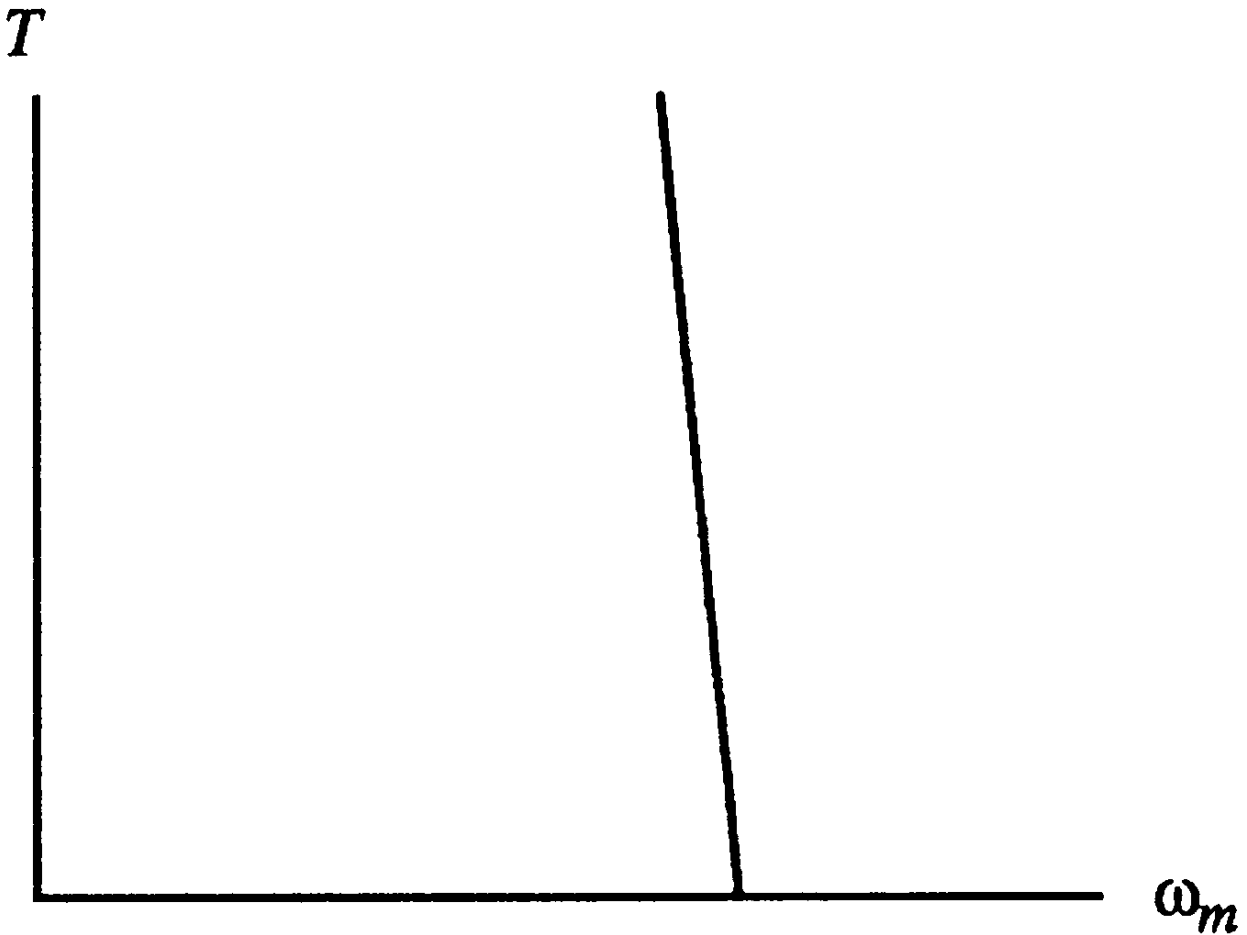


Solving Eq. (74.23) for  $I_a$  and substituting into Eq. (74.25) yields

$$T = \frac{k_f I_f V_t}{R_a} - \frac{k_f^2 I_f^2}{R_a} \omega_m \quad (74.26)$$

This relationship is the torque speed characteristic of the motor and is plotted in [Fig. 74.9](#). It is clear that the motor runs at nearly a constant speed as load changes.

**Figure 74.9** Torque-speed curve for a shunt-wound DC motor.



### Defining Terms

**Armature winding:** The power or high current winding in an electric machine. This contrasts to the field winding, which generally has lower current and/or voltage level.

**Commutator:** A mechanical rectifier used to convert alternating voltages, generated internal to a DC machine, into DC voltages at the terminals of the DC machine.

**Infinite bus:** An idealization of a large AC power system in which the rms voltage is constant, as is the system frequency and voltage phase angle, regardless of the load level supplied by the power system.

**Squirrel cage winding:** A series of conducting bars embedded in axial slots on the surface of the rotor of an induction motor that are shorted together at each end. The name arises from the

physical appearance of the winding, which resembles a squirrel cage.

**Torque-speed curve:** The value of torque in N·m produced by a motor as a function of the rotor speed.

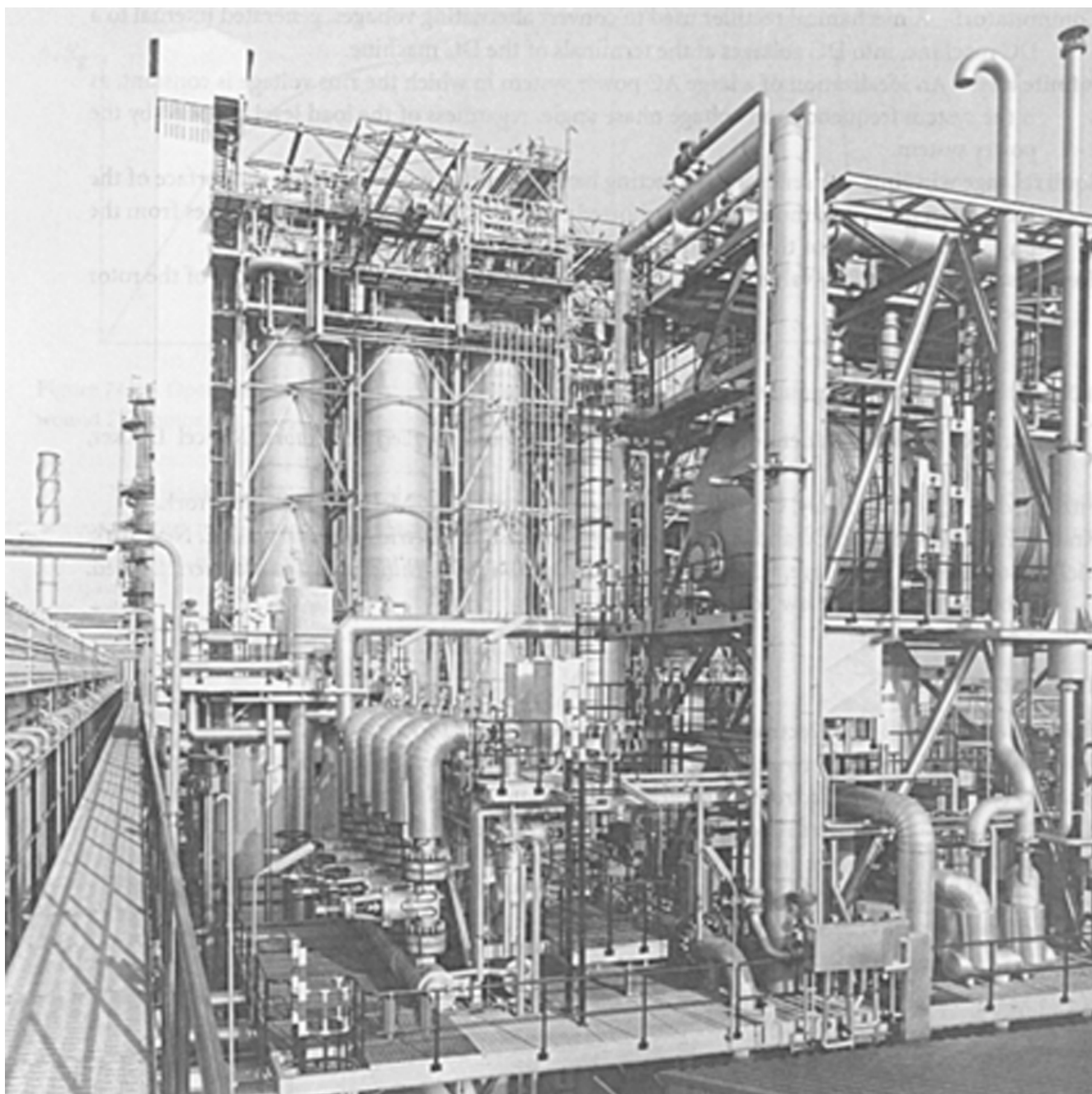
## References

- Engelman, R. H. and Middendorf, W. H. 1994. *Handbook of Electric Motors*. Marcel Dekker, New York.
- Fitzgerald, A. E. and Kingsley, C. *Electric Machinery*, 2nd ed. 1961. McGraw-Hill, New York.
- Krause, P. C., Wasynczyk, O., and Sudhoff, S. 1994. *Analysis of Electric Machinery*. IEEE, New York.
- McPherson, G. and Laramore, R. D. 1990. *An Introduction to Machines and Transformers*, 2nd ed. John Wiley & Sons, New York.

## Further Information

Institute of Electrical and Electronics Engineering (professional society)  
*Institute of Electrical and Electronics Engineers Transactions on Energy Conversion* (journal)  
*Institute of Electrical and Electronics Engineers Transactions on Industry Applications* (journal)  
Small Motors Manufacturers Association (professional society)  
*Institute of Electrical Engineering, Part B* (journal)  
*Electric Machines and Power Systems* (journal)  
*Electric Power Systems Research* (journal)

Falconer, J. L. "Kinetics and Reaction Engineering"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



The oil crises in 1973 and 1979 created a powerful incentive for producing the greatest amount of light products from crude oil. The Shell Pernis complex oil refinery located in the Netherlands was able to treat the heavy oil residue remaining after atmospheric distillation by redistilling it under high vacuum. By redistilling the residue, valuable products such as gasoline, gasoil, and lubricating oil distillates could be manufactured. However, even after this conversion method, a black, viscous mass still remained and to create a usable product from this material was a costly endeavor.

Shell laboratories made a start with the development of a better economic process called hydroconversion. This process is based on the fact that the vacuum residue contains much less hydrogen than the lighter products. The aim, therefore, is to reduce the size of the molecules and to increase the ration of hydrogen atoms to carbon atoms. This ratio can be increased either by removing some carbon from the molecules or by attaching hydrogen to the atoms with the aid of a catalyst. Shell chose the latter approach and called its process HYCON. For additional information on HYCON see page 811. (Photo courtesy of Shell Group.)

# X

## Kinetics and Reaction Engineering

---

**John L. Falconer**

*University of Colorado*

**75 Reaction Kinetics** *K. H. Lin*

Fundamentals • Analysis of Kinetic Data

**76 Chemical Reaction Engineering** *H. S. Fogler*

The Algorithm • Pressure Drop in Reactors • Multiple Reactions • Heat Effects • Summary

**77 The Scaleup of Chemical Reaction Systems from Laboratory to Plant** *J. B. Cropley*

General Considerations in the Rational Design of Chemical Reactors • Protocol for the Rational Design of Chemical Reactors

CHEMICAL REACTORS ARE THE MOST IMPORTANT PART of a chemical plant. In most cases, multiple reactions take place, multiple reactors are required, and catalysts are used to obtain sufficient rates and desired selectivities. Improvements in reaction rates and selectivities to desired products can have significant influences on other parts of the chemical plant such as the separations processes. Thus, design of the chemical reactor can control the economics of a plant even though the reactor is not the most expensive part. The chemical reactor also determines the amounts of waste products that form and thus the plant's effect on the environment. Because highly exothermic reactions are often carried out in chemical reactors on a large scale, the reactor is also the biggest safety hazard in the plant.

The first chapter of Section X is concerned with chemical kinetics and the analysis of kinetic data. The most important aspect of chemical kinetics is the rate at which a chemical reaction takes place and the selectivity to reaction products. Also of interest is how the rate and selectivity depend on concentrations, temperature, and other reaction conditions. Reaction rates and the products that form for a given set of reactants and reaction conditions cannot be predicted; such kinetic information must be measured. These measurements are particularly sensitive to temperature because most chemical reactions exhibit an exponential dependence on temperature. They also must be made in the absence of transport effects such as diffusion and mass transfer. Moreover, many large-scale chemical processes use catalysts, and the composition and preparation of the catalyst can have a large influence on both the rate of reaction and the product distribution. The basic methods for analyzing laboratory kinetic data to determine rate expressions are presented in this section, and the mechanisms by which reactions occur on a molecular scale are also discussed.

The second chapter discusses the design and analysis of chemical reactors. Reaction engineering involves determining how the type of reactor, its size, and its operating conditions affect production rates and distribution of products. Industrial processes almost always involve multiple reactions taking place simultaneously, and the desired product is often not the more favored

thermodynamically. Batch, semibatch, continuous stirred tank (CSTR), and plug flow reactors are discussed. These ideal reactor types are often used in combination and are used to model real reactors, which may deviate significantly from ideal behavior. Once kinetic data have been obtained in the laboratory, the kinetic rate expression, thermodynamic equilibrium, and reaction stoichiometry can be used in material balances (conservation of mass) and energy balances (first law of thermodynamics) to obtain equations that predict how a large-scale reactor will behave. Large-scale reactors must take account of pressure drop, nonideal flow patterns, and transport limitations.

The last chapter describes scale-up of a reaction system from the laboratory to a large-scale plant using laboratory data and correlations, and presents a protocol for rational design. This could be called scale-down since it starts from a potential type of commercial reactor, and then a laboratory system is designed to generate the necessary kinetic data. A statistically valid kinetic model is then developed from the data, and this is used with the material and energy balances to develop a model that simulates the reactor. This model of a reactor can then be used to develop an optimal design, which is then validated in a pilot plant. This chapter also discusses industrial reactor types, reactor mass balances, and design of laboratory reactors. In addition to the basic ideal reactors, fluidized beds, shell-and-tube packed beds, multiphase packed beds, and slurry reactors are discussed.



Lin, K. H. "Reaction Kinetics"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## 75.1 Fundamentals

Basic Terms and Equations • Rate Constant and Elementary Reactions • Complex (or Multiple) Reactions • Uncatalyzed Heterogeneous Reactions • Homogeneous and Heterogeneous Catalytic Reactions

## 75.2 Analysis of Kinetic Data

Data Acquisition • Evaluation of Reaction Mechanism • Development of Rate Equation • Determination of Rate Constant and Arrhenius Parameters

### K. H. Lin

*Oak Ridge National Laboratory*

This chapter presents a brief overview of reaction kinetics primarily for engineers who are not directly involved in the investigation of reaction kinetics or in the design of chemical reactors. For a comprehensive treatise on reaction kinetics, the reader should consult the references at the end of the chapter.

In contrast to the static and equilibrium concept of thermodynamics, reaction kinetics is concerned with dynamics of chemical changes. Thus, reaction kinetics is the science that investigates the rate of such chemical changes as influenced by various process parameters and attempts to understand the mechanism of the chemical changes. For any reaction system to change from an initial state to a final state, it must overcome an energy barrier. The presence of such an energy barrier is commonly manifested in the observed relationship between the reaction rate and temperature, which will be discussed in some detail in the section to follow.

## 75.1 Fundamentals

---

### Basic Terms and Equations

One of the key terms in reaction kinetics is the **rate of reaction**,  $r_A$ , the general definition of which is given by

$$r_A = \frac{1}{y} \frac{dN_A}{dt} \quad (75.1)$$

which expresses the rate  $r_A$  as the amount of a chemical component of interest being converted or produced per unit time per unit quantity of a reference variable  $y$  [e.g., volume of reacting mixture ( $V$ ) or of reactor ( $V_R$ ), mass ( $W$ ), surface area ( $S$ ), etc.]. It delineates the time

dependence of chemical changes involving component  $A$  as a derivative. In homogeneous fluid reactions,  $y$  is normally represented by  $V$  or  $V_R$ , whereas the mass ( $W$ ) or the surface area ( $S$ ) of the solid reactant may be taken as  $y$  in heterogeneous solid-fluid reactions. Here,  $N_A$  refers to the amount of component  $A$ , and  $t$  is time. By convention,  $r_A$  is negative when  $A$  is a reactant and positive when  $A$  is a product. Molal units are commonly used as the amount of  $N_A$ , but other units such as mass, radioactivity, pressure, and optical property are also used. When  $V$  remains constant (as in a liquid-phase batch reactor), Eq. (75.1) is simplified to

$$r_A = dC_A/dt \quad (75.2)$$

where  $C_A$  represents the concentration of component  $A$ .

## Rate Constant and Elementary Reactions

In general, the rate of reaction in terms of component  $i$  is a function of the concentration of all components participating in the reaction,  $C$ ; the temperature,  $T$ ; the pressure,  $P$ ; and other parameters,  $m$ :

$$r_i = \text{function } (C, T, P, m) \quad (75.3)$$

Thus, the rate expression for a simple irreversible reaction in terms of the reacting components may assume the following form:

$$-r_A = k(C_A)^{n_1}(C_B)^{n_2} \dots (C_i)^{n_i} \quad (75.4)$$

The proportionality constant,  $k$ , is the **rate constant** that is markedly influenced by the temperature and may be subject to the influence of pressure, pH, kinetic isotopes, and so on, and the presence of catalysts. The exponents  $n_1, n_2, \dots, n_i$  are *orders of reaction* with respect to individual reacting components  $A, B, \dots, i$ . The *overall order of reaction* refers to the sum of  $n_1, n_2, \dots, n_i$ , which does not have to be an integer and may be determined empirically.

The units and value of rate constant  $k$  vary with the units of  $C$ , the specific component that  $k$  refers to, and the reaction order. The effect of temperature on  $k$  was first described by Arrhenius through the following equation:

$$k = Ae^{-\frac{E}{RT}} \quad (75.5)$$

where  $A$ , termed the *frequency factor*, has the same units as  $k$ ,  $E$  is **activation energy**, and  $R$  is the gas law constant.  $E$  was considered by Arrhenius as the amount of energy that a reacting system must have in excess of the average initial energy level of reactants to enable the reaction to proceed.

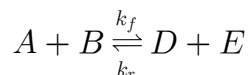
A *simple* or *elementary* reaction is one in which the order of reaction is identical to the *molecularity* (the number of molecules actually participating in the reaction). Under these circumstances, the chemical stoichiometric equation represents the true reaction mechanism, and

the rate equation may therefore be derived directly from the stoichiometric equation. Thus, for an elementary reaction  $n_1A + n_2B = n_3D$ , the rate equation in terms of disappearance of A would be  $-r_A = k(C_A)^{n_1}(C_B)^{n_2}$ . The values of  $n_1$  and  $n_2$  in this equation are positive integers.

## Complex (or Multiple) Reactions

A reaction that proceeds by a mechanism involving more than a single reaction path or step is termed a *complex reaction*. Unlike elementary reactions, the mechanisms of complex reactions differ considerably from their stoichiometric equations. Most industrially important reactions are complex reactions, the mechanisms of which can often be determined by assuming that the overall reaction consists of several elementary reaction steps. The resulting overall rate expression is then compared with the experimental data, and the procedure is repeated until a desired degree of agreement is obtained. Each of the elementary reaction steps may proceed *reversibly*, *concurrently*, or *consecutively*.

A *reversible reaction* is one in which conversion of reactants to products is incomplete at equilibrium because of an increasing influence of the reverse reaction as the forward reaction approaches equilibrium. For a reversible reaction of the type

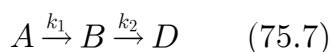


the net forward rate of reaction in terms of disappearance of A is

$$-r_A = k_f C_A C_B - k_r C_D C_E \quad (75.6)$$

which assumes both forward and reverse reactions to be elementary.

A simple example of *consecutive reaction* is illustrated by



Again, assuming an elementary reaction for each reaction step, the following rate equations result:

$$-r_A = k_1 C_A \quad (75.8)$$

$$r_D = k_2 C_B \quad (75.9)$$

$$r_B = k_1 C_A - k_2 C_B \quad (75.10)$$

*Parallel or simultaneous reactions* are those involving one or more reactants undergoing reactions of more than one scheme, as in



The rate equations may assume the following forms:

$$-r_A = k_1(C_A)^a + k_2(C_A)^b \quad (75.11)$$

$$r_B = k_1(C_A)^a \quad (75.12)$$

$$r_D = k_2(C_A)^b \quad (75.13)$$

Under a constant-volume condition,  $r_B = dC_B/dt$  and  $r_D = dC_D/dt$ . Therefore, the relative rate of formation of  $B$  and  $D$  is derived from Eqs. (75.12) and (75.13) as

$$dC_B/dC_D = (k_1/k_2)(C_A)^{a-b} \quad (75.14)$$

The ratio  $dC_B/dC_D$  is termed the *point selectivity*, which is the ratio of the *rate* of formation of product  $B$  to the *rate* for product  $D$ . The *overall* (or *integrated*) *selectivity* is obtained by integration of this ratio, and it represents the ratio of the overall *amount* of product  $B$  to that of product  $D$ . Equation (75.14) implies that the relative rate of formation of  $B$  is proportional to  $C_A$  when  $a > b$ , whereas it is inversely proportional to  $C_A$  when  $a < b$ .

In a chemical process involving complex reactions that consist of several reaction steps, one or more steps may represent major factors in governing the overall rate of reaction. Such reaction steps are termed the **rate-controlling steps**. The rate-controlling reaction steps are observed in homogeneous complex reactions and heterogeneous reactions.

## Uncatalyzed Heterogeneous Reactions

Heterogeneous reactions involve more than one phase (e.g., gas-liquid, gas-solid, liquid-solid, and gas-liquid-solid) and are generally more complicated than homogeneous reactions due to interaction between physical and chemical processes; that is, reactants in one phase have to be transported (physical process) to the other phase, containing other reactants where the reactions take place.

In a gas-solid reaction, for example, the reaction may proceed in several steps, as follows:

1. Reactants in the gas phase diffuse to the gas-solid interface.
2. When there is a layer of solid product and/or inert material at the interface (e.g., ash), the reactants from the gas phase would have to diffuse through this layer before they can reach the unreacted solid core containing other reactants.
3. The chemical reaction takes place between the reactants from the gas phase and those in the unreacted solid core.
4. The reaction products diffuse within the solid phase and/or diffuse out of the solid phase into the bulk of gas phase.

The step that controls the overall reaction rate will be determined by the nature of the phases and specific reactions involved and by process conditions. Thus, the overall reaction rate is subject to the influence of parameters that affect both the physical and chemical processes, including (a)

patterns of phase contact, (b) the reactor geometry, (c) fluid dynamic factors (e.g., velocity and degree of turbulence), (d) interfacial surface area, (e) mass transfer factors, (f) chemical kinetics of reactions involved, and (g) process parameters (e.g., temperature and pressure). Some of these parameters may interact with one another. For example, in a reaction involving two distinct fluid phases (e.g., gas-liquid or liquid-liquid), parameters (d) and (e) would be affected by parameter (c).

The overall reaction rate expression of a heterogeneous reaction is fairly complex, since it considers all of these parameters. Further, the form of rate equation varies with the type of heterogeneous reaction system and with the nature of the controlling step. Some examples of the industrially significant uncatalyzed heterogeneous reactions are given in [Table 75.1](#).

**Table 75.1** Examples of Uncatalyzed Heterogeneous Reactions

<p><i>Gas-liquid reactions</i></p> <ul style="list-style-type: none"> <li>• Production of ammonium nitrate by reaction between ammonia gas and nitric acid</li> <li>• Hydrogenation of vegetable oil with hydrogen gas</li> <li>• Production of nitric acid by absorption of nitric oxide in water</li> </ul> <p><i>Gas-solid reactions</i></p> <ul style="list-style-type: none"> <li>• Gasification of coal</li> <li>• Production of hydrogen gas by reaction of steam with iron</li> <li>• Production of volatile uranium chloride by reaction of uranium oxide with chlorine gas</li> </ul> <p><i>Liquid-liquid reactions</i></p> <ul style="list-style-type: none"> <li>• Aqueous sulfuric acid treatment of petroleum liquid</li> <li>• Nitration of organic solvents with aqueous nitric acid</li> <li>• Production of soaps by reaction of aqueous alkalies and fatty acids</li> </ul>	<p><i>Liquid-solid reactions</i></p> <ul style="list-style-type: none"> <li>• Reaction of aqueous sulfuric acid with phosphate rock</li> <li>• Ion exchange process</li> <li>• Recovery of uranium by leaching of uranium ores with sulfuric acid</li> </ul> <p><i>Solid-solid reactions</i></p> <ul style="list-style-type: none"> <li>• Production of calcium carbide by reaction of carbon with lime</li> <li>• Production of Portland cement by reaction of limestone with clay</li> <li>• Production of glass by melting a mixture of calcium carbonate, sodium carbonate, and silica</li> </ul> <p><i>Gas-liquid-solid reaction</i></p> <ul style="list-style-type: none"> <li>• Liquefaction of coal by reaction of hydrogen with coal-oil slurry</li> </ul>
--	---

## Homogeneous and Heterogeneous Catalytic Reactions

A catalytic reaction is a chemical reaction the rate of which is modified in the presence of a catalyst. A **catalyst** is a substance that may or may not change chemically during the reaction and is regenerated at the end of the reaction. The catalytic reaction proceeds appreciably faster than does an uncatalyzed reaction, presumably because an intermediate compound, formed between the catalyst and some reactants, reacts with other reactants by a mechanism that requires a lower activation energy to form desired products. In *homogeneous catalysis* the catalyst forms a homogeneous phase with the reaction mixture. In *heterogeneous catalysis*, however, the catalyst is present in a phase different from that of the reaction mixture.

### Homogeneous Catalysis

Most homogeneous catalysis takes place in the liquid phase. Perhaps the most widely studied type of liquid-phase catalysis is the acid-base catalysis that exerts influence on the rates of many important organic reactions, including (a) esterification of alcohols, (b) hydrolysis of esters, and (c) inversion of sugars. One of the industrially important gas-phase catalytic reactions is the oxidation of  $\text{SO}_2$  to  $\text{SO}_3$  in the production of sulfuric acid, catalyzed by nitric oxide in the lead chamber.

### Solid-Catalyzed Reaction

This reaction, the most common type of heterogeneous catalysis, finds extensive applications in many important industrial processes that produce inorganic and organic chemicals. Well-known examples of such chemicals include  $\text{HNO}_3$ ,  $\text{HCl}$ , ammonia, aniline, butadiene, ethanol, formaldehyde, methanol, organic polymers, and petrochemicals. The generally accepted, simplified mechanism of solid-catalyzed fluid (gas or liquid) phase reactions is outlined as follows:

1. Reactants diffuse from the main body of the fluid phase to the exterior surface of catalyst pellets, and subsequently into catalyst pores.
2. Reactants are adsorbed onto both the catalyst exterior and pore surfaces.
3. Products are formed from interaction of the reactants on the surfaces (catalyst exterior and pore).
4. Products thus formed are desorbed (or released) from the surfaces; those formed on the pore surface are released to the fluid phase within the pores and then diffuse out of pores to the exterior surface of the catalyst pellet.
5. Products then diffuse from the exterior surfaces into the bulk of the fluid phase.

The relative importance of these steps in influencing the overall reaction rate depends upon a variety of factors, among which are thermal factors, fluid dynamic factors, properties of the catalyst, and diffusion characteristics of the reactants and products. Besides the process steps described earlier, there are various deactivation processes that cause loss of catalytic efficiency, such as fouling and poisoning.

## 75.2 Analysis of Kinetic Data

---

### Data Acquisition

Because chemical reactions involved in most industrial processes are complex, development of the database for the design of the chemical reactor facility can be quite time consuming. Accordingly, selection of experimental methods and equipment for acquisition of kinetic data is crucial in determining the development cost as well as the accuracy and reliability of the data obtained. In essence, the selection process is concerned with both the methods and the equipment to conduct the reactions and to monitor the progress of the reactions. The type of equipment to be used is generally determined by the experimental method for data acquisition. Acquisition of the kinetic

data for homogeneous reactions is frequently carried out using a batch reactor because of its relatively simple design and versatility. In heterogeneous reactions a flow reactor is often utilized for the data acquisition. For detailed discussion on various methods and equipment for obtaining the experimental kinetic data, the reader is referred to the references at the end of the chapter.

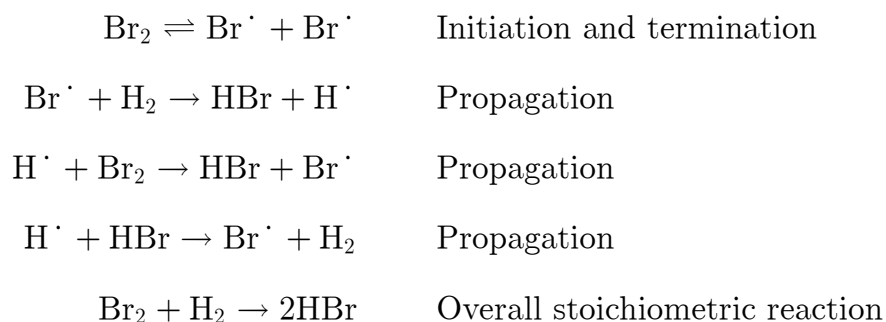
## Evaluation of Reaction Mechanism

Understanding of the reaction mechanism is important in the selection and design of an industrial reactor for a specific reaction. Full elucidation of the mechanism, however, is not always possible, in which case derivation of the rate equation may have to resort to the empirical method (see the following subsection). No simple standardized method is available for evaluation of the reaction mechanism. Nevertheless, a trial-and-error method is often used based on the experimental kinetic data acquired—including analysis of the reaction mixture to determine the distribution of the residual reactants, intermediates, and final products—following these steps:

1. Assume a simple elementary reaction mechanism and a corresponding stoichiometry and derive a rate equation from the assumed mechanism.
2. Evaluate the experimental data based on the proposed reaction mechanism and the corresponding rate equation using the integral method (described in the section to follow) first since it is relatively easy to use.
3. If the experimental data do not agree with the proposed mechanism, possibly suggesting a nonelementary reaction, propose a new mechanism that consists of several elementary reaction steps with formation of intermediate compounds.
4. Develop rate equations for individual elementary reaction steps and combine the individual rate equations to represent the overall rate expression.
5. If the experimental kinetic data do not fit into the rate equation developed above, assume an alternate mechanism, and repeat step 4. Continue this process until a desired degree of agreement is reached between the experimental data and the rate expression.
6. Evaluation of the reaction mechanism may also be accomplished using the differential method (see the next subsection), especially for complicated reactions. The method of approach is similar to that using the integral method, but it requires more accurate and extensive experimental data.

An example of a nonelementary complex reaction consisting of several elementary reaction steps is the formation of hydrogen bromide from hydrogen and bromine:





Based on the above reaction mechanism, the following rate equation has been derived:

$$r_{\text{HBr}} = \frac{k_1 C_{\text{H}_2} C_{\text{Br}_2}^{0.5}}{1 + k_2 (C_{\text{HBr}} / C_{\text{Br}_2})} \quad (75.15)$$

## Development of Rate Equation

For reactions with simple mechanism under isothermal conditions, either the integral method or the differential method (discussed in a later section) may be used in the derivation of rate equations. It is assumed that the data representing the extent of an isothermal reaction are available in terms of the time variation of a selected component A .

### Integral Method

An elementary reaction mechanism is first assumed; for example, a second-order isothermal homogeneous reaction under *constant-volume* conditions,  $A + B \rightarrow C + D$ , results in the rate equation of the form

$$-r_A = \frac{dC_A}{dt} = k C_A C_B \quad (75.16)$$

and the integrated rate equation becomes

$$kt = \frac{1}{C_{B_0} - C_{A_0}} \ln \frac{C_{A_0} (C_A + C_{B_0} - C_{A_0})}{C_A C_{B_0}} \quad (75.17)$$

where  $C_{A_0}$  and  $C_{B_0}$  represent the initial concentrations of reactants A and B , respectively. One way to confirm the proposed reaction mechanism is to compute values of  $k$  at various values of  $C_A$  and  $t$  . If the values of  $k$  remain nearly constant, the proposed mechanism is accepted. Otherwise, another mechanism is assumed and the process is repeated until a desired degree of agreement is reached.

When the volume of the reaction mixture varies with the extent of reaction, the rate equation becomes more complicated. In this case, if it is assumed that the volume varies linearly with the

extent of reaction, derivation of the rate equation could be simplified. Using the fractional conversion  $x_A$  to replace  $C_A$  as the variable, the volume  $V$  is expressed as

$$V = V_0(1 + f_A x_A) \quad (75.18)$$

Here,  $V_0$  represents the initial volume of the reaction mixture and  $f_A$  is the fractional change in  $V$  between no conversion and complete conversion with respect to reactant  $A$ , as defined by

$$f_A = (V_{x_A=1} - V_{x_A=0})/V_{x_A=0} \quad (75.19)$$

Thus, for a *variable-volume* reaction, the reaction rate defined by Eq. (75.1) assumes the following form,

$$-r_A = -\frac{1}{V} \frac{dN_A}{dt} = -\frac{1}{V_0(1 + f_A x_A)} \frac{d}{dt} N_{A_0}(1 - x_A) \quad (75.20)$$

which simplifies to

$$-r_A = \frac{C_{A_0}}{1 + f_A x_A} \frac{dx_A}{dt} \quad (75.21)$$

and the integrated rate equation is

$$t = C_{A_0} \int_0^{x_A} \frac{dx_A}{(1 + f_A x_A)(-r_A)} \quad (75.22)$$

where  $-r_A$  stands for the rate expression for the assumed reaction mechanism to be evaluated. For example, for a first-order homogeneous reaction,

$$-r_A = kC_A = kC_{A_0} \frac{1 - x_A}{1 + f_A x_A} \quad (75.23)$$

Introducing this expression for  $-r_A$  into Eq. (75.22),

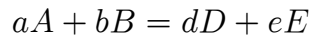
$$kt = -\ln(1 - x_A) \quad (75.24)$$

The steps to be taken to evaluate the assumed reaction mechanism for the variable-volume case are identical to the constant-volume case.

### Differential Method

This approach is based on the direct application of the differential rate equation in the analysis of the experimental kinetic data to determine the reaction mechanism. The method requires more accurate and extensive experimental data than the integral method. The basic principle of the method is illustrated by an example of assumed  $n$  th-order reaction under the isothermal and

constant-volume condition:



With initially equimolal concentrations of  $A$  and  $B$  (i.e.,  $C_A = C_B$ ), the rate equation may take the following form,

$$-r_A = kC_A^a C_B^b = kC_A^{a+b} = kC_A^n \quad (75.25)$$

which is rearranged into

$$\log(-r_A) = \log(-dC_A/dt) = \log k + n \log C_A \quad (75.26)$$

The assumption of the  $n$  th-order reaction mechanism is confirmed if a log-log plot of  $(dC_A/dt)$  versus  $C_A$  results in a straight line.

The next step is to evaluate the values of the rate constant  $k$  and the overall order of reaction  $n$  from the plot. With the known values of  $k$  and  $n$ , the reaction orders with respect to  $A$  and  $B$  can be determined by the method that follows. Rearranging Eq. (75.25),

$$-r_A = kC_A^a C_B^{n-a} = kC_B^n (C_A/C_B)^a \quad (75.27)$$

On further rearrangement,

$$\log \left( -\frac{r_A}{kC_B^n} \right) = a \log(C_A/C_B) \quad (75.28)$$

A log-log plot of Eq. (75.28) yields the reaction order  $a$  with respect to  $A$ , whereas the reaction order  $b$  is obtained as the difference between  $n$  and  $a$ .

### Empirical Method

This method, which finds uses when the reaction mechanism appears to be complex, is often based on a mathematical approach using the curve-fitting procedure. The method involves a trial-and-error technique to fit the experimental data to a relatively simple form of the empirical equation, including (a) linear form,  $y = a + bx$ , (b) semilogarithmic form,  $y = ae^{bx}$ , (c) logarithmic form,  $y = c + ax^n$ , and so forth. The initial step usually consists of plotting the experimental data on graph papers of different coordinates that may produce a straight line. Upon selection of a proper form of the empirical equation, the constants in the empirical equation are determined either by the graphic means or by the analytical technique using the method of averages or the method of least squares. Computer software is available for performing the curve-fitting procedure.

## Determination of Rate Constant and Arrhenius Parameters

The rate constant  $k$  can be obtained by using either a differential form [e.g., Eq. (75.16)] or an integrated form [e.g., Eq. (75.17)] of the rate equation. It is an average of values calculated at various experimental kinetic data points (i.e., reactant concentrations at various reaction times).

*Arrhenius parameters* consist of the activation energy  $E$  and the frequency factor  $A$ . The value of  $E$  may be calculated from the rate constants at two distinct but adjacent temperatures,  $T_1$  and  $T_2$ , as follows:

$$k_1 = Ae^{-E/RT_1} ; \quad k_2 = Ae^{-E/RT_2}$$

Combining the above two equations,

$$E = R \frac{\ln(k_2/k_1)}{1/T_1 - 1/T_2} \quad (75.29)$$

The value of  $A$  is calculated from one of the Arrhenius equations shown above.

## Defining Terms

**Activation energy:** A parameter associated with the Arrhenius equation and considered by Arrhenius as the energy in excess of the average energy level of reactants required to enable the reaction to proceed.

**Catalyst:** A substance that accelerates the reaction, presumably by making available a reaction path that requires a lower activation energy. The catalyst may or may not change chemically during the reaction and is regenerated at the end of the reaction.

**Rate constant:** A proportionality constant in the rate equation. The rate constant is markedly influenced by temperature and, to lesser degree, by pressure and the presence of catalysts. The units and value of the rate constant depend on the specific chemical component to which it refers, the units for concentration (or other quantity) of the component, and the reaction order.

**Rate-controlling step:** The slow steps that tend to control the overall rate of reaction. In a complex reaction consisting of several chemical reaction steps (and physical process steps in heterogeneous reaction), the reaction rate (and physical process rate) of one or more steps may be much slower than other steps.

**Rate equation:** A functional expression describing the relationship between the rate of reaction and the amounts (e.g., concentrations) of selected chemical components participating in the reaction at any time under isothermal condition.

**Rate of reaction:** The amount of a chemical component of concern being converted or produced per unit time per unit quantity of a reference variable. Examples of the reference variable include the volume of reacting mixture, the reactor volume, the mass of solid (solid-fluid reaction), and the surface area of solid.

## References

- Carberry, J. J. 1976. *Chemical and Catalytic Reaction Engineering*. McGraw-Hill, New York.
- Connors, K. A. 1990. *Chemical Kinetics<sup>3</sup>/<sub>4</sub>The Study of Reaction Rates in Solution*. VCH, New York.
- Kataakis, D. and Gordon, G. 1987. *Mechanisms of Inorganic Reactions*. John Wiley & Sons, New York.
- Lin, K. H. 1984. Reaction kinetics, reactor design (section 4). In *Perry's Chemical Engineers' Handbook* 4.52. McGraw-Hill, New York.
- Moore, J. W. and Pearson, R. G. 1981. *Kinetics and Mechanism*, 3rd ed. John Wiley & Sons, New York.

## Further Information

The following professional journals provide good sources for examples of basic and applied reaction kinetic studies on specific reactions of industrial importance:

- AIChE Journal*. Published monthly by the American Institute of Chemical Engineers, New York, NY.
- Chem. Eng. Sci.* Published semimonthly by Elsevier Science, Oxford, U.K.
- Ind. Eng. Chem. Res.* Published monthly by the American Chemical Society, Washington, D.C.
- J. Am. Chem. Soc.* Published biweekly by the American Chemical Society, Washington, D.C.
- J. Catal.* Published monthly by Academic Press, Orlando, FL.
- J. Chem. Soc.<sup>3</sup>/<sub>4</sub>Faraday Trans.* Published semimonthly by The Royal Society of Chemistry, Cambridge, U.K.
- Trans. Inst. Chem. Eng. (London)<sup>3</sup>/<sub>4</sub>Chem. Eng. and Design*. Published bimonthly by the Institute of Chemical Engineers, Basinstoke, U.K.

Fogler, H. S. "Chemical Reaction Engineering"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Chemical Reaction Engineering

## 76.1 The Algorithm

Mole Balances • Rate Laws • Stoichiometry

## 76.2 Pressure Drop in Reactors

## 76.3 Multiple Reactions

## 76.4 Heat Effects

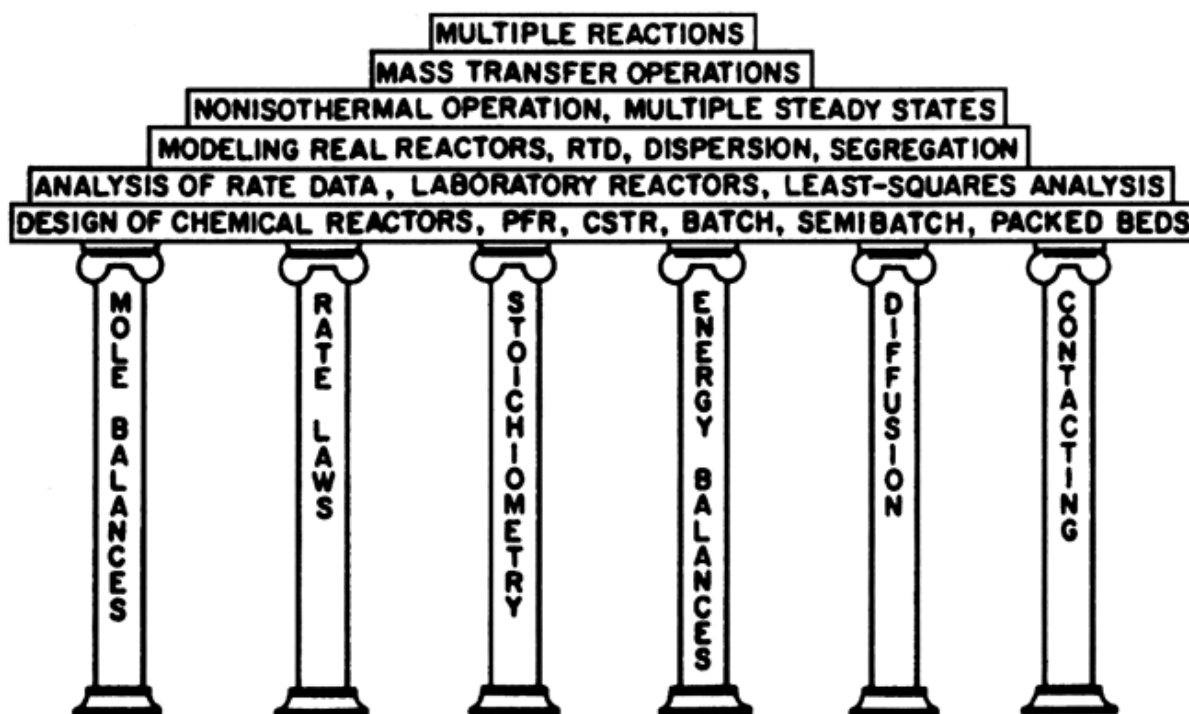
## 76.5 Summary

### H. Scott Fogler

University of Michigan

Chemical reaction engineering (CRE) sets chemical engineers apart from other engineers. Students and professionals can easily learn the elements of CRE because it has a very logical structure. The six basic pillars that hold up what you might call the "temple" of chemical reaction engineering are shown in Fig. 76.1 .

**Figure 76.1** Pillars of the temple of chemical reaction engineering. (Source: Fogler, H. S. 1992. *The Elements of Chemical Reaction Engineering*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.)



The pillar structure shown in Fig. 76.1 allows one to develop a few basic concepts and then to arrange the parameters (equations) associated with each concept in a variety of ways. Without such a structure, one is faced with the possibility of choosing or perhaps memorizing the correct equation from a multitude of equations that can arise for a variety of reactions, reactors, and sets of conditions. This chapter shall focus on five types of chemical reactors commonly used in industry: **batch**, **semibatch**, **CSTR**, **plug flow**, and **packed bed reactors**. Table 76.1 describes each of these reactors.

**Table 76.1** Comparison of Five Types of Chemical Reactors

Type of Reactor	Characteristics	Usage	Advantages	Disadvantages
Batch	<ul style="list-style-type: none"> <li>Reactor is charged (filled) via two holes in the top of the tank; while reaction is carried out, nothing else is put in or taken out until reaction is done; tank easily cooled or heated by jacket</li> </ul>	<ul style="list-style-type: none"> <li>Small-scale production</li> <li>Intermediate or one-shot productions</li> <li>Pharmaceuticals</li> <li>Fermentations</li> </ul>	<ul style="list-style-type: none"> <li>High conversion per unit volume for one pass</li> <li>Same reactor can be used to produce one product one time and a different product the next</li> </ul>	<ul style="list-style-type: none"> <li>High operating cost (labor)</li> <li>Product quality more variable than with continuous operation</li> </ul>
Semibatch	<ul style="list-style-type: none"> <li>Either one reactant is charged and the other is fed continuously (at small concentrations) or else one of the products can be removed continuously (to avoid side reactions)</li> </ul>	<ul style="list-style-type: none"> <li>Small-scale production</li> <li>Competing reactions</li> </ul>	<ul style="list-style-type: none"> <li>Good selectivity; feed can be controlled so as to minimize side runs.</li> </ul>	<ul style="list-style-type: none"> <li>High operating labor cost</li> <li>Product quality more variable than with continuous operation</li> </ul>
Continuously stirred tank reactor (CSTR)	<ul style="list-style-type: none"> <li>Run at steady state with continuous flow of reactants and products; the feed assumes a uniform composition throughout the reactor, exit stream has the same composition as in the tank</li> </ul>	<ul style="list-style-type: none"> <li>When agitation is required</li> <li>Series configuration for different concentration streams</li> </ul>	<ul style="list-style-type: none"> <li>Continuous operation</li> <li>Good temperature control</li> <li>Good control</li> <li>Simplicity of construction</li> <li>Low operating (labor) cost</li> </ul>	<ul style="list-style-type: none"> <li>Lowest conversion per unit volume</li> <li>Bypassing and channeling possible with poor agitation</li> </ul>
Plug flow reactor (PFR)	<ul style="list-style-type: none"> <li>Arranged as one long reactor or many short reactors in a tube bank; no radial variation in reaction rate (concentration); concentration changes with length down the reactor</li> </ul>	<ul style="list-style-type: none"> <li>Large-scale production</li> <li>Homogeneous reactions</li> <li>Heterogeneous reactions</li> <li>Continuous production</li> <li>High temperature</li> </ul>	<ul style="list-style-type: none"> <li>Highest conversion per unit volume</li> <li>Low operating labor cost</li> <li>Continuous operation</li> <li>Good heat transfer</li> </ul>	<ul style="list-style-type: none"> <li>Undesired thermal gradients may exist</li> <li>Poor temperature control</li> <li>Shutdown, cleaning may be expensive</li> </ul>
Tubular packed bed reactor (PBR)	<ul style="list-style-type: none"> <li>Tubular reactor that is packed with solid catalyst particles</li> </ul>	<ul style="list-style-type: none"> <li>Used primarily in heterogeneous gas phase reactions with a catalyst</li> </ul>	<ul style="list-style-type: none"> <li>Highest conversion per unit mass of catalyst</li> <li>Low operating cost</li> <li>Continuous operation</li> </ul>	<ul style="list-style-type: none"> <li>Undesired thermal gradients may exist</li> <li>Poor temperature control</li> <li>Channeling may occur</li> </ul>

By using an algorithm to formulate CRE problems, we can formulate and solve CRE problems in a very logical manner. Step 1 in the CRE algorithm is to begin by choosing the mole balance for one of the five types of reactors shown. In step 2 we choose the rate law and in step 3 we specify whether the reaction is gas or liquid phase. Finally, in step 4 we combine steps 1, 2, and 3 and obtain an analytical solution or solve the equations using an *ordinary differential equation (ODE) solver* [Sacham and Cutlip, 1988].

## 76.1 The Algorithm

We now address each of the individual steps in the algorithm to design isothermal reactors: (1) mole balances, (2) rate laws, (3) stoichiometry, and (4) combine.

### Mole Balances

The general mole balance equation (GBE) on species  $j$  in a system volume  $V$  is:



$$\begin{bmatrix} \text{Molar flow} \\ \text{rate} \\ \text{IN} \end{bmatrix} - \begin{bmatrix} \text{Molar flow} \\ \text{rate} \\ \text{OUT} \end{bmatrix} + \begin{bmatrix} \text{Molar rate} \\ \text{of} \\ \text{GENERATION} \end{bmatrix} = \begin{bmatrix} \text{Molar rate} \\ \text{of} \\ \text{ACCUMULATION} \end{bmatrix}$$

$$F_{j0} - F_j + \int_0^v r_j dV = \frac{dN_j}{dt} \quad (76.1)$$

We now make use of the definition of conversion,  $X$ , with respect to the limiting reactant, which we shall call species  $A$ ,

$$\begin{array}{cc} \text{Batch} & \text{Flow} \\ X = (N_{A0} - N_A)/N_{A0} & X = (F_{A0} - F_A)/F_{A0} \end{array}$$

and apply the GME to each of the following reactors: batch, continuous stirred tank reactors (CSTR), plug flow reactor (PFR), and packed bed reactor (PBR). The CSTR, PFR, and PBR are all operated at steady state (i.e.,  $dN_j/dt = 0$ ) and it is assumed that the PBR and PFR are in plug flow (no radial gradients or dispersion) and that the contents of the CSTR are well mixed. There is no in-flow or out-flow ( $F_{j0} = F_j = 0$ ) in the batch reactor. When these conditions and the definition of conversion are applied to the general mole balance, the design equations (76.2 to 76.8) in [Table 76.2](#) result.

**Table 76.2** Design Equations

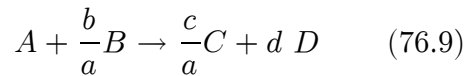
Reactor	Differential	Algebraic	Integral
Batch	$N_{A0} \frac{dX}{dt} = -r_A V \quad (76.2)$		$t = N_{A0} \int_0^x \frac{dX}{-r_A V} \quad (76.3)$
CSTR		$V = \frac{F_{A0} X}{-r_A} \quad (76.4)$	
PFR	$F_{A0} \frac{dX}{dV} = -r_A \quad (76.5)$		$V = F_{A0} \int_0^x \frac{dX}{-r_A} \quad (76.6)$
PBR	$F_{A0} \frac{dX}{dW} = -r'_A \quad (76.7)$		$W = F_{A0} \int_0^x \frac{dX}{-r'_A} \quad (76.8)$

In order to evaluate the design equations given in [Table 76.2](#) we must determine the form of the rate of formation,  $r_A$ . We do this with the aid of a rate law.

## Rate Laws

The power law model is one of the most commonly used forms for the rate law. It expresses the rate of reaction as a function of the concentrations of the species involved in the reaction.

For the irreversible reaction in which  $A$  is the limiting reactant,



the rate law is

$$-r_A = kC_A^\alpha C_B^\beta \quad (76.10)$$

We say the reaction is  $\alpha$  order in  $A$ ,  $\beta$  order in  $B$ , and overall order  $= \alpha + \beta$ . For example, if the reaction  $A + B \rightarrow C + D$  is said to be second order in  $A$  and first order in  $B$  and overall third order, then the rate law is

$$-r_A = kC_A^2 C_B \quad (76.11)$$

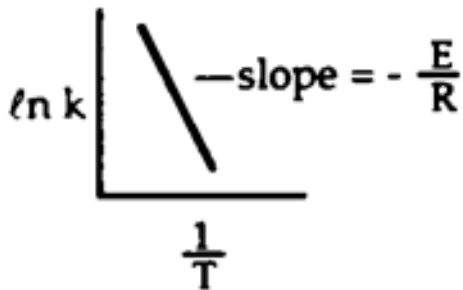
The temperature dependence of specific reaction rate,  $k$ , is given by the Arrhenius equation,

$$k = Ae^{-E/RT} \quad (76.12)$$

where  $A$  is the frequency factor and  $E$  the activation energy. Taking the natural log of both sides of Eq. (76.12),

$$\ln k = \ln A - \frac{E}{R} \left( \frac{1}{T} \right) \quad (76.13)$$

we see the slope of a plot of  $\ln k$  versus  $(1/T)$  will be a straight line equal to  $(-E/R)$ .



The specific reaction rate at temperature  $T$  is commonly written in terms of the specific reaction rate,  $k_1$ , at a reference temperature  $T_1$  and the activation energy  $E$ . That is,

$$k = k_1(T_1) \exp \left[ \frac{E}{R} \left( \frac{1}{T_1} - \frac{1}{T} \right) \right] \quad (76.14)$$

**Example.** The following reaction is carried out in a constant volume batch reactor:



Determine the appropriate linearized concentration-time plots for zero-, first-, and second-order reactions.

**Solution.** Use the algorithm to determine the concentration of A as a function of time.

$$\text{Mole balance: } \frac{dN_A}{dt} = r_A V \quad (76.15)$$

$$\text{Rate law: } -r_A = kC_A^\alpha \quad (76.16)$$

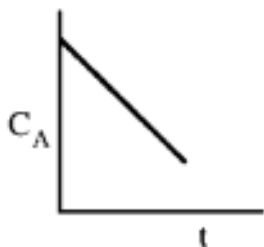
$$\text{Stoichiometry: } V = V_0 \quad C_A = N_A/V_0 \quad (76.17)$$

$$\text{Combine: } -\frac{dC_A}{dt} = kC_A^\alpha \quad (76.18)$$

Solving Eq. (76.18) for a first-, second-, and third-order rate law, we can arrive at the following linearized concentration-time plots.

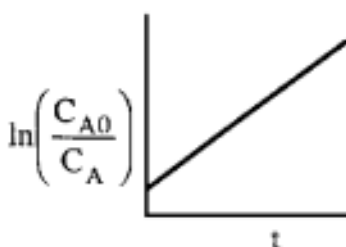
**Zero order,  $\alpha = 0$**

$$C_A = C_{A0} - kt$$



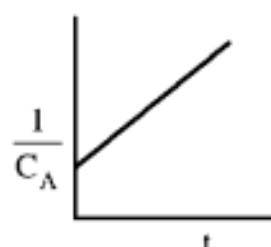
**First Order,  $\alpha = 1$**

$$\ln C_{A0}/C_A = kt$$



**Second Order,  $\alpha = 2$**

$$\frac{1}{C_A} - \frac{1}{C_{A0}} = kt$$



For reversible reactions at equilibrium the rate law must reduce to a thermodynamically consistent equation for the equilibrium constant.

## Stoichiometry

Now that we have the rate law as a function of concentration (i.e.,  $-r_A = kC_A^\alpha C_B^\beta$ ), we need to express the concentrations of the reacting species as functions of conversion in order to evaluate any one of the reactor design equations.

### Concentration

We start by defining concentration for a flow system and a batch system. For a flow system,

$$C_i = \frac{F_i}{v} \quad (76.19)$$

where  $v$  is the volumetric flow rate. For a batch system,

$$C_i = \frac{N_i}{V} \quad (76.20)$$

The next step is to express  $N_i$  and  $F_i$  as a function of conversion using a stoichiometric table.

### The Stoichiometry Table

Using our definition of conversion we can construct the following stoichiometric table.

Stoichiometry	$A + \frac{b}{a}B \rightarrow \frac{c}{a}C + \frac{d}{a}D$			
Batch systems Species	Symbol	Initial	Change	Remaining
$A$	$A$	$N_{A0}$	$-N_{A0}X$	$N_A = N_{A0}(1 - X)$
$B$	$B$	$N_{A0}\Theta_B$	$-\frac{b}{a}N_{A0}X$	$N_B = N_{A0}\left(\Theta_B - \frac{b}{a}X\right)$
$C$	$C$	$N_{A0}\Theta_C$	$+\frac{c}{a}N_{A0}X$	$N_C = N_{A0}\left(\Theta_C + \frac{c}{a}X\right)$
$D$	$D$	$N_{A0}\Theta_D$	$+\frac{d}{a}N_{A0}X$	$N_D = N_{A0}\left(\Theta_D + \frac{d}{a}X\right)$
Inert	$I$	$N_{A0}\Theta_I$	—	$N_I = N_{A0}\Theta_I$
		$N_{T0}$		$N_T = N_{T0} + \delta N_{A0}X$
where $\delta = \frac{d}{a} + \frac{c}{a} - \frac{b}{a} - 1$ , $\varepsilon = y_{A0}\delta$ , and $\Theta_i = \frac{N_{i0}}{N_{A0}} = \frac{y_{i0}}{y_{A0}} = \frac{C_{i0}}{C_{A0}}$				

For flow systems the number of moles of species  $i$ ,  $N_i$  in this table are simply replaced by the molar flow rates of species  $i$ ,  $F_i$ .

### Expressing Concentration as a Function of Conversion in Batch System

Constant volume batch:  $V = V_0$ .

$$C_B = \frac{N_B}{V} = \frac{N_B}{V_0} = \frac{N_{A0}}{V_0} \left( \Theta_B - \frac{b}{a}X \right)$$

$$C_B = C_{A0} \left( \Theta_B - \frac{b}{a}X \right) \quad (76.21)$$

### Expressing Concentration as a Function of Conversion in Flow System

For flow systems, the stoichiometric table is the same, except replace  $N_i$  by  $F_i$ . Because there is hardly ever a volume change with reaction, the concentration of  $A$  in a *liquid* flow system is as follows. For liquid systems,

$$C_A = \frac{F_A}{v_0} = \frac{F_{A0}}{v_0}(1 - X) = C_{A0}(1 - X) \quad (76.22)$$

For gas systems,

$$C_A = \frac{F_A}{v} \quad (76.23)$$

In ideal gas systems the gas volumetric flow rate,  $v$ , can change with conversion, temperature, and pressure according to the following equation:

$$v = v_0 \left( \frac{F_T}{F_{T0}} \right) \frac{P_0}{P} \frac{T}{T_0} \quad (76.24)$$

Taking the ratio of  $F_T/F_{T0}$  and then using the stoichiometric table, we arrive at the following equation for the volumetric flow rate at any point in the reactor.

$$v = v_0(1 + \varepsilon X) \frac{P_0}{P} \frac{T}{T_0} \quad (76.25)$$

Substituting this result and Eq. (76.22) into Eq. (76.23) gives

$$C_A = \frac{C_{A0}(1 - X)}{(1 + \varepsilon X)} \frac{P}{P_0} \frac{T_0}{T} \quad (76.26)$$

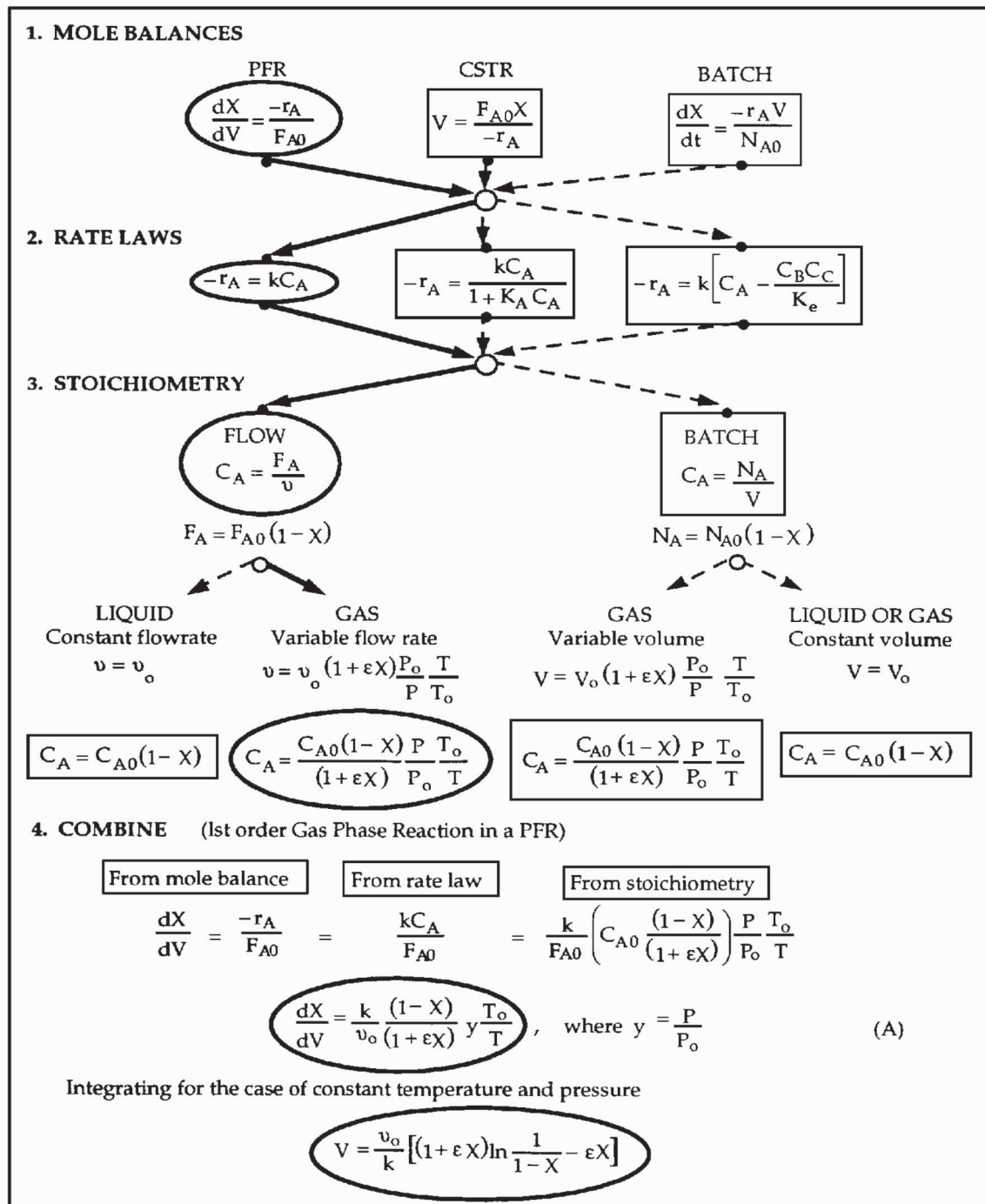
We now will apply the algorithm described earlier to a specific situation. Suppose we have, as shown in Fig. 76.2, mole balances for three reactors, three rate laws, and the equations for concentrations for both liquid and gas phases. In Fig. 76.3 the algorithm is used to formulate the equation to calculate the PFR reactor volume for a first-order gas-phase reaction. The pathway to arrive at this equation is shown by the ovals connected to the dark lines through the algorithm. The dashed lines and the boxes represent other pathways for other solutions. For the reactor and reaction specified, we follow these steps:

1. Choose the *mole balance* on species  $A$  for a PFR.
2. Choose the *rate law* for an irreversible first-order reaction.
3. Choose the equation for the concentration of  $A$  in the gas phase (*stoichiometry*).
4. Finally, *combine* to calculate the volume necessary to achieve a given conversion or calculate the conversion that can be achieved in a specified reaction volume.

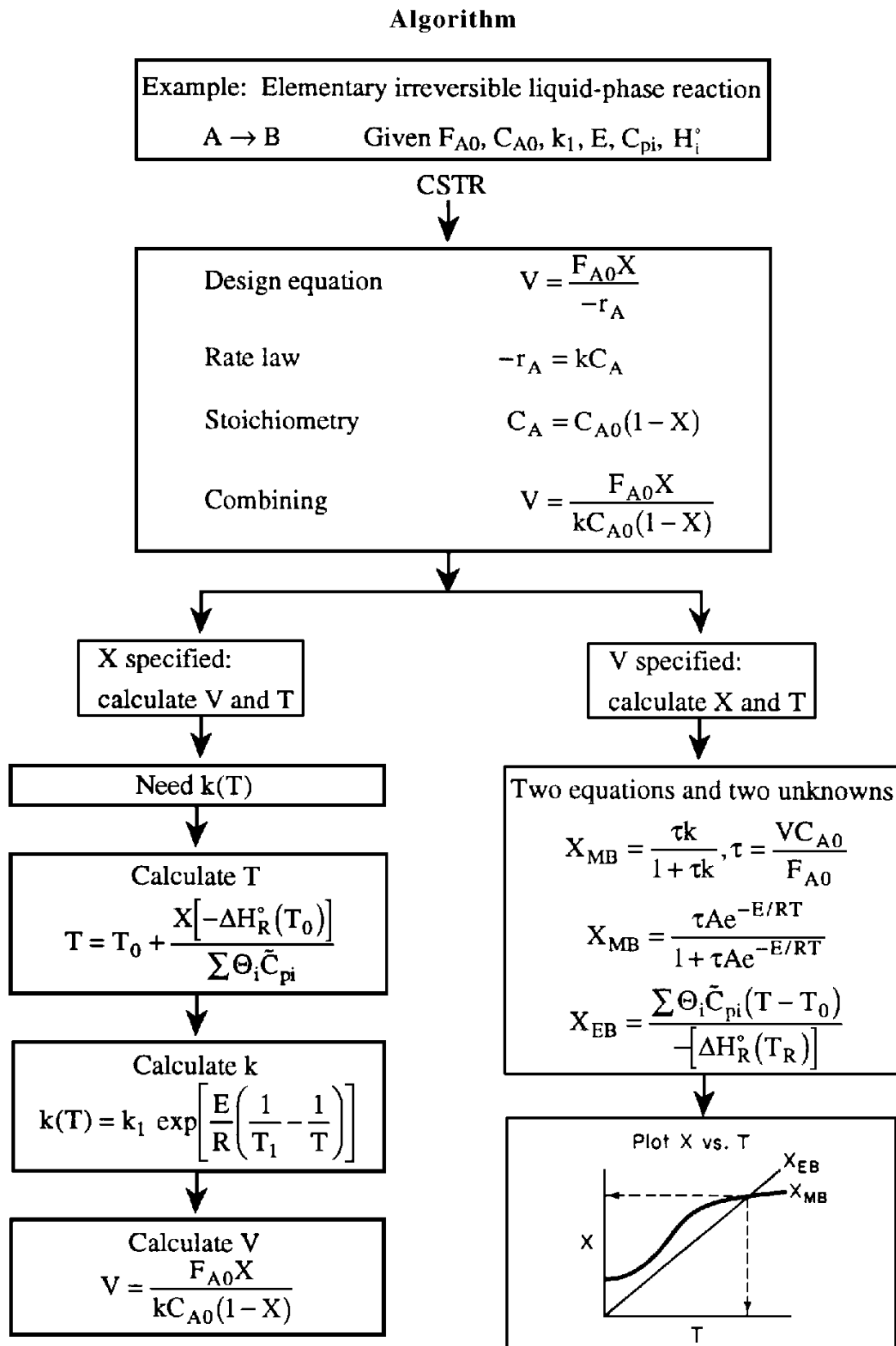
For the case of isothermal operation with no pressure drop, we were able to obtain an analytical

solution, given by equation (A) in Fig. 76.2, which gives reactor volume necessary to achieve a conversion  $X$  for a gas phase reaction carried out isothermally in a PFR. However, in the majority of situations, analytical solutions to the ordinary differential equations appearing in the combine step are not possible. By using this structure, one should be able to solve reactor engineering problems through reasoning rather than memorization of numerous equations together with the various restrictions and conditions under which each equation applies (i.e., whether there is a change in the total number of moles, etc.). In perhaps no other area of engineering is mere formula plugging more hazardous; the number of physical situations that can arise appears infinite, and the chances of a simple formula being sufficient for the adequate design of a real reactor are vanishingly small.

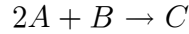
**Figure 76.2** Algorithm for isothermal reactors. (Source: Fogler, H. S. 1996. *Elements of Chemical Reaction Engineering*. 3rd ed. Prentice Hall, Englewood Cliffs, NJ.)



**Figure 76.3** Algorithm for nonisothermal CSTR design. (Source: Fogler, H. S. 1992. *The Elements of Chemical Reaction Engineering*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.)



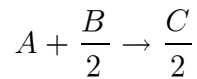
**Example.** The elementary gas phase reaction



$$-r_A = k_A C_A^2 C_B$$

is carried out at constant  $T$  (500 K) and  $P$  (16.4 atm) with  $k_A = 10 \text{ dm}^6/\text{mol}^2 \cdot \text{s}$ . Determine the CSTR reactor volume necessary to achieve 90% conversion when the feed is 50% mole  $A$  and 50%  $B$ .

**Solution.** The feed is equal molar in  $A$  and  $B$ ; therefore,  $A$  is the limiting reactor and taken as our basis of calculation:



Mole balance is given as:

$$V = \frac{F_{A0} X}{-r_A} \quad (76.27)$$

Rate law is given as:

$$-r_A = k_A C_A^2 C_B \quad (76.28)$$

Stoichiometry is found as follows:

$$C_A = C_{A0} \cdot \frac{(1 - X)}{(1 + \varepsilon X)} \frac{P}{P_0} \frac{T_0}{T} = C_{A0} \frac{(1 - X)}{(1 + \varepsilon X)}$$

$$\varepsilon = y_{A0} \delta = 0.5 \left[ \frac{1}{2} - \frac{1}{2} - 1 \right] = -0.5 \quad (76.29)$$

$$C_A = C_{A0} (1 - X) / (1 - 0.5X)$$

$$C_{A0} = \frac{y_{A0} P_0}{RT_0} = \frac{(0.5)(16.4 \text{ atm})}{\frac{0.082 \text{ atm m}^3}{\text{kmol K}} \cdot 500 \text{ K}} = 0.2 \frac{\text{kmol}}{\text{m}^3} = 0.2 \frac{\text{mol}}{\text{dm}^3} \quad (76.30)$$



$$C_B = C_{A0} \frac{\Theta_B - \frac{1}{2}X}{(1 + \varepsilon X)} = C_{A0} \frac{(1 - 0.5X)}{(1 - 0.5X)} = C_{A0} \quad (76.31)$$

For the combine step,

$$\begin{aligned} -r_A &= k_A C_A^2 C_B = k_A C_{A0}^3 \frac{(1 - X)^2}{(1 - 0.5X)^2} \\ &= 0.08 \frac{\text{mol}}{\text{dm}^3 \cdot \text{s}} \frac{(1 - X)^2}{(1 - 0.5X)^2} \end{aligned} \quad (76.32)$$

For a CSTR,

$$\begin{aligned} V &= \frac{F_{A0} X}{-r_A} = \frac{(5 \text{ mol/s})(0.9)[1 - 0.5(0.9)]^2}{(0.08) \frac{\text{mol}}{\text{dm}^3 \cdot \text{s}} (1 - 0.9)^2} \\ &= 1701 \text{ dm}^3 \end{aligned} \quad (76.33)$$

## 76.2 Pressure Drop in Reactors

---

If pressure drop is not accounted for in gas phase reactions, significant underdesign of the reactor size can result. This variation is handled in the stoichiometry step, in which concentration is expressed as a function of conversion, temperature, and total pressure. The change in total pressure is given by the Ergun equation [Fogler, 1992]:

$$\frac{dP}{dL} = -\frac{G(1 - \phi)}{\rho g_C D_p \phi^3} \left[ \frac{150(1 - \phi)\mu}{D_p} + 1.75G \right] \quad (76.34)$$

For isothermal operation the density is (assuming ideal gas)

$$\rho = \frac{\rho_0}{(1 + \varepsilon X)} \frac{P}{P_0} \quad (76.35)$$

The catalyst weight,  $W$ , and length down the reactor,  $L$ , are related by the equation

$$W = LA_c(1 - \Phi)\rho_{\text{cat}}$$

Substituting back in the Ergun equation,

$$\frac{dP}{dW} = -\frac{\alpha_P}{2} \frac{(1 + \varepsilon X)}{\frac{1}{P_0} \left( \frac{P}{P_0} \right)} \quad (76.36)$$

where

$$\alpha_p = \frac{\frac{G(1 - \Phi)}{\rho_0 g_c D_p \Phi^3} \left[ \frac{150(1 - \Phi)\mu}{D_p} + 1.75G \right]}{A_c(1 - \Phi)\rho_{\text{cat}} P_0}$$

We now need to solve this differential equation to obtain the pressure as a function of the weight of catalyst the gas has passed over. We can obtain an analytical solution of  $\varepsilon = 0$ . Otherwise, we must solve the equation numerically and simultaneously with the mole balance. For an analytical solution,

$$\frac{d(P/P_0)^2}{dW} = -\alpha_p(1 + \varepsilon X) \quad (76.37)$$

For POLYMATH solution, letting  $y = P/P_0$ ,

$$\frac{dy}{dW} = -\frac{\alpha_p(1 + \varepsilon X)}{2y} \quad (76.38)$$

**Example.** To understand the effect pressure drop has on gas phase reaction in a packed bed, we analyze the reaction  $A \rightarrow B$  carried out in a packed bed reactor (PBR). Mole balance (PBR) is

$$F_{A0} \frac{dX}{dW} = -r'_A$$

Wherever pressure drop occurs in a PBR we must use the differential form of the mole balance to separate variables. Pressure drop only affects  $C_A$ ,  $C_B$ , and so on, as well as  $-r'_A$ .

Rate law is second order in A and irreversible, according to the formula  $-r'_A = kC_A^2$ .

Stoichiometry is given by

$$C_A = C_{A0} \frac{(1 - X)}{(1 + \varepsilon X)} \frac{P}{P_0} \frac{T_0}{T}$$

For  $\varepsilon = 0$  and isothermal operation,

$$\frac{P}{P_0} = (1 - \alpha_p W)^{1/2}$$

$$C_A = C_{A0}(1 - X) \frac{P}{P_0} = C_{A0}(1 - X) (1 - \alpha_p W)^{1/2}$$

Combining,

$$F_{A0} \frac{dX}{dW} = -r'_A = kC_{A0}^2 (1 - X)^2 (1 - \alpha_p W)$$

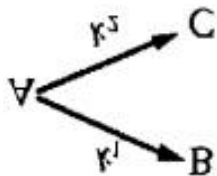
By integrating with limits  $W = 0, X = 0$  we obtain the desired relationship between conversion and catalyst weight.

$$\frac{X}{1 - X} = \frac{kC_{A0}^2}{F_{A0}} \left[ W - \frac{\alpha_p W^2}{2} \right]$$

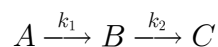
## 76.3 Multiple Reactions

---

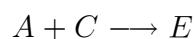
There are three basic types of multiple reactions: series, parallel, and independent. In *parallel reactions* (also called *competing reactions*) the reactant is consumed by two different reactions to form different products:



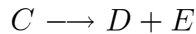
In *series reactions*, also called *consecutive reactions*, the reactant forms an intermediate product, which reacts further to form another product:



Multiple reactions involve a combination of both series and parallel reactions, such as

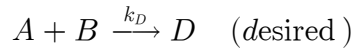


*Independent* reactions are of the type



and occur in feed stocks containing many reactants. The cracking of crude oil to form gasoline is an example of an independent reaction.

To describe selectivity and yield we consider the following competing reactions.



The rate laws are

$$r_D = k_D C_A^{\alpha_1} C_B^{\beta_1} \quad (76.39)$$

$$r_U = k_U C_A^{\alpha_1} C_B^{\beta_2} \quad (76.40)$$

We want the rate of  $D$ ,  $r_D$ , to be high with respect to the rate of formation  $U$ ,  $r_U$ . Taking the ratio of these rates, we obtain a rate *selectivity parameter*,  $S$ , which is to be maximized:

$$S_{DU} = \frac{r_D}{r_U} \quad (76.41)$$

Substituting Eqs. (76.39) and (76.40) into Eq. (76.41) and letting  $a = \alpha_1 - \alpha_2$  and  $b = \beta_2 - \beta_1$ , where  $a$  and  $b$  are both positive numbers, we have

$$S_{DU} = \frac{r_D}{r_U} = \frac{k_1 C_A^a}{k_2 C_B^b}$$

To make  $S_{DU}$  as large as possible we want to make the concentration of  $A$  high and the concentration of  $B$  low. To achieve this result, use the following:

A semibatch reactor in which  $B$  is fed slowly into a large amount of  $A$

A tubular reactor with side streams of  $B$  continually fed to the reactor

A series of small CSTRs with  $A$  fed only to the first reactor and  $B$  fed to each reactor

Another definition of selectivity used in the current literature is given in terms of the flow rates leaving the reactor:

$$\tilde{S}_{DU} = \text{Selectivity} = \frac{F_D}{F_U} = \frac{\text{Exit molar flow rate of desired product}}{\text{Exit molar flow rate of undesired product}} \quad (76.42)$$

For a batch reactor the selectivity is given in terms of the number of moles of  $D$  and  $U$  at the

end of the reaction time:

$$\tilde{S}_{DU} = \frac{N_D}{N_U} \quad (76.43)$$

One also finds that the reaction yield, like the selectivity, has two definitions: one based on the ratio of reaction rates and one based on the ratio of molar flow rates. In the first case the yield at a point can be defined as the ratio of the reaction rate of a given product to the reaction rate of the key reactant A [Carbery, 1967]:

$$Y_D = \frac{r_D}{-r_A} \quad (76.44)$$

In the case of reaction yield based on molar flow rates, the yield is defined as the ratio of moles of product formed at the end of the reaction to the number of moles of the key reactant, A, that have been consumed. For a batch system,

$$\tilde{Y}_D = \frac{N_D}{N_{A0} - N_A} \quad (76.45)$$

For a flow system,

$$\tilde{Y}_D = \frac{F_D}{F_{A0} - F_A} \quad (76.46)$$

Because of the various definitions for selectivity and yield, when reading literature dealing with multiple reactions, check carefully to ascertain the definition intended by the author.

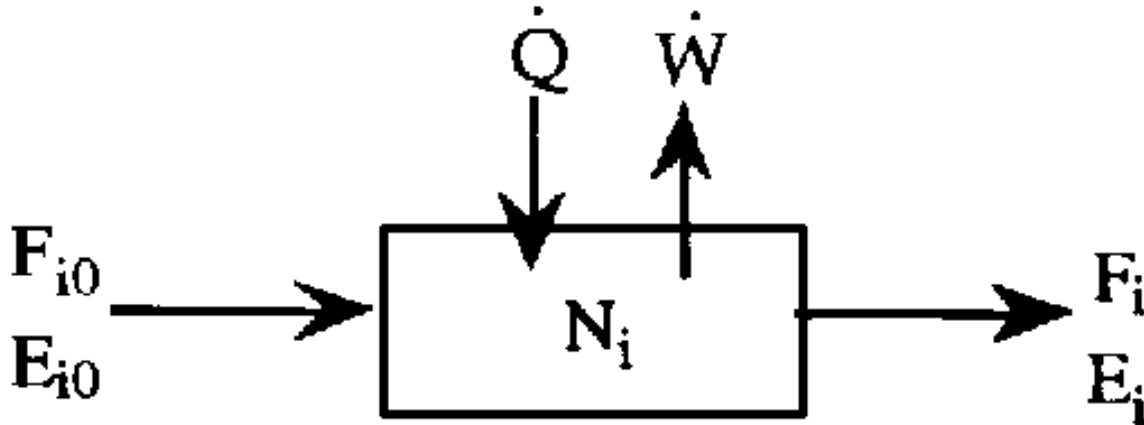
## 76.4 Heat Effects

For nonisothermal reaction in CRE we must choose which form of the energy balance to use (e.g., PFR, CSTR) and which terms to eliminate (e.g.,  $Q = 0$  for adiabatic operation). The structure introduced to study these reactors builds on the isothermal algorithm by introducing the Arrhenius equation,  $k = Ae^{-E/RT}$  in the **rate law** step, which results in one equation with two unknowns,  $X$  and  $T$ , when we finish with the combine step. For example, using again the PFR mole balance and conditions in Fig. 76.2[Eq. (A)], we have, for constant pressure,

$$\frac{dX}{dV} = \frac{k(1-X)}{v_0(1+\varepsilon X)} \frac{T_0}{T} \quad (76.47)$$

$$\frac{dX}{dV} = \frac{Ae^{-E/RT} (1-X)}{v_0(1+\varepsilon X)} \left( \frac{T_0}{T} \right) \quad (76.48)$$

We can now see the necessity of performing an energy balance on the reactor to obtain a second equation relating  $X$  and  $T$ . We will use energy balance to relate  $X$  and  $T$ .



$$\begin{aligned}
 \left[ \begin{array}{c} \text{Rate of} \\ \text{accumulation} \\ \text{of energy} \\ \text{within the} \\ \text{system} \end{array} \right] &= \left[ \begin{array}{c} \text{Rate of flow} \\ \text{of heat to} \\ \text{the system} \\ \text{from the} \\ \text{surroundings} \end{array} \right] - \left[ \begin{array}{c} \text{Rate of work} \\ \text{done by} \\ \text{the system} \\ \text{on the} \\ \text{surroundings} \end{array} \right] \\
 &+ \left[ \begin{array}{c} \text{Rate of energy} \\ \text{added to the} \\ \text{system by mass} \\ \text{flow into the} \\ \text{system} \end{array} \right] - \left[ \begin{array}{c} \text{Rate of} \\ \text{energy leaving} \\ \text{system by mass} \\ \text{flow out of} \\ \text{the system} \end{array} \right]
 \end{aligned}$$

$$\frac{dE}{dt} = \dot{Q} - \dot{W} + \sum F_{in} E_{in} - \sum F_{out} E_{out}$$

Neglecting potential and kinetic energy, the energy  $E_i$  is just the internal energy of species  $i$ , substituting for  $W$  in terms of the flow work and the shaft work,  $W_S$ , and using the definition of enthalpy gives for steady state operation

$$\dot{Q} - \dot{W}_s + \sum F_{i0} H_{i0} - \sum F_i H_i = 0 \quad (76.49)$$

For adiabatic operation, no work done on the system and constant heat capacity and  $\Delta C_p = 0$ , the energy balance reduces to

$$T = T_0 + \frac{(-\Delta H_R) X}{\sum \theta_i C_{pi} + \Delta C_p X} \quad (76.50)$$

We now use this relationship to solve adiabatic reactor design problems.

The procedure for nonisothermal reactor design can be illustrated by considering the first-order irreversible liquid-phase reaction  $A \rightarrow B$ . The CSTR design equation is

$$V = \frac{F_{A0} X}{-r_A}$$

Rate law is found by

$$-r_A = kC_A \quad (76.51)$$

with the Arrhenius equation:

$$k = Ae^{-E/RT}$$

Stoichiometry for the liquid phase (i.e.,  $v = v_0$ ) is given by

$$C_A = C_{A0}(1 - X)$$

Combining yields

$$V = \frac{v_0}{Ae^{-E/RT}} \left( \frac{X}{1 - X} \right) \quad (76.52)$$

Continuing from this point requires two distinct cases. For the first case, the variables  $X$ ,  $v_0$ ,  $C_{A0}$ , and  $F_{i0}$  are specified and the reactor volume,  $V$ , must be determined. The procedure is as follows:

1. Evaluate Eq. (76.50) to find the temperature,  $T$ , for the conditions specified.
2. Calculate  $k$  from the Arrhenius equation.
3. Calculate the reactor volume,  $V$ , from Eq. (76.52).

For the second case, the variables  $v_0$ ,  $C_{A0}$ ,  $V$ , and  $F_{i0}$  are specified and the exit temperature,  $T$ , and conversion,  $X$ , are unknown quantities. The procedure is as follows:

1. Solve the energy balance for  $X$  as a function of  $T$ . If adiabatic, Eq. (76.50) becomes

$$X_{EB} = \frac{\sum \Theta_i \tilde{C}_{pi} (T - T_0)}{-[\Delta H_R^\circ(T_R)]} \quad (76.53)$$

2. Solve the mole balance [Eq.(76.52)]for  $X$  as a function of  $T$ .

$$X_{MB} = \frac{\tau Ae^{-E/RT}}{1 + \tau Ae^{-E/RT}}$$

where  $\tau = V/v_0$

3. Plot the previous two steps on the same graph to determine the intersection. At this point the values of  $X$  and  $T$  satisfy both the energy balance and mole balance. As an alternative, one may equate the equations for  $X$  from the previous two steps and solve numerically.

An energy balance on a PFR with heat exchange yields the second equation we need relating our independent variables  $X$  and  $T$ :

$$\frac{dT}{dV} = \frac{[U A_c (T_a - T) + (r_A)(\Delta H_R)]}{F_{A0} C_{P_A}} \quad (76.54)$$

The differential equation describing the change of temperature with volume (i.e., distance) down the reactor,

$$\frac{dT}{dV} = g(X, T) \quad (76.55)$$

must be coupled with the mole balance, Eq. (76.5),

$$\frac{dX}{dV} = \frac{-r_A}{F_{A0}} f(X, T) \quad (76.56)$$

and solved simultaneously. A variety of numerical integration schemes and ODE solvers can be used to solve these two equations simultaneously.

#### SHELL'S HYCON PROCESS

As long ago as 1967, the laboratory of the Royal Dutch/Shell Group in Amsterdam headed the development of the HYCON process, by which even the heaviest oil fraction could be converted into premium products, such as gasoline, kerosene, and gasoil.

This had not really been necessary before the first oil crisis in 1973. Oil then was still cheap and in plentiful supply and there was a ready market for all products.

After the oil crises in 1973 and 1979, however, prices had risen to such a level as to provide a powerful incentive for producing the greatest amount of light products from the crude oil, which had, moreover, been getting gradually heavier.

Initially, this was no problem for complex refineries such as Shell Pernis. In contrast with other refineries, Shell Pernis was able to treat the heavy residue remaining after atmospheric distillation by redistilling it under high vacuum. In this way, valuable products such as gasoline, gasoil, and lubricating oil distillates could be manufactured.

But even after this conversion method, a black, viscous mass still remained, which could only be used as a component of heavy fuel oil after adding gasoil. Adding gasoil, however, was a costly solution for a product that was earning less and less. In the Shell laboratories, meanwhile, people were hard at work on a process with better economic prospects: HYCON.

This process is based on the fact that the vacuum residue contains much less hydrogen than the



lighter products. The aim, therefore, is to reduce the size of the molecules and to increase the ratio of hydrogen atoms to carbon atoms.

This ratio can be increased either by removing some carbon from the molecules or by attaching hydrogen to the atoms with the aid of a catalyst. Shell chose the latter approach and called its process HYCON (hydroconversion).

This all sounds much simpler than it actually is. Reducing the size (cracking) of the large molecules is only possible at high temperatures, but the problem with heating is that it produces a heavy deposit of carbon on the catalyst. This problem, and also the addition of hydrogen to the molecules, was solved by carrying out the process under high pressure. Another difficulty is that the asphaltenes (substances with a complex molecular structure) in the residue react very slowly. The reaction mixture therefore has to remain in the reactors for a long time. All in all, this means: several reactors in series, large amounts of catalyst, and high pressures and temperatures in the reactors.

There was another problem that had to be solved in the laboratories. The residue contains heavy metals, such as vanadium and nickel, which poison the catalyst. The solution for this is first of all to pass the residue through a number of reactors containing a specially developed catalyst which removes most of the metals. A technique was also developed to regenerate this catalyst during the process.

The HYCON process required the largest capital investment ever made by Shell in the Netherlands. HYCON in Pernis was put into operation at the beginning of 1989. (Courtesy of the Shell Group.)

## 76.5 Summary

---

By arranging chemical reaction engineering in a structure analogous to a French menu, we can study a multitude of reaction systems with very little effort. This structure is extremely compatible with a number of user-friendly ordinary differential equation (ODE) solvers. Using **ODE solvers** such as POLYMATH, the student is able to focus on exploring reaction engineering problems rather than crunching numbers. Thus, the teacher is able to assign problems that are more open ended and give students practice at developing their creativity. Practicing creativity is extremely important, not only in CRE, but in every course in the curriculum if students are to compete in the world arena and succeed in solving the relevant problems that they will be faced with in the future.

### Nomenclature

$A$	frequency factor (appropriate units)
$A_C$	cross-sectional area, $\text{m}^2$
$C_i$	concentration of species $i$ ( $i = A, B, C, D$ ), $\text{mol/dm}^3$
$C_{Pi}$	heat capacity of species $i$ , $\text{J/g/K}$

$D_P$	particle diameter, m
$E$	activation energy, J/mol
$F_i$	entering molar flow rate of species $i$ , mol/s
$G$	superficial gas velocity g/m <sup>2</sup> /s
$g_c$	conversion factor
$k$	specific reaction rate (constant), appropriate units
$K_A$	adsorption equilibrium constant (dm <sup>3</sup> /mol)
$K_e$	equilibrium constant, appropriate units
$L$	length down the reactor, m
$N_t$	number of moles of species $i$ , mol
$P$	pressure, kPa
$r_t$	rate of formation of species $i$ per unit volume, mol/s/dm <sup>3</sup>
$r_t'$	rate of formation of species $i$ per unit mass of catalyst, mol/s/g
$R$	ideal gas constant, J/mol/K
$t$	time, s
$T$	temperature, K
$U$	overall heat transfer coefficient, J/(dm <sup>3</sup> s K)
$V$	volume, dm <sup>3</sup>
$W$	catalyst weight, g
$X$	conversion
$y$	pressure drop parameter ( $P/P_0$ )
$y_A$	mole fraction of $A$
$a$	ambient temperature
$A$	refers to species $A$
cat	catalyst density kg/m <sup>3</sup>
EB	energy balance
MB	mole balance
$T$	total number of moles
0	entering or initial condition
$\alpha$	reaction order
$\alpha_P$	pressure drop parameter, g <sup>-1</sup>
$\beta$	reaction order
$\Delta H_R$	heat of reaction, J/mole $A$
$\delta$	change in the total number of moles per mole of $A$ reacted
$\epsilon$	volume change parameter = $y_{A0} \delta$
$\phi$	porosity
$\mu$	viscosity, cp
$\rho$	density, g/dm <sup>3</sup>
$\nu$	volumetric flow rate, dm <sup>3</sup> /s

$$\Theta_i \quad N_i/N_{A0}$$

## Defining Terms

**Batch reactor:** A closed vessel (tank) in which there is no flow in or out of the vessel during the time the reaction is taking place.

**Continuous stirred tank reactor (CSTR):** A reactor in which the reactant and products flow continuous into and out of (respectively) the tank. A reactor where the contents are well mixed.

**ODE solver:** A user-friendly software package that solves ordinary differential equations, for example, Mathematica, POLYMATH, Matlab.

**Packed bed reactor:** Usually a tubular reactor packed with solid catalyst pellets.

**Plug flow reactor:** Usually a tubular reactor used for gas phase reactions in which it is assumed there are no radial gradients in temperature or concentration as well as no dispersion of reactants.

**Semibatch reactor:** A reactor (vessel) in which one of the reactants is placed in the reactor and a second reactant is slowly added to the reactor.

## References

Fogler, H. S. 1992. *The Elements of Chemical Reaction Engineering*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.

Shacham, M. and Cutlip, M. B. 1988. Applications of a microcomputer computation package in chemical engineering. *Chemical Engineering Education*. 12(1):18

Carbery, J. J. 1967. Applied kinetics and chemical reaction engineering, Chemical engineering education, ed. R. L. Goring and V. W. Weekman, p.89. *American Chemical Society*, Washington, DC.

## Further Information

### Professional Organizations

The American Institute of Chemical Engineers (three national meetings per year), 345 E. 47th St., New York, NY 10017. Phone (212) 705-7322.

The American Chemical Society (several national meetings each year), 1155 16th St., Washington, DC 20036. Phone (202) 872-4600.

### Special Meetings and Conferences

The Engineering Foundation Conferences on Chemical Reaction Engineering, 345 E. 47th St., New York, NY 10017. Phone (212) 705-7835.

International Symposia on Chemical Reaction Engineering (even years), sponsored by American Institute of Chemical Engineers, American Chemical Society, Canadian Society for Chemical Engineering, and the European Federation of Chemical Engineering.

**Professional Journals**

*AIChE Journal*. Published monthly by the American Institute of Chemical Engineers, New York, NY.

*Chem. Eng. Sci.* Published semimonthly by Elsevier Science, Oxford, U.K.

Cropley, J. B. "The Scaleup of Chemical Reaction Systems from..."  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# The Scaleup of Chemical Reaction Systems from Laboratory to Plant

---

## 77.1 General Considerations in the Rational Design of Chemical Reactors

Reaction Kinetics Models and Reactor Models

## 77.2 Protocol for the Rational Design of Chemical Reactors

Step 1: Select the Type of Reactor for the Commercial Process • Step 2: Design the Laboratory to Generate Reaction Kinetics Data • Step 3: Use Statistically-Valid Experimental Programs and Models • Step 4: Develop Computer Programs for Reactor Simulation and Design • Step 5: Develop the Economically Optimum Reactor Design • Step 6: Validate the Design in a Pilot Plant Reactor

### J. B. Cropley

*Union Carbide Corporate Fellow(Retired)*

*Scaleup* is one of those overworked terms that has come to mean almost anything and everything, depending on who is using it. In this article we will use it very little, and then only to indicate the generic process of commercializing new chemical technology. The alternative to scaleup is rational design, which utilizes mathematical relationships and computer simulation to develop the best design for the reactor. The mathematical relationships describe both the reaction kinetics and the attributes of the reactor and its associated auxiliary equipment.

Geometric scaleup was practiced routinely in the chemical industry as a design protocol until a few decades ago, but, today, rational design—based on laboratory data and correlations—has largely replaced it for most types of industrial chemical reaction systems. To understand why this is so, it is necessary to note how the chemical industry has changed over the years.

Forty or 50 years ago, merely producing a chemical on an industrial scale was usually sufficient to ensure a profit for the manufacturer. Chemical processes were labor intensive, but capital and energy were cheap and the selling price of a pound of finished product was typically several times the raw material cost. Furthermore, the environment had not yet been discovered either by industry or by the public at large.

It was really unnecessary to design reactors rationally for most kinds of processes in that era, because any questions about productive capacity could be addressed simply by making the reactor larger, raw material selectivities usually were not economically critical, and the large quantities of energy that were expended in complex distillation trains to remove by-products and impurities were both practicable and inexpensive.

In contrast, the petrochemical industry today utilizes chemical reactions that produce the desired products much more directly and cleanly, with as little waste as possible. Raw material cost is

frequently the largest component of the final product cost, and the crude product must not contain unexpected by-products that the refining system cannot remove adequately and efficiently. The failure to meet tight product specifications can produce chemicals that either cannot be sold at all without expensive reprocessing or can be sold only for little more than their value as fuel.

In any case, both today's marketplace and concerns for the environment demand that chemical reaction systems produce no more than extremely small amounts of waste or off-specification product per pound of refined salable product. Reactors must be accurately designed and operated because today's chemistry frequently is strongly dependent upon carrying out just the desired amount of reaction in order to avoid the production of unwanted by-products by over-reaction. The energy efficiencies of refining systems strongly depend upon their receiving crude product of uniform and predictable composition, because product specifications are usually tight and must be met at minimum cost. Simply put, today's chemical reaction systems must operate as intended.

## 77.1 General Considerations in the Rational Design of Chemical Reactors

---

### Reaction Kinetics Models and Reactor Models

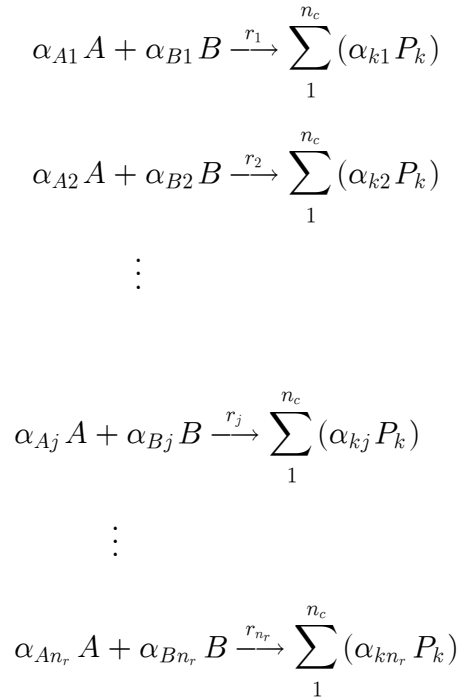
The rational design of any type of reactor involves the marriage of one or more reaction kinetics models and a reactor model. It is important to recognize exactly what these two kinds of models describe:

- *Reaction kinetics models* describe the response of reaction rates to the reaction environment—that is, to temperature and the concentrations of virtually everything in the system—reactants, products, by-products, catalysts, and contaminants. For design purposes, it is necessary and sufficient that the kinetic model reflect the reaction stoichiometry accurately and that it predict reaction rates accurately. It is not important that it reflect the actual reaction mechanism.
- *Reactor models* describe how the reaction environment is shaped by the geometry of the reactor, by physical processes like fluid dynamics and heat and mass transport, and by process variables and conditions such as mean reactor residence time and residence time distribution, flow rate, pressure, and temperature.

These distinctions are subtle, but important. It is sufficient to remember that kinetics models contain *only* temperature and concentration terms, whereas reactor models may contain these as well as everything else that influences the conduct of the reaction.

### Kinetics of a Simple Hypothetical System of Reactions

Real systems will have their own individual structures and characteristics and will reflect the particular stoichiometry of the reaction system at hand. In this chapter, we will use a general, somewhat simplified set of reactions and reactor equations for illustration. Consider the general group of  $n_r$  reactions presented below, in which chemical species  $A$  and  $B$  react to produce several products  $P_k$  according to the following scheme:



Kinetic models for the rational design of reaction systems will usually be of one of two basic mathematical forms, each given in moles/volume/time and describing the response of the reaction rates  $r_j$  to temperature and concentration. For exponential models,

$$r_j = K_{o_j} e^{-\frac{E_{a_j}}{RT}} C_A^{a_j} C_B^{b_j} C_{P_1}^{P_{1j}} \dots C_{P_n}^{p_{nj}} \quad (77.1)$$

For hyperbolic models,

$$r_j = \frac{K_{o_j} e^{-\frac{E_{a_j}}{RT}} C_A C_B}{1 + K_{A_j} C_A + K_{B_j} C_B + \sum_1^{n_c} (K_{P_{kj}} C_{P_k})} \quad (77.2)$$

Each of these two types of kinetic models has its own preferred uses. Their development will be discussed in a later section of this chapter. For now, assume that either of these types is to be used in a combined kinetics and reactor model to predict reaction rates, as described in the following section.

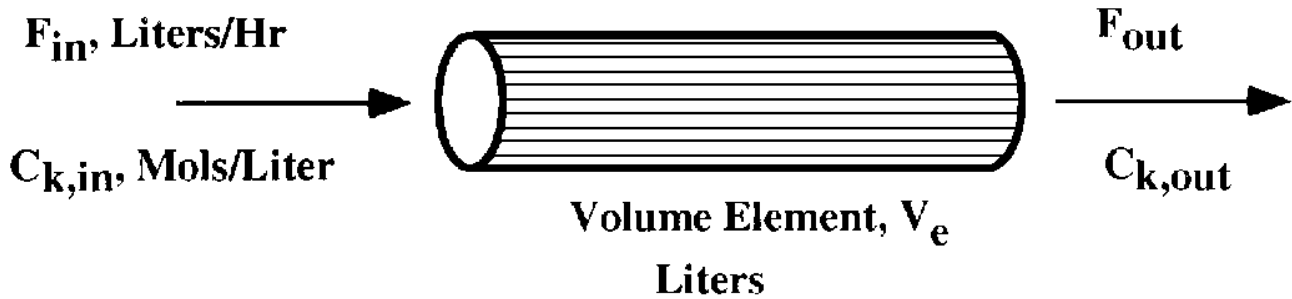
### Combined Kinetics and Reactor Models: The General Continuity Equation

Virtually all combined reaction kinetics and reactor models utilize some form of the general continuity equation for flow and reaction in an element of volume of some kind of reactor, such as that shown in [Fig. 77.1](#). In its simplest form, this equation states that the quantity of each component  $k$  that enters the volume element of the reactor must either leave, react, or accumulate:



$$F_{e_{in}} C_{k_{in}} = F_{e_{out}} C_{k_{out}} - V_e \sum_1^{n_r} (\alpha_{kj} r_j) + \frac{d(V_e C_{k_e})}{dt} \text{ mol/time} \quad (77.3)$$

**Figure 77.1** Flow and concentration in a volume element of a chemical reactor.



Different types of idealized reactors can be represented by this equation, simply by noting what terms are not appropriate and eliminating them. Thus, a simple *batch reactor* has no flow in or out, and therefore terms containing  $F$  drop out. So, for a simple batch reactor,

$$\frac{d(V_e C_{k_e})}{dt} = \sum_1^{n_r} (\alpha_{kj} r_j) V_e \quad (77.4)$$

which may be further simplified by canceling the  $V_e$  terms as well if the reaction volume is constant. Note that the volume element in this case is simply the entire filled volume of the reactor, because the concentration is assumed to be uniform throughout.

A steady state *continuous stirred tank reactor (CSTR)* may have the same flow rate in and out, and the entire contents of the reactor comprise the volume element, as in the batch reactor. The time-derivative term is absent because the reactor is at steady state. The concentrations of all species in the reactor are the same as in the outlet. Thus we have, for the CSTR,

$$F C_{k_{in}} = F C_{k_{out}} - V \sum_1^{n_r} (\alpha_{kj} r_j) \quad (77.5)$$

which can be rearranged to

$$\sum_1^{n_r} \alpha_{kj} r_j = \frac{F(C_{k_{out}} - C_{k_{in}})}{V} \text{ mol/volume/time} \quad (77.6)$$

Weight of catalyst,  $W_c$ , replaces reactor volume,  $V$ , for a catalytic reaction:

$$\sum_1^{n_r} \alpha_{kj} r_j = \frac{F(C_{k_{out}} - C_{k_{in}})}{W_c} \text{ mol/wt. catalyst/time} \quad (77.7)$$

This simple relationship makes the CSTR the preferred type of reactor for many kinds of kinetics studies, in which the net rates of formation or disappearance [that is,  $\sum_1^{n_r} (\alpha_{kj} r_j)$ ] for each individual component can be observed directly.

For design purposes it is more convenient to integrate the unsteady state form of the general continuity equation for the CSTR until the steady state concentrations are attained. (The model will behave very much as the real reactor would in this respect.) This procedure avoids the need to use constrained nonlinear estimation to predict the reactor outlet concentrations. Again, assuming that flows in and out of the reactor are the same and that volume is constant,

$$\frac{d(C_k)}{dt} = \frac{F(C_{k_{in}} - C_{k_{out}})}{V} + \sum_1^{n_r} (\alpha_{kj} r_j) \text{ mol/h} \quad (77.8)$$

The reactor model will comprise an equation like Eq. (77.8) for each component  $k$  in the CSTR. Another advantage of the unsteady state model is that stoichiometry is automatically preserved without the need for any constraints. It may be used readily to simulate a system comprising a large number of components and reactions in a multistage system of CSTRs, which is otherwise mathematically intractable. And, of course, it may be used to study the dynamic behavior of the system as well. Therefore, it is the preferred design approach for multistage CSTR systems.

In the *ideal steady state plug flow reactor*, all elements of fluid that enter the reactor together travel down its length and exit together, having thus stayed in the reactor for identical lengths of time. The volume element will be only a differential slice of the reactor cross section, denoted by  $dV$ . There will be no accumulation term, since the reactor is at steady state. The differential concentration difference across the differential volume element will be denoted by  $dC$ . Thus, Eq. (77.3) is once again applicable and simplifies to

$$F_e dC_k = \sum_1^{n_r} (\alpha_{kj} r_j) dV_e \quad (77.9)$$

If  $F$  changes as the reaction proceeds (as with many gas-phase reactions), then this can be accommodated down the length of the plug flow reactor by modifying the above equation to

$$d(F_e C_k) = \sum_1^{n_r} (\alpha_{kj} r_j) dV_e$$

whence

$$F_e \frac{dC_k}{dV_e} + C_k \frac{dF_e}{dV_e} = \sum_1^{n_r} (\alpha_{kj} r_j) dV_e$$

$$\frac{dC_k}{dV_e} = \frac{\sum_1^{n_r} (\alpha_{kj} r_j)}{F_e} - \frac{C_k}{F_e} \frac{dF_e}{dV_e} \quad (77.10)$$

Equation (77.10) describes the rate of change of concentration  $C$  of species  $k$  with volume down a plug flow reactor as a function of the reaction rates, concentration, and flow. It also reflects the change in molar flow as the reaction proceeds.

A complete *isothermal* plug flow reactor model can readily be constructed using as many Eqs. (77.10) as there are components  $k$ , and as many kinetic models as there are reactions  $j$ . For *nonisothermal* reactors, differential equations that describe the temperature changes down the length of the reactor can be constructed in an analogous fashion, using the molar heat generation for each reaction  $j$  and its corresponding reaction rate  $r_j$  and heat transfer terms appropriate to the reactor type and geometry. For a multitube plug flow reactor with coolant on the outside of the tubes, the equation for reaction temperature (in degrees/volume) is as follows:

$$\frac{dT_r}{dV_e} = \frac{\sum_1^{n_r} (r_j \Delta H_j) - \frac{4U}{D_t} (T_r - T_c)}{F_e \rho_p c_p} \quad (77.11)$$

For coolant temperature,

$$\frac{dT_c}{dV_e} = \frac{\frac{4U}{D_t} (T_r - T_c) (\text{Mode})}{F_c \rho_c c_c} \quad (77.12)$$

Combined model equations like these are simplified in the sense that they do not account for departures from ideality because of nonideal mixing patterns in the case of stirred reactors, radial and axial diffusion effects in the case of tubular catalytic reactors, or the very specialized phenomena in fluidized beds and fixed-bed multiphase reactors. Yet they are surprisingly applicable in many industrial applications—and will certainly be preferred to geometric scaleup in almost all cases.

## 77.2 Protocol for the Rational Design of Chemical Reactors

---

Given the distinctions between kinetics models and reactor models, as well as the characteristics of combined reaction and reactor models, we can now establish a general protocol for rational design that will apply to many types of chemical reactors in industrial situations. The protocol comprises several steps, each of which is discussed in the following sections.

### Step 1: Select the Type of Reactor for the Commercial Process

First, select one or more potentially useful types of reactors for the commercial process. In many instances, the preferred reactor type will be known from past experience with the same or similar reactions. Even so, the scaleup characteristics and requirements for two important types of reaction—batch reactions and solid-catalyzed reactions—merit special attention here. The reader is referred to the large open literature for additional information. Texts by Froment and Bischoff [1979], and Levenspiel [1972] are classic and are particularly recommended.

## Scaleup of Laboratory Batch Reactions

***Plant-Size Batch Reactors.*** Many reactions are conveniently studied in laboratory batch reactors, but batch reactors often are not preferred for full-scale operations, particularly if the reaction is rapid or if the planned plant will be quite large. Plant-size batch reactors are costly to operate, simply because they must be shut down, emptied, and recharged after a fairly short time—typically after only a few hours. This means that each reactor produces chemicals only on a part-time basis. As a consequence, batch reactors are usually preferred only for fairly high-priced specialty chemicals like pharmaceuticals that are produced at fairly low volumes—say, under 50 000 000 pounds per year—so that the required number and size of batch reactors are reasonable.

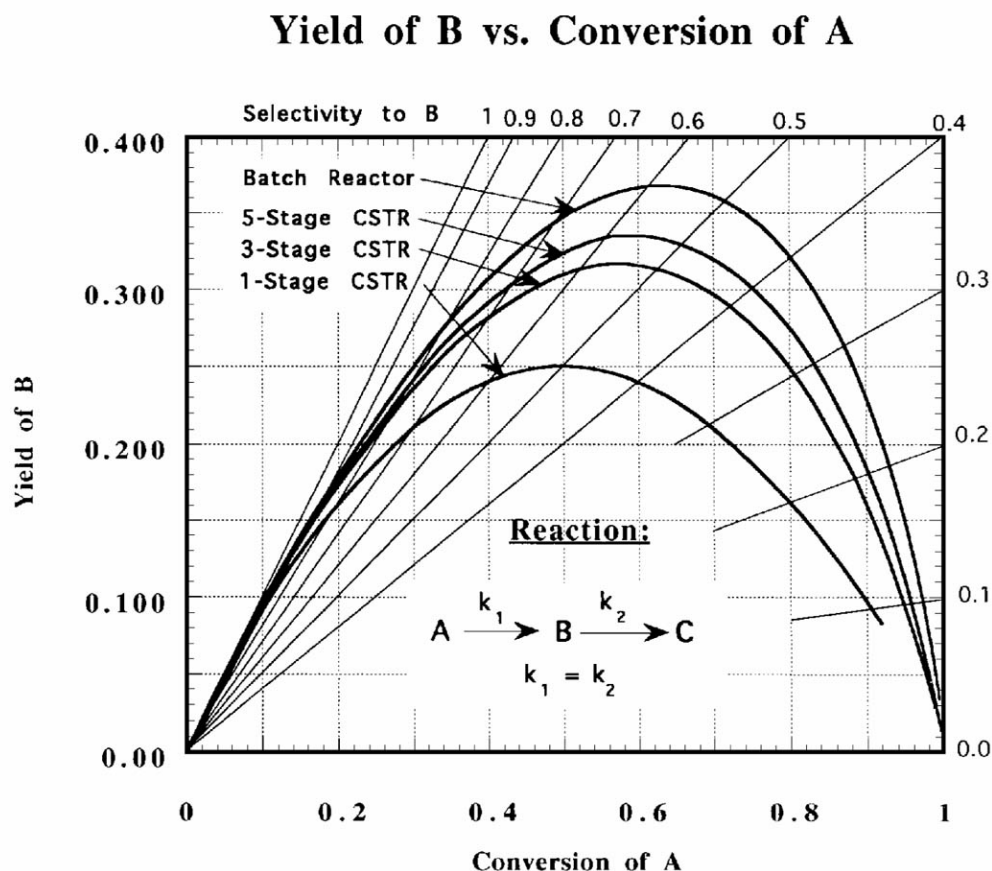
***Ideal Plug Flow Reactors (PFRs).*** PFRs have the same residence-time distribution as ideal batch reactors—that is, each element of feed is exposed to reaction conditions for exactly the same length of time. They are particularly useful for high-volume, low-priced commodity chemicals, for which the laboratory batch reactor is preferred. If the reaction time is short—say, under an hour—it may be practicable to use some type of large-scale plug flow reactor (PFR) for the plant. A baffled column is a common example. More often than not, however, batch times are at least several hours, and a plug flow continuous reactor would be too large and too costly for full-scale plant use. However, PFRs are routinely used in industry for solid-catalyzed reactions, as discussed later. The difference here is that the catalyst dramatically accelerates the reaction so that large-scale plug flow reactors are quite practicable.

***Single and Multistage CSTRs.*** The single-stage continuous stirred tank reactor (CSTR) is relatively inexpensive and provides good temperature control, but it has a broad residence-time distribution. This means that some of the feed may be under reaction for a very short time and some may be in the reactor for an extended time. Also, concentrations of reactants and products throughout a CSTR will be the same as their exit concentrations, so that the reactions are conducted at minimum reactant concentration and maximum product concentration. This will mean a relatively slow reaction rate with maximum exposure of products to further reaction, which may lead to relatively low production rates and relatively high by-product formation rates.

***Residence-Time Distribution and the Effects of Staging.*** To overcome some of the disadvantages of both the PFR and CSTR reactors for reactions that are ideally carried out by batch in the laboratory, several CSTRs are frequently connected in series. Their residence-time distributions will be intermediate between the very narrow distributions of the batch reactor or PFR and the very broad distributions of the CSTR. Such a multistage reactor may be a good compromise, but it will not operate identically to the laboratory batch reactor.

The effect of staging on a chemical reaction may be inferred from Fig. 77.2, in which a simple sequential reaction of  $A$  going to  $B$  (desired) and then to  $C$  (undesired) is assumed. The figure shows that the batch reactor has the highest conversion of  $A$  and the highest yield of  $B$ , and that the CSTR has the lowest. The three- and five-stage CSTRs are intermediate between the batch and CSTR reactors. In practice, multistage systems of two to five CSTRs are common. Froment and Bischoff [1979], Levenspiel [1972], and others have written extensively on residence-time distribution and its impact on product yields and selectivities. These concepts are important to the successful commercialization of batchwise laboratory reaction technology.

**Figure 77.2** Effects of residence-time distribution.



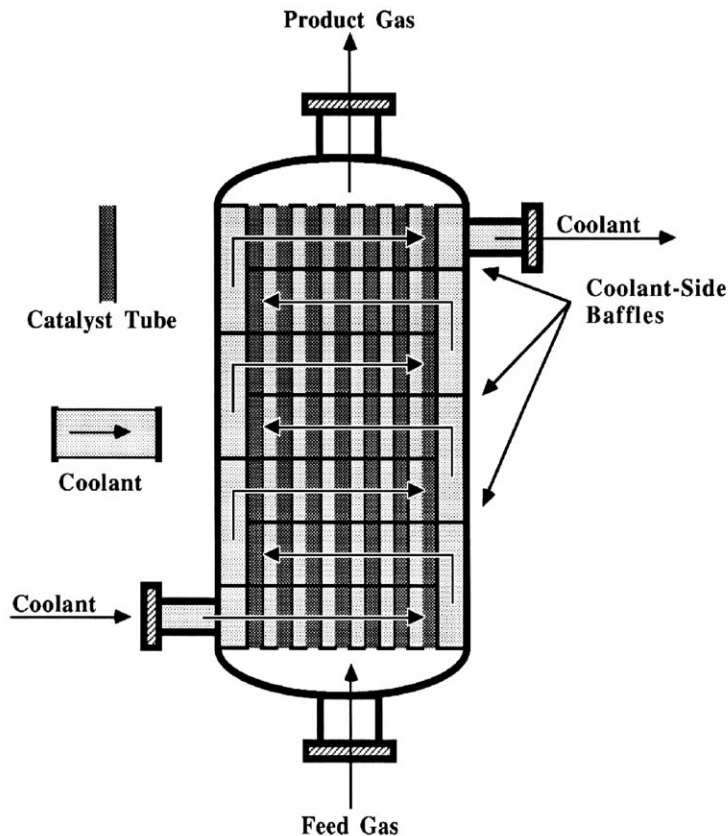
### Scaleup of Solid-Catalyzed Reactions

Solid catalysts are widely used in the chemical industry in several types of packed-bed reactors, slurry reactors, and fluidized bed reactors. Commercial processes frequently utilize solid-catalyzed gas-phase reactions, but it is not uncommon for them to use both gas and liquid streams, which greatly complicates the hydrodynamics. Reactors for solid-catalyzed reactions are all subject to scaleup problems because of differences in mass- and heat-transport in large-scale equipment and laboratory equipment.

*Packed-bed reactors* are probably used more than any other type for solid-catalyzed reactions. They tend to have residence-time distributions that closely approach plug flow, especially for plant-scale single-phase systems in which the bed length is at least 5 to 10 meters. Their narrow residence-time distribution makes them preferred from the standpoint of being able to control product distribution in systems of sequential reactions. Equations (77.10)–(77.12) are appropriate here.

*Shell-and-tube packed-bed reactors* have excellent heat removal characteristics, particularly if the tube diameters are fairly small—say, one to one-and-a-half inches. A schematic of a typical plant-scale shell-and-tube reactor is shown in Fig. 77.3. Although relatively expensive, such reactors can usually be designed rationally with confidence.

**Figure 77.3** Conceptual shell-and-tube fixed-bed reactor. (Source: Reproduced with permission of the American Institute of Chemical Engineers from Cropley, J. B. 1990. Development of optimal fixed bed catalytic reaction systems. *Chemical Engineering Progress* 86(2): 32–39. American Institute of Chemical Engineers, New York, February 1990. ©1990 AIChE. All rights reserved.)



*Adiabatic packed-bed reactors* are attractive because of their relatively low cost. (At their simplest, they can be little more than an empty tank filled with catalyst pellets.) However, they suffer from the effects of an uncontrolled reaction temperature and tend to drift towards the maximum possible attainable temperature, which usually means that the limiting reactant is completely consumed or that equilibrium has been reached. This is a desirable property for some reactions (e.g., hydrogenations), but usually the chemical selectivity suffers as a result. Unwanted

by-products may result, even in hydrogenation systems. The adiabatic reactor usually is not recommended where temperature control is important.

*Multiphase packed-bed reactors* (e.g., trickle beds) are widely used, but the transport of reactants and products between the flowing gas and liquid and the solid catalyst is a major uncertainty. A lot of work remains to be done before the rational design of these reactors can be undertaken with confidence. Instead, back-and-forth experimentation and mathematical analysis of the hydrodynamics is necessary. Suffice it to note here that the transport characteristics of multiphase reactors are different in each of the several hydrodynamic regimes that may be experienced (e.g., trickle flow, bubbling flow, pulsing flow, slug flow, mist flow). A priori calculations using relationships from the open literature can often predict whether reaction kinetics or mass transfer will limit the reactor's performance, and the preliminary design sometimes can be developed rationally thereafter. Pilot-scale experimental verification of the performance of the final design in the *same hydrodynamic region* as the plant reactor is nonetheless essential to avoid surprises. There is an abundant literature on this type of reactor, and the excellent text by Ramachandran and Chaudhari [1983] is a good place to obtain an introduction to this complex technology.

*Fluidized-bed reactors* are widely used in fluidized catalytic cracking, the manufacture of polyethylene and polypropylene, the manufacture of silicones from silicon, and some other commercial reactions. The fluid and solid dynamics of these systems are extremely complex, and available models are best described as learning models, rather than predictive models for reaction system design. Scaleup is typically done incrementally in a series of pilot-plant reactors. Despite the well-known advantages of these systems for some purposes (excellent temperature control, absence of diffusional and transport restrictions), the decision to use fluidized systems for new applications must not be taken lightly. Although the solid phase may usually be assumed to be well mixed, the residence-time distribution of the fluid phase is complex and largely unpredictable. It follows that the performance of a scaled-up reactor is also largely unpredictable if the fluid-phase residence-time distribution is important.

It is worth noting that, with few exceptions, all successful commercial fluidized bed reaction systems involve a solid phase that is always in some kind of rapid transition. Fluidized catalytic-cracking catalyst becomes coked and inactive in only 3 to 5 seconds and is continuously removed from the fluidized riser reactor and regenerated by burning in dilute oxygen. The solid phase is in fact the product polyolefin in fluidized systems like Union Carbide's Unipol™ polyolefins process. Dimethyldichloro silane (an intermediate in the manufacture of silicone oils and other products) is manufactured by reacting silicon metal and methyl chloride in a fluidized-bed reactor. Here, the silicon reactant is in the form of a fine metal powder that is continually consumed by the reaction and that comprises the solid phase in the reactor. An exception to the above observation is the Badger process for the manufacture of phthalic anhydride in a fluidized-bed reactor, in which the solid is a catalyst that is not undergoing any rapid change. It is carried out in a fluidized bed because the temperature control is important to the process.

The technology of fluidization has been studied in depth since before World War II and continues to be studied still because of its importance in those industries that depend upon it. AIChE has published several volumes on fluidization in its symposium series, and more appear



periodically. Some excellent texts exist on fluidization phenomena; that by Kunii and Levenspiel [1969] is considered a classic.

*Slurry reactors* are attractive for many gas-liquid systems that are catalyzed by solid catalysts, but their design is nearly as complicated as for fluidized-bed systems. There is an extensive open literature on this type of reactor. The Air Products Company has published a number of reports on work done under contract for the U.S. Department of Energy and has a large pilot plant at LaPorte, Texas, for the manufacture of methanol and higher alcohols from CO and hydrogen. These reports and the text on multiphase reactors by Ramachandran and Chaudhari [10] are recommended.

## Step 2: Design the Laboratory to Generate Reaction Kinetics Data

The second step in the protocol for rational design is to design the laboratory reaction system specifically to generate kinetics rate data. By now it should be clear that the kinetic model is the link between the laboratory operations and the large-scale plant design. The experimental reactor for kinetics studies will ordinarily not look at all like the final plant reactor. Rather, it will be designed to obtain the kinetic data necessary for kinetic model development. Three primary types of laboratory reactor are suitable for design-quality kinetics studies:

- The batch reactor
- The long-tube reactor with sample points along its length
- The continuous stirred tank reactor, or CSTR

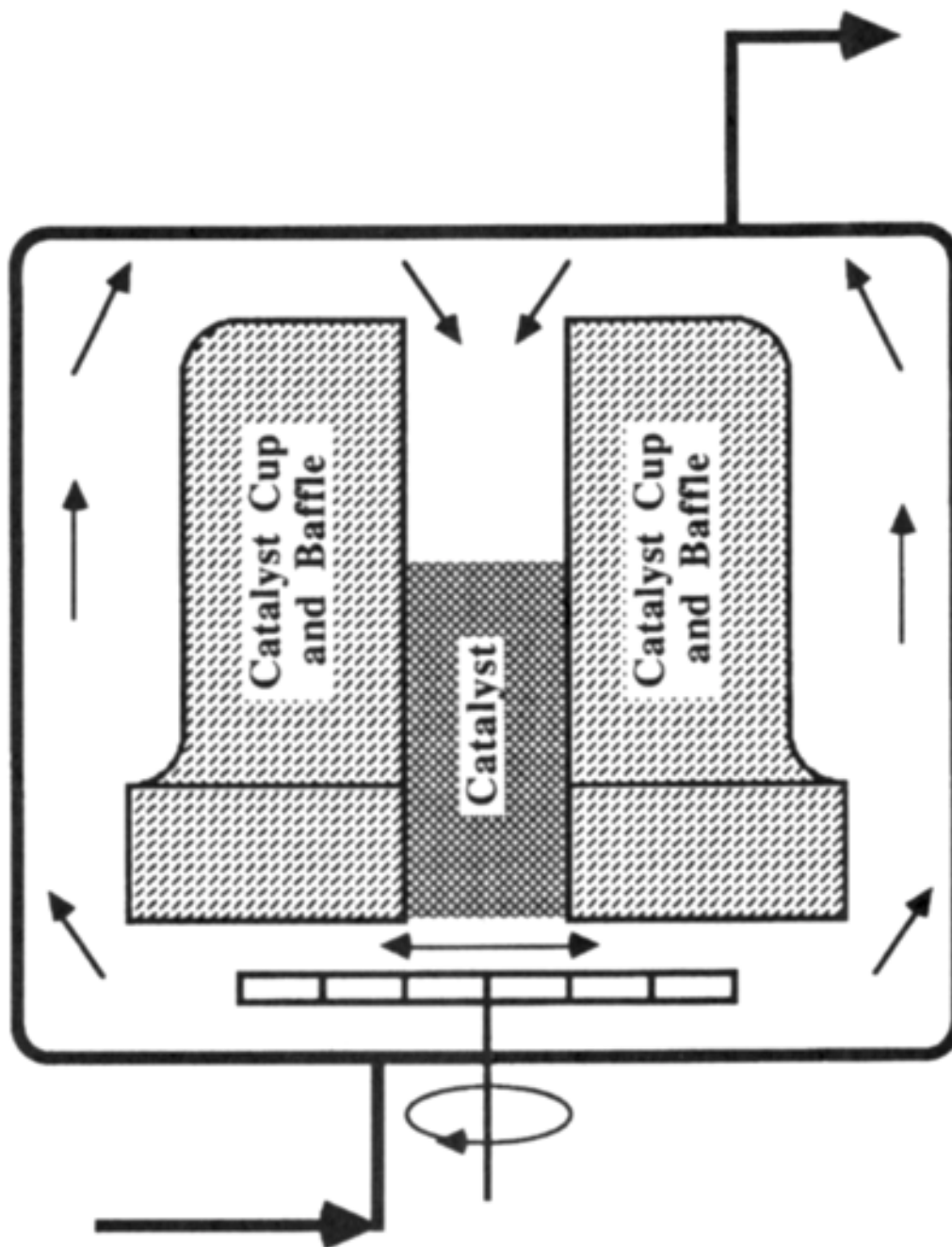
The following paragraphs discuss each of these types, with the emphasis upon what kinds of reactions are suitable for each and any special considerations that may apply.

*Laboratory batch reactors* are found in virtually limitless variety in most industrial chemical laboratories. For kinetics studies it is imperative that chemical analyses be obtained at several times during the reaction period. This sometimes poses problems if the reaction starts before the desired reaction temperature has been reached or if the reaction proceeds so quickly that multiple samples are not practical. Likewise, the loss of reacting volume due to sampling may be important, particularly if a solid catalyst is involved. Both of these problems are best handled during mathematical analysis of the data by forcing the simulated time-temperature and time-reacting volume profiles to be the same as the observed ones.

Batch laboratory reactors are suitable for either uncatalyzed or catalyzed liquid-phase or gas-liquid reactions. In special situations they may be useful for gas-phase reactions as well. They are especially useful for slurry-catalyzed reactions, but they can also be used with pelleted catalysts, provided that provision is made to retain the pellets in a basket or container through which the fluid passes. Both the Berty reactor [see Fig. 77.4 and (Berty, 1974)] and the Carberry rotating basket reactor [Levenspiel, 1972] have been widely used for batchwise kinetic studies of catalytic reactions, although their best use is probably in continuous gas-phase kinetics studies.

**Figure 77.4** The Berty reactor for experimental catalyst testing and kinetics. (Source: Reproduced with permission of the American Institute of Chemical Engineers from Cropley, J. B. 1990. Development of optimal fixed bed catalytic reaction systems. *Chemical Engineering Progress* 86(2): 32–39. American





*Long-tube reactors* for kinetics studies will typically be of a length to promote both plug flow characteristics and good mass and heat transfer so that these physical processes do not mask the chemical reaction rates. Tube length per se is not critical, although it is highly desirable that the length-to-diameter ratio be at least 100:1 to avoid the effects of axial mixing and departure from plug flow behavior. Tube diameter is not particularly important, but should be small enough to ensure good heat transfer. If solid-catalyst pellets are to be used, the tube diameter should be no more than four or five pellet diameters to avoid radial temperature gradients. It is important that provision be made for samples to be taken at multiple points down the tube length. Reactors like these are particularly well suited for kinetics studies using plant-scale catalyst pellets and for the

study of the kinetics of sequential by-product formation, in which the desired product reacts to form unwanted by-products.

Perhaps the most difficult requirement from the standpoint of experimental reactor design is to provide for multiple sampling points down the length of the tube. It is relatively easy to make such a reactor using lengths of stainless steel tubing connected by tubing tees, which can then be used simultaneously as sample taps and for the insertion of thermocouples. Tubing diameter will typically be between 0.25 and 0.50 inches, and the length will be five to six feet. Sample taps at 12-inch intervals will provide good composition and temperature profiles. Excellent temperature control can be attained in such a reactor if it is immersed in a thermostated heat transfer fluid or fluidized sand-bath heater.

*Continuous stirred tank reactors (CSTRs)* are ideally suited for kinetics studies for many types of reactions, owing largely to the ease with which reaction rates can be measured directly. Kinetics models are readily developed from CSTR data, as discussed by Cropley [1978]. The Berty and Carberry reactors cited previously are especially useful for continuous studies of catalytic kinetics, using real catalyst pellets. There is an abundant literature on their use; Cropley [1990] describes an overall strategy for their use in catalytic reactor design.

### Step 3: Use Statistically-Valid Experimental Programs and Models

The third element of the protocol is to utilize statistically valid experimental programs and data analysis for kinetic model development. Usually, kinetic data should be generated from statistically designed experimental programs, such as the factorial or central composite design, for which an abundant literature is available. The writings of Hendrix [1979] are recommended. Cropley [1978] describes the heuristic development of both exponential and hyperbolic kinetic models from a statistically designed data set. Table I from that article is an example of a central composite statistical design for the study of the kinetics of a fictitious catalytic reaction (the oxidation of Dammitol to Valualdehyde). The experimental reactor was assumed to be a CSTR like the Berty reactor, which permitted kinetic rates to be observed directly, as in Eq. (77.7). The synthetic data in that table were developed from the following "true" model, with the incorporation of 20% normally distributed random error [reproduced with permission of the American Chemical Society from (Cropley, 1978)].

$$r_{\text{Val}} = \frac{4.67(10^{11})e^{-20000/RT} (P_{\text{O}_2})^{0.5} (P_{\text{Dam}})^{1.0}}{1 + 5.52(10^{-4})e^{5000/RT} (P_{\text{O}_2})^{0.5} + 7.64(10^{-4})e^{5000/RT} (P_{\text{Val}})^2} \text{ Gmols/Kgcat/H}$$

(77.13)

If a CSTR is used as the experimental reactor, the experimental design should include the independent control of any product species that might influence the reaction rates to avoid systematic bias in the kinetic parameters, as discussed by Cropley [1987].

In the heuristic study cited above, log-linear multiple regression was used for the development of exponential models, and the Nelder-Mead nonlinear search algorithm [Nelder and Mead,

1965] was used for estimation of the nonlinear parameters in the hyperbolic models. (Although Nelder-Mead tends to converge more slowly than some other algorithms, it is robust, stable, and reliable for nonlinear estimation and optimization studies.)

## **Step 4: Develop Computer Programs for Reactor Simulation and Design**

The fourth element of the protocol is to develop computer programs for reactor simulation and design. The typical reaction system design model will comprise several dozen differential equations and a number of nonlinear constraints as well. Consider very carefully just how complex the final models should be. The determination of the true economic optimum may require that the reactor be simulated iteratively many thousands of times, and the model at this point should be no more complex than necessary in order to minimize computer time. Therefore, simulation models for reactor design and optimization should in most cases be based primarily upon the kind of simple idealized reactors represented by Eqs. (77.3)–(77.12), as discussed earlier. Even so, simulation models can become unwieldy. For example, a five-stage CSTR model for a system of four chemical components and temperature will comprise 25 nonlinear differential equations similar to Eq. (77.8). Any of a number of numerical integration techniques may be used for the solution of the differential equations; the fourth-order Runge-Kutta algorithm is perhaps the most common.

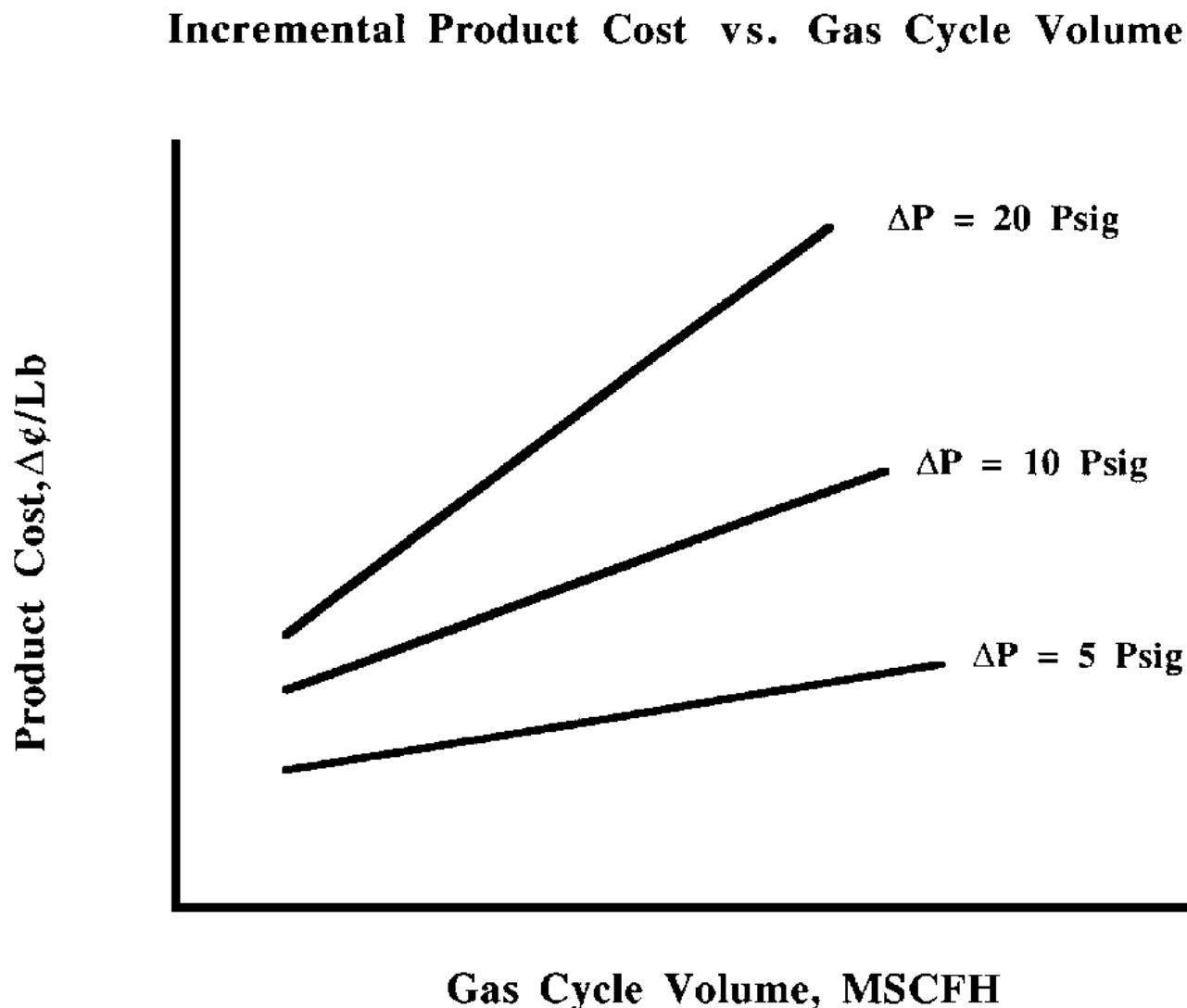
## **Step 5: Develop the Economically Optimum Reactor Design**

The fifth protocol for rational design is to develop economically optimum reactor designs. The potentially useful reactor types should be compared to one another at their individual economic optimum design and operating conditions. Arrive at these for each type by varying the simulated reactor size and geometry (e.g., diameter, length and number of tubes, or agitator horsepower and vessel diameter and height) and operating conditions (like pressure, flow rate, reactant concentration, coolant temperature, and so on). A suitable optimization procedure, such as the Nelder-Mead algorithm cited previously [Nelder and Mead, 1965], should be used.

### **The Objective Function for Optimization**

The design process will include the computation of objective economic or reactor performance criteria for the comparison of optimized design alternatives. This objective function must be reasonably accurate, yet simple enough to be evaluated easily and quickly for each iteration. It is convenient and quite accurate to use simple linear relationships for optimization criteria like that in Fig. 77.5, which illustrates the dependence of incremental product cost on cycle gas flow rate and catalyst bed pressure drop for a hypothetical process. Several similar relationships can comprise an easily used objective function for optimization. Relationships like this can be developed easily from detailed economic analyses of a small number of base case designs. Their development was discussed by Cropley [1990].

**Figure 77.5** Elements of product cost. (Source: Reproduced with permission of the American Institute of Chemical Engineers from Cropley, J. B. 1990. Development of optimal fixed bed catalytic reaction systems. *Chemical Engineering Progress* 86(2): 32–39. American Institute of Chemical Engineers, New York, February 1990. ©1990 AIChE. All rights reserved.)



### Explicit and Implicit Constraints

The optimization package will normally utilize both explicit and implicit constraints to relate the reactor to the rest of the process. In a typical design, for example, reactor inlet pressure and reactor gas flow rate might be *explicitly* constrained to reasonable ranges. Likewise, the maximum reactor temperature might be *implicitly* constrained not to exceed a stipulated maximum. Constraints like these are at the very heart of the optimization process. At the same time, their formulation is both esoteric and beyond the scope of this article, and they are not extensively treated in the open literature. It is recommended that an optimization or numerical analysis specialist be consulted in their development.

## Step 6: Validate the Design in a Pilot Plant Reactor

The final step in the protocol for rational design is to validate the overall optimum design by operation of a pilot reactor system. This is the *only time* that scaleup, per se, will be considered in the design. Here again, this may not necessarily involve a pilot reactor that is geometrically similar to the final plant design, although it may be in some cases. It will be important to design the pilot reactor to be able to confirm the predicted reaction yields and selectivities under design conditions. It is vital that all important recycle streams be incorporated into the pilot plant system—which will then be, to the extent possible, a scaled-down version of the integrated commercial process. Typically, recycle streams include previously unreacted raw materials, by-products that can be reverted to useful product, or potential pollutants that can simply be destroyed by further reaction in the reactor. But trace components and minor species can build up in recycle systems to many times their single-pass concentrations, and they may have adverse effects on catalyst life, equilibrium conversion, crude product quality, and so on. It is important that the effects of recycled species be incorporated into the kinetics model for operational monitoring and control. Unanticipated effects of recycle streams probably account for a significant fraction of scaleup problems in commercial systems.

Finally, please note that this chapter has not discussed the *size* of the pilot plant reactor—size per se really is not an issue. What *is* important is that it function in such a way as to test the rational design before it is built. More discussion of process optimization and validation can be found in Cropley [1990].

## Nomenclature

### Uppercase Symbols

$A, B$	Reactants
$C_k$	Concentration of $k$ th chemical component, in mols/volume units
$D_t$	Tube diameter, in linear units
Dam	Fictitious component Dammitol
$F_e, F_c$	Flow rate of process fluid ( $e$ ) or of coolant ( $c$ ) through the volume element, in volume/time units
$\Delta H_j$	Heat of reaction of $j$ th reaction in heat/mol units
$K_{oj}, E_{aj}$	Arrhenius kinetic parameters for reaction $j$
$K_{Aj},$ $K_{Bj},$ $K_{Pkj}$	Kinetic parameters associated with reactants $A$ and $B$ and products $P_k$ for reaction $j$
Mode	(-1) if coolant flows countercurrent to process stream flow (+1) if flows are cocurrent (0) if coolant is isothermal (infinite flow, boiling, etc.)

$P_k$	Reaction products
$R$	Gas constant, typically 1.987 cal/Gmol/K for kinetics models
$T_r, T_c$	Reaction and coolant temperature, respectively
$U$	Overall heat transfer coefficient, in heat/area/time/temperature units
Val	Fictitious component Valualdehyde
$V_e$	An element of volume of any reactor
$W_c$	Weight of catalyst in the reactor, in weight units

## Lowercase and Greek Symbols

$\alpha_{kj}$	Stoichiometric coefficient for species $k$ in reaction $j$
$c_c$	Heat Capacity of coolant stream, in heat/mass/temperature units
$c_p$	Heat capacity of flowing process stream, in heat/mass/temperature units
$n_c$	Total number of products $k$ in the kinetic model
$n_r$	Number of reactions $j$ in the kinetic model
$n_t$	Number of tubes in a tubular reactor
$\rho_p, \rho_c$	Density of process stream ( $p$ ) or coolant ( $c$ ), in mass/volume units
$r$	Kinetic reaction rate, mols/volume/time (or mols/wt. catalyst/time for catalytic reactions)
$t$	Time, in units consistent with $F, r, U$

## Defining Terms

**Chemical reaction:** The chemical transformation of one or more reactant species into one or more chemical products, usually with the evolution or absorption of heat.

**Chemical reactor:** The vessel in which a chemical reaction is conducted.

**Continuous stirred tank reactor:** A well-mixed continuous reactor, characterized by a broad residence time distribution.

**Conversion:** The fraction of a feed component that undergoes chemical transformation in the reactor.

**Plug flow reactor:** A type of reactor in which all entering elements of fluid have the same residence time. The residence time distribution is thus extremely narrow.

**Residence time:** The amount of time that a reactive mixture spends in a chemical reactor.

**Residence-time distribution (RTD):** The spread of residence times that different elements entering a reactor spend in it. RTD is one of the distinguishing characteristics of different reactor types.

**Selectivity:** The fraction of all of a feed component that is converted to form a specified product.

**Yield:** The fraction of all of a feed component entering a reactor that is converted to a specified product. Yield = (Selectivity)(Conversion).

## References

- Berty, J. M. 1974. Reactor for vapor-phase catalytic studies. *Chem. Eng. Prog.* 70: 57–584.  
 Cropley, J. B. 1978. Heuristic approach to complex kinetics. In *Chemical Reaction Engineering* 3/4 Houston, D. Luss and V. Weekman, eds. ACS Symposium Series 65. Paper 24.  
 Cropley, J. B. 1987. Systematic errors in recycle reactor kinetics studies. *Chemical Engineering Pro*

- 50. American Institute of Chemical Engineers, New York.
- Cropley, J. B. 1990. Development of optimal fixed bed catalytic reaction systems. *Chemical Engineering Progress*, 68(1), 39–45.
- 39. American Institute of Chemical Engineers, New York.
- Froment, G. F. and Bischoff, K. B. 1979. *Chemical Reactor Analysis and Design*. John Wiley & Sons, New York.
- Hendrix, C. D. 1979. What every technologist should know about experimental design. *Chemtech*, 9(1), 17–19.
- 174. American Chemical Society, Washington, DC.
- Kunii, D. and Levenspiel, O. 1969. *Fluidization Engineering*. John Wiley & Sons, New York.
- Levenspiel, O. 1972. *Chemical Reaction Engineering*, 2nd ed. John Wiley & Sons, New York.
- Nelder, J. A. and Mead, J. R. 1965. A simplex method for function minimization. *Computer Journal*, 8(3), 315–318.
- Ramachandran, P. A. and Chaudhari, R. V. 1983. *Three-Phase Catalytic Reactors*. Gordon and Breach Science, New York.

## Further Information

### Professional Organizations

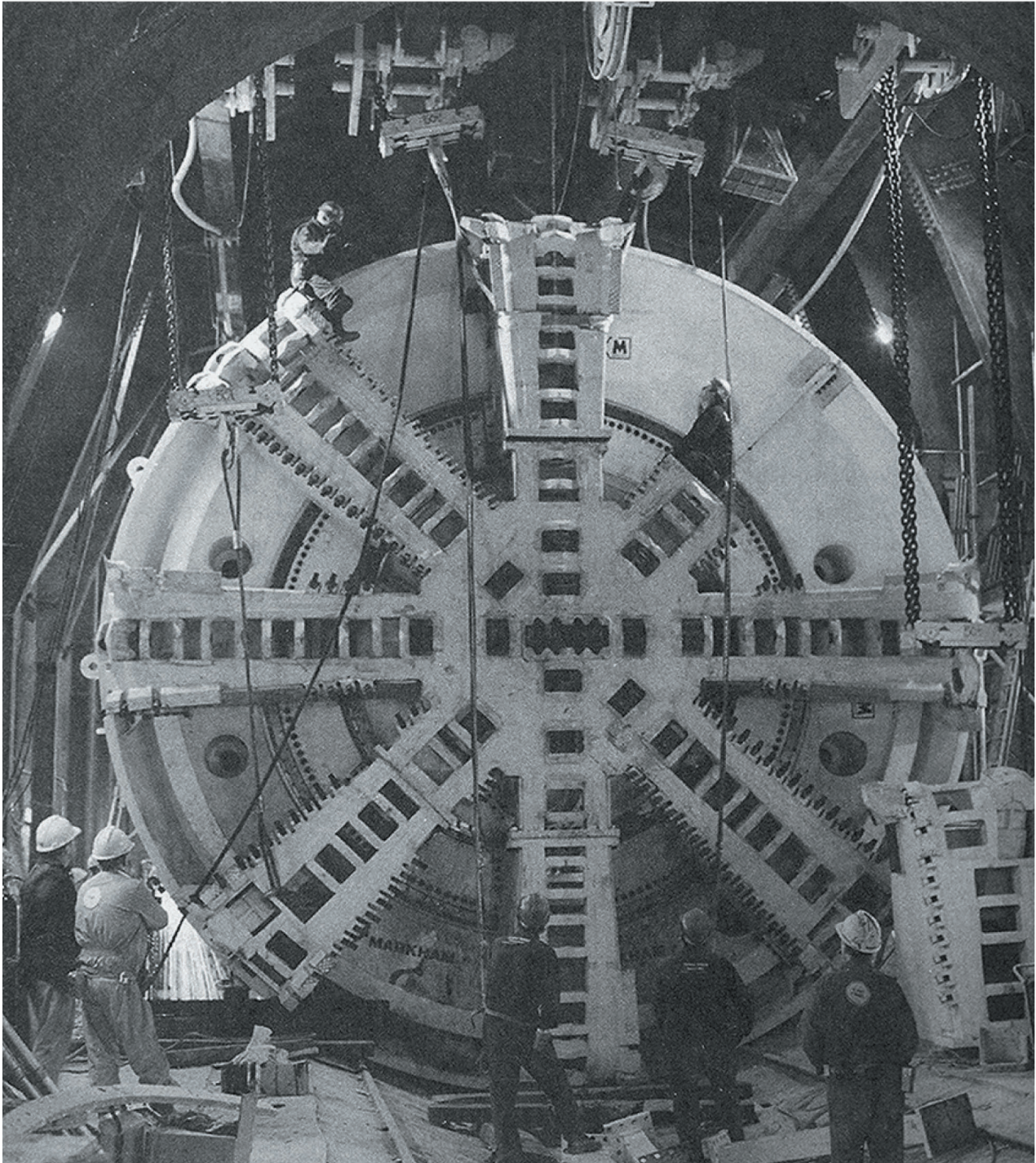
- The American Institute of Chemical Engineers (three national meetings per year). 345 E. 47th St., New York, NY 10017. Phone (212) 705-7322.
- The American Chemical Society (several national/regional meetings each year). 1155 16th St., N.W., Washington, DC, 20036. Phone (202) 872-4600.

### Special Meetings and Conferences

- The Engineering Foundation Conferences on Chemical Reaction Engineering. 345 East 47th Street, New York, NY 10017. Phone (212) 705-7835.
- International Symposia on Chemical Reaction Engineering (even years). Sponsored by American Institute of Chemical Engineers, American Chemical Society, Canadian Society for Chemical Engineering, and the European Federation of Chemical Engineering.

Harr, M. E. "Geotechnical"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000





This tunnel-boring machine (TBM) was one of the largest built to bore the 24 miles (38 km) under the English Channel for the Eurotunnel project. Today the Eurotunnel is a massive transportation system which links Britain and France.

This TBM is 28.17 ft (8.78 m) in diameter and weighs 1575 tonnes. The teeth of the TBM's revolving cutting heads are made of a very hard material called tungsten carbide. This allowed the TBM to cut its way through the chalk marl at a rate of up to 0.62 miles (1 km) per month. At full speed the TBM was pumping out 2400 tonnes of spoil per hour on the British side of the tunnel. A total of 8 million cubic meters of spoil was removed and disposed of. It required incredible geotechnical ingenuity to build the Eurotunnel, which now transports thousands of vehicles and tens of thousands of people each day. (Courtesy of ©Eurotunnel 1994. Photo by QA Photos, Hythe. Published with permission.)

# XI

## Geotechnical

---

**Milton E. Harr**

*Purdue University*

### 78 **Soil Mechanics** *B. M. Das*

Weight–Volume Relationship Hydraulic Conductivity Effective Stress Consolidation Shear Strength

A SOIL BODY, IN ITS GENERAL FORM, is a complex conglomeration of discrete particles, in a compact array of varying shapes and orientations. These may range in magnitude from the microscopic elements of clay to the macroscopic boulders of a rock fill. Soil mechanics deals with the action and reaction of soil bodies when acted upon by energy sources.

Today the world is in the midst of a scientific revolution that requires new types of structures for which past experience is either inadequate or absent. Earth structures of unprecedented height and size are becoming commonplace engineering problems. Confronted with these problems, engineers are finding it necessary to rely on rational, scientific methods of analysis with correspondingly less emphasis on empirical rules.

In an engineering handbook it is difficult to present the subtle details in such a specialized area as soil mechanics. However, Dr. Das has covered some of the fundamentals quite well.

Das, B. M. "Soil Mechanics"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

78.1 Weight–Volume Relationship

78.2 Hydraulic Conductivity

78.3 Effective Stress

78.4 Consolidation

Time Rate of Consolidation

78.5 Shear Strength

**Braja M. Das**

*California State University, Sacramento*

Soil mechanics is the branch of science that deals with the study of the physical properties of soil and the behavior of soil masses while being subjected to various types of forces. Soils engineering is the application of the principles of soil mechanics to practical problems.

## 78.1 Weight- Volume Relationship

---

The three phases of a soil sample are solid, water, and air, as shown in [Fig. 78.1](#). Thus, a given soil sample of weight  $W$  can be expressed as

$$W = W_s + W_w + W_a \quad (78.1)$$

where  $W_s$  is the weight of the soil solids,  $W_w$  is the weight of the water, and  $W_a$  is the weight of air. Assuming that  $W_a \approx 0$ ,

$$W = W_s + W_w \quad (78.2)$$

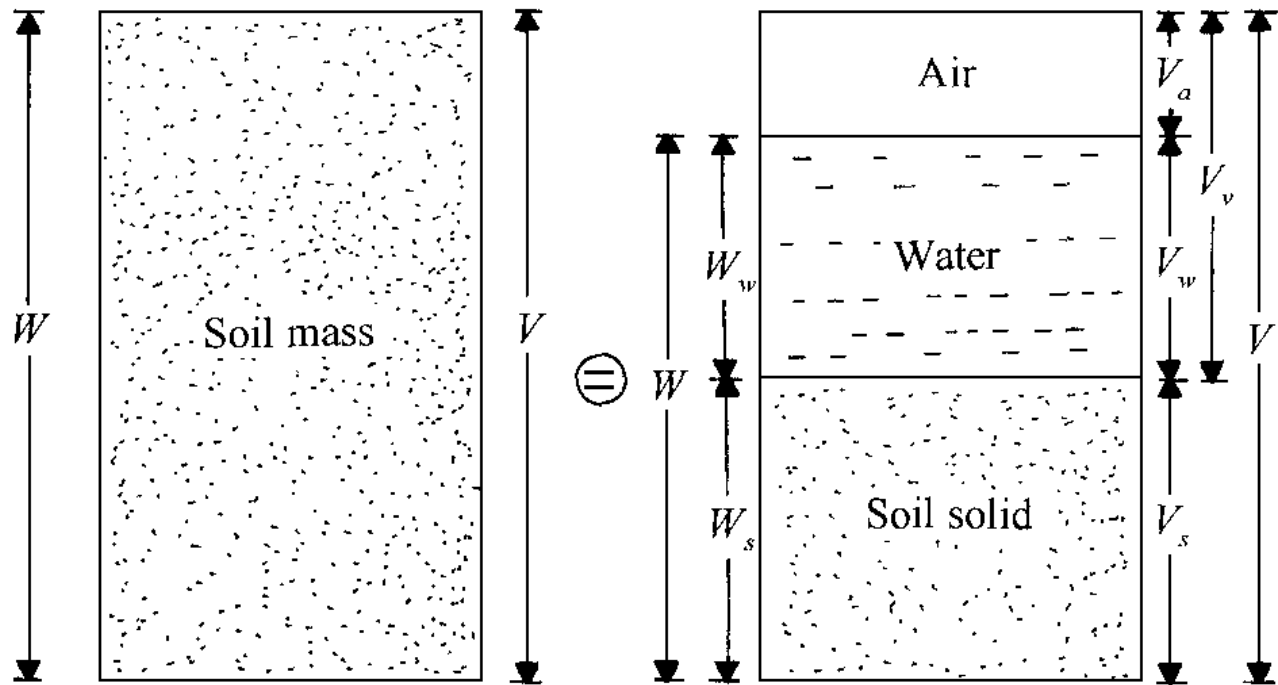
As shown in [Fig. 78.1](#), the total volume of the soil sample is  $V$ . The volumes occupied by solid, water, and air are, respectively,  $V_s$ ,  $V_w$ , and  $V_a$ . Thus, the volume of void,  $V_v$ , is

$$V_v = V_w + V_a \quad (78.3)$$

Common weight and volume relationships are given in [Table 78.1](#).



**Figure 78.1** Weight-volume relationship.



**Table 78.1** Weight-Volume Relationships

Volume Relationships	
Void ratio, $e = \frac{V_v}{V_s} = \frac{n}{1 - n}$	
Porosity, $n = \frac{V_v}{V} = \frac{e}{1 + e}$	
Degree of saturation, $S = \frac{V_w}{V_v} = \frac{wG_s}{e}$	
Weight Relationships	
Moisture content, $w = \frac{W_w}{W_s}$	
Moist unit weight, $\gamma = \frac{W}{V}$	
$\gamma = \frac{(1_w)G_s\gamma_w}{1 + e}$	
$\gamma = \frac{(G_s + Se)\gamma_w}{1 + e}$	
$\gamma = G_s\gamma_w(1 - n)(1 + w)$	
Dry unit weight, $\gamma_d = \frac{\gamma}{1 + w}$	
$\gamma_d = \frac{G_s\gamma_w}{1 + e}$	
$\gamma_d = G_s\gamma_w(1 - n)$	
$\gamma_d = \frac{G_s\gamma_w}{1 + wG_s/S}$	

$$\begin{aligned}\text{Saturated unit weight, } \gamma_{\text{sat}} &= \frac{(G_s + e)\gamma_w}{1 + e} \\ \gamma_{\text{sat}} &= [(1 - n)G_s + n]\gamma_w \\ \gamma_{\text{sat}} &= \gamma_d + n\gamma_w\end{aligned}$$

$G_s$  = Specific gravity of soil solids

$\gamma_w$  = Unit weight of water (62.4 lb/ft<sup>3</sup>; 9.81 kN/m<sup>3</sup>)

## 78.2 Hydraulic Conductivity

In 1856, Darcy published a relationship for the discharge velocity of water through saturated soil (usually referred to as Darcy's law), according to which

$$v = ki \quad (78.4)$$

where  $v$  is the discharge velocity,  $i$  is the **hydraulic gradient**, and  $k$  is the *coefficient of permeability* or *hydraulic conductivity*. The unit of hydraulic conductivity is  $LT^{-1}$ . See [Table 78.2](#).

**Table 78.2** Typical Range of Hydraulic Conductivity

Soil Type	$k$ (cm/s)	Relative Hydraulic Conductivity
Clean gravel	$10^2 - 10^0$	High
Coarse sand	$10^0 - 10^{-2}$	High to medium
Fine sand	$10^{-2} - 10^{-3}$	Medium
Silt	$10^{-3} - 10^{-5}$	Low
Clay	Less than $10^{-6}$	Very low

One of the most well-known relationships for hydraulic conductivity is the Kozeny-Carman equation, which is of the form

$$k = \frac{1}{\eta} \frac{1}{s_p t^2 s_s^2} \frac{n^3}{(1 - n)^2} \quad (78.5)$$

where  $\eta$  is viscosity,  $s_p$  is pore shape factor,  $t$  is tortuosity,  $s_s$  is specific surface per unit volume, and  $n$  is porosity. [Table 78.2](#) gives the general magnitude of the hydraulic conductivity.

## 78.3 Effective Stress

In saturated soil the **total stress** in a given soil mass can be divided into two parts—a part that is carried by water present in continuous void spaces, which is called *pore water pressure*, and the remainder that is carried by the soil solids at their points of contact, which is called the *effective*

stress. Or

$$\sigma = \sigma' + u \quad (78.6)$$

where  $\sigma$  is total stress,  $\sigma'$  is effective stress, and  $u$  is pore water pressure.

In partially saturated soil, water in the void space is not continuous, and the soil is a three-phase system—that is, solid, pore water, and pore air. In that case,

$$\sigma = \sigma' + u_a - \chi(u_a - u_w) \quad (78.7)$$

where  $u_a$  is pore air pressure and  $u_w$  is pore water pressure.

## 78.4 Consolidation

---

Consolidation is the *time-dependent* volume change of saturated clayey soil due to the expulsion of water occupying the void spaces. When a load is applied to a saturated compressible soil mass, the increase in stress is initially carried by the water in the void spaces (that is, increase in pore water pressure) due to its relative incompressibility. With time, the water is squeezed out and the stress increase is gradually transferred to effective stress. The effective stress increase results in consolidation settlement of the clayey soil layer(s).

A clay is said to be *normally consolidated* when the present *effective* overburden pressure ( $p_o$ ) is the maximum pressure to which the soil has been subjected in the past. When the present effective overburden pressure of a clay is less than that which it experienced in the past, it is referred to as an *overconsolidated* clay. The maximum past effective overburden pressure is called the *preconsolidation pressure* ( $p_c$ ).

The *primary consolidation* settlement ( $S$ ) of a saturated clay layer of thickness  $H$  (Fig. 78.2) due to a load application can be calculated by the following relationships. In Fig. 78.2, for the clay layer,  $p_o$  is the effective overburden pressure before the load application,  $\Delta p$  is the increase in stress at the middle of the clay layer due to the load application, and  $e_o$  is the initial void ratio. For normally consolidated clay ( $p_o = p_c$ ),

$$S = \frac{C_c H}{1 + e_o} \log \left( \frac{p_o + \Delta p}{p_o} \right) \quad (78.8)$$

For overconsolidated clay with  $p_o + \Delta p \leq p_c$ ,

$$S = \frac{C_s H}{1 + e_o} \log \left( \frac{p_o + \Delta p}{p_o} \right) \quad (78.9)$$

For overconsolidated clay with  $p_o < p_c < p_o + \Delta p$ ,

$$S = \frac{C_s H}{1 + e_o} \log \left( \frac{p_c}{p_o} \right) + \frac{C_c H}{1 + e_o} \log \left( \frac{p_o + \Delta p}{p_c} \right) \quad (78.10)$$

where  $C_c$  is the compression index and  $C_s$  is the swelling index. The magnitudes of  $C_c$  and  $C_s$  can be obtained from laboratory consolidation tests. In the absence of the laboratory results the following empirical approximations can be used. From Skempton [1944],

$$C_c = 0.009(LL - 10) \quad (78.11)$$

where LL is liquid limit, in percent. From Nishida [1956],

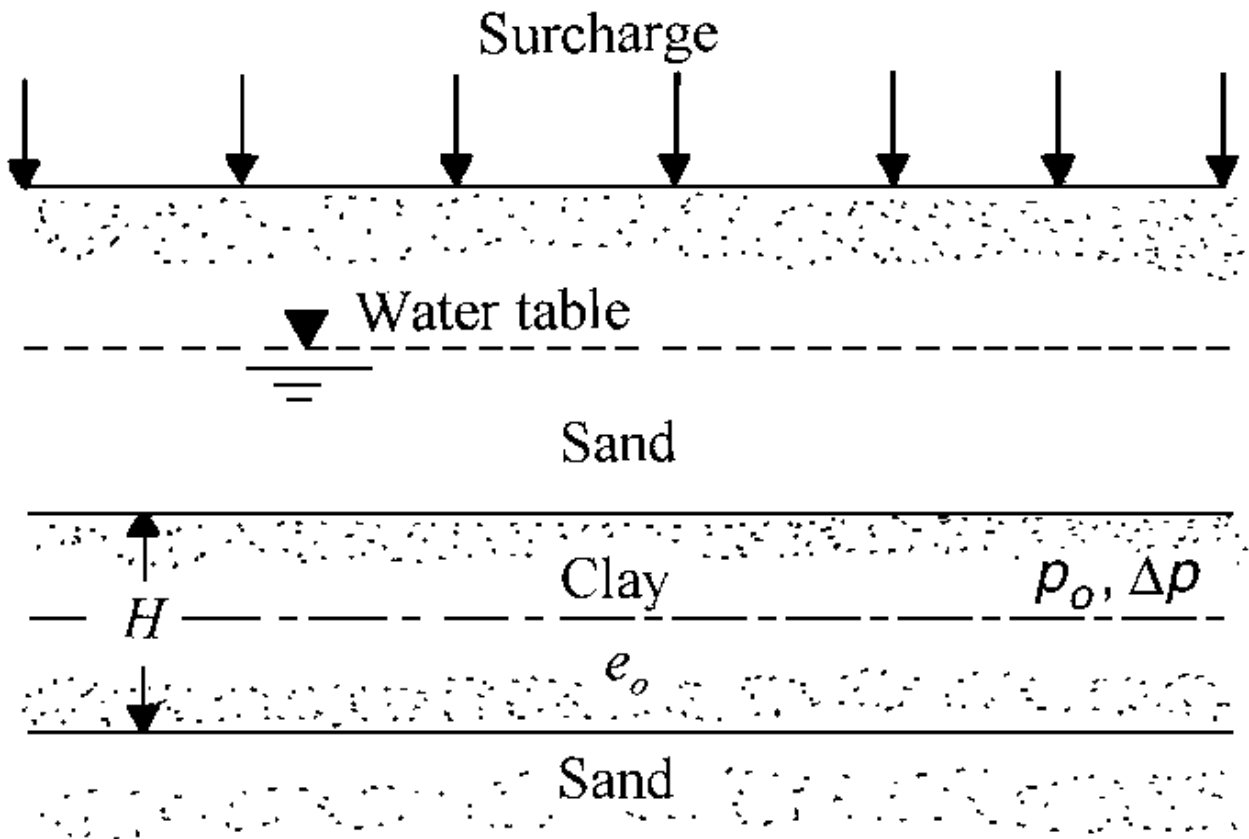
$$C_c = 1.15(e_o - 0.27) \quad (78.12)$$

And, finally, from Rendon-Herrero [1980],

$$C_c = 0.156e_o + 0.0107 \quad (78.13)$$

$$C_s \approx 0.1-0.2 C_c \quad (78.14)$$

**Figure 78.2** Consolidation settlement of a clay layer.





## Time Rate of Consolidation

The average degree of consolidation ( $U$ ) of a clay layer due to an applied load can be defined as

$$U = \frac{S_t}{S} \quad (78.15)$$

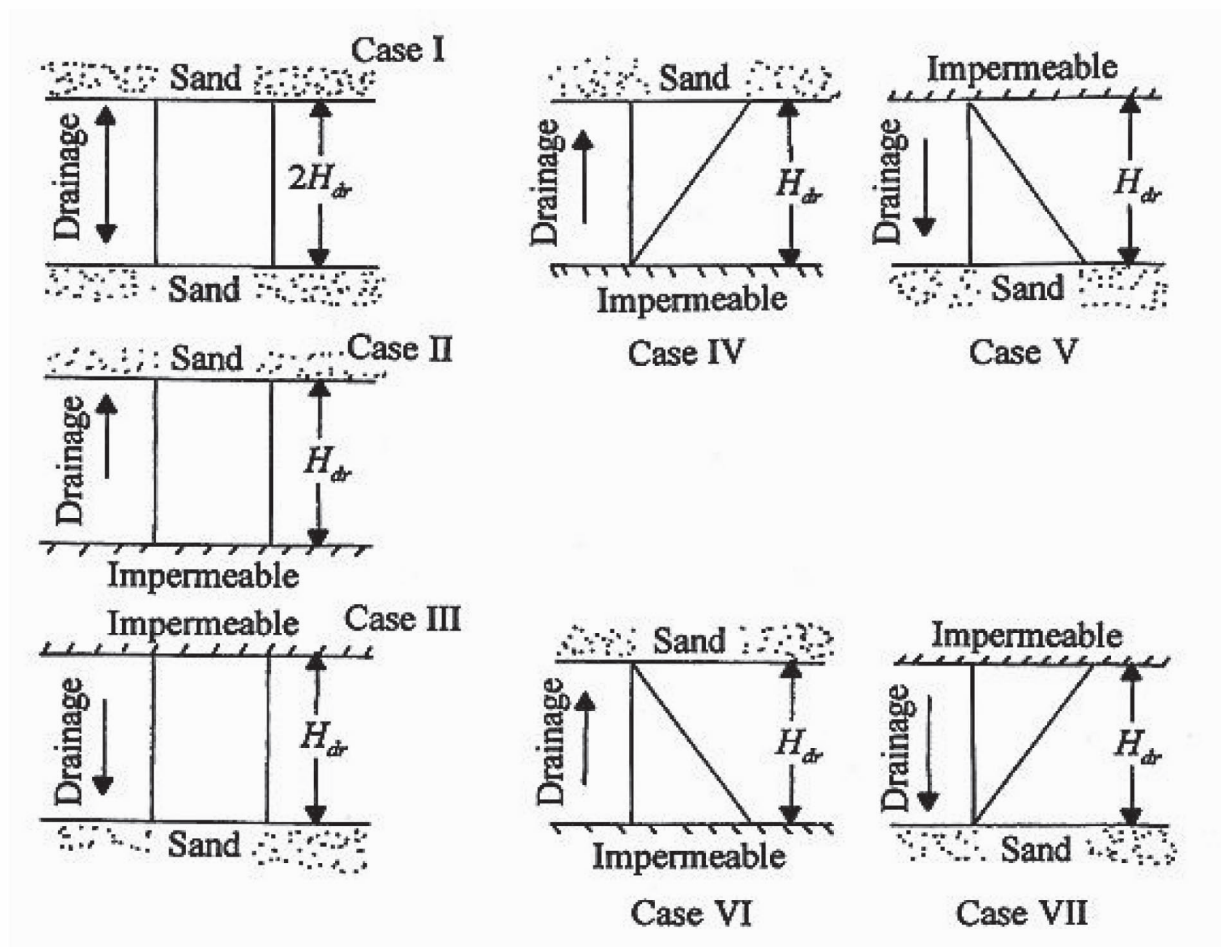
where  $S_t$  is primary consolidation settlement after time  $t$  of load application, and  $S$  is the ultimate consolidation settlement. The degree of consolidation is a function of time factor  $T_v$ , which is a nondimensional quantity. Or

$$U = f(T_v) \quad (78.16)$$

$$T_v = \frac{C_v t}{H_{dr}^2} \quad (78.17)$$

where  $C_v$  is the **coefficient of consolidation**,  $t$  is the time after load application, and  $H_{dr}$  is the length of the smallest drainage path. [Figure 78.3](#) shows the definitions of  $H_{dr}$  and the initial excess pore water pressure ( $u_o$ ) distribution for seven possible cases. The variations of  $U$  with  $T_v$  for these cases are shown in [Table 78.3](#).

**Figure 78.3** Definition of  $H_{dr}$ .



**Table 78.3** Variation of Average Degree of Consolidation With Time Factor

Average Degree of Consolidation, $U$ (%)	Time factor, $T_v$		
	Case I, II, III	Case IV, V	Case VI, VII
0	0	0	0
10	0.008	0.003	0.004 7
20	0.031	0.009	0.100
30	0.071	0.025	0.158
40	0.126	0.048	0.221
50	0.197	0.092	0.294
60	0.287	0.160	0.383
70	0.403	0.271	0.500
80	0.567	0.440	0.665
90	0.848	0.720	0.940
100	$\infty$	$\infty$	$\infty$

## 78.5 Shear Strength

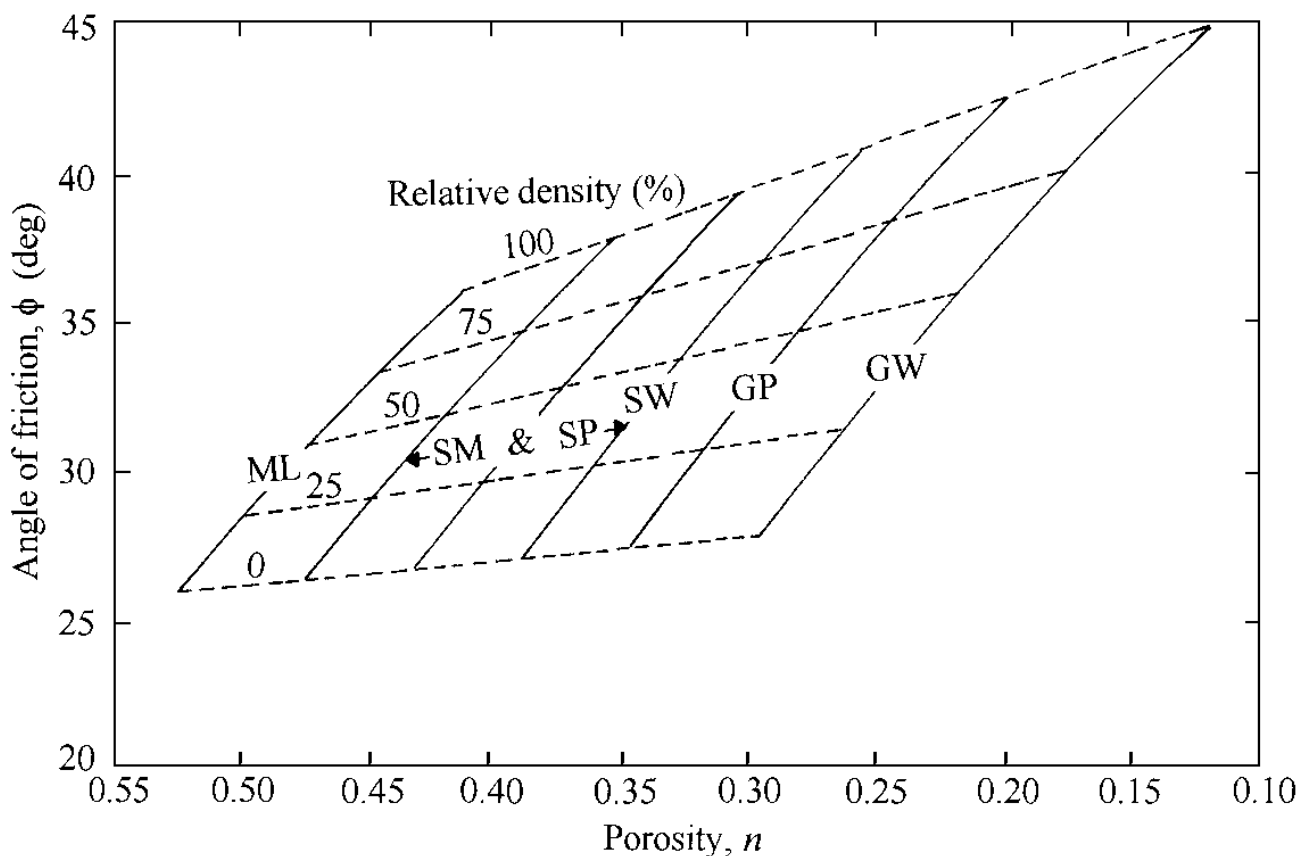
Shear strength of a soil mass is the internal resistance per unit area that the soil mass can offer to resist failure and sliding along any plane inside it. For most soil mechanics problems, it is sufficient to approximate the shear strength by the *Mohr-Coulomb failure criteria*, or

$$s = c + \sigma' \tan \phi \quad (78.18)$$

where  $c$  is the cohesion,  $\sigma'$  is the effective normal stress, and  $\phi$  is the drained friction angle. For normally consolidated clays and sands,  $c \approx 0$ .

Figure 78.4 shows an approximate correlation for  $\phi$  with porosity and **relative density** for coarse-grained soils. Table 78.4 gives typical values of  $\phi$  for sand and silt.

**Figure 78.4** Approximate correlation for  $\phi$  of coarse-grained soil with porosity and relative density. Note: GW—well-graded gravel, GP—poorly graded gravel, SW—well-graded sand, SP—poorly graded sand, SM—silty sand, ML—silt with low plasticity. (After Department of the Navy, 1971.)



**Table 78.4** Typical Values of  $\phi$  for Sand and Silt

Soil	$\phi$ (deg)
Sand	
Loose	28–35
Medium	30–40
Dense	35–45
Silt	25–35

## Defining Terms

**Coefficient of consolidation:** The coefficient of consolidation is defined by the relationship

$$C_v = \frac{k}{\gamma_w \left( \frac{\Delta e}{\Delta p} \frac{1}{1 + e_o} \right)}$$

where  $\Delta e$  is the change in void ratio due to a pressure increase in  $\Delta p$ .

**Hydraulic gradient:** The ratio of the loss of head to the length of flow over which the loss occurred.

**Relative density:** Relative density is defined as  $(e_{\max} - e)/(e_{\max} - e_{\min})$ , where  $e_{\max}$  and  $e_{\min}$  are, respectively, the maximum and minimum possible void ratios for a soil, and  $e$  is the in situ void ratio.

**Total stress:** The total stress at a given elevation is the force per unit gross cross-sectional area due to soil solids, water, and surcharge.

## References

- Darcy, H. 1856. *Les Fontaines Publiques de la Ville de Dijon*. Dalmont, Paris.
- Department of the Navy. 1971. *Soil Mechanics, Foundations, and Earth Structures* 3/4NAVFAC DM-7. U.S. Government Printing Office, Washington, DC.
- Nishida, Y. 1956. A brief note on compression index of soils. *J. Soil Mech. Found. Div., ASCE*. 82(3):1027-1–1027-14.
- Rendon-Herrero, O. 1980. Universal compression index equation. *J. Geotech. Engr. Div., ASCE*. 106(11):1179–1200.
- Skempton, A. W. 1944. Notes on the compressibility of clays. *J. Geol. Soc. Lond.* 100:119–135.

## Further Information

Das, B. M. 1994. *Principles of Geotechnical Engineering*, 3rd ed. PWS, Boston, MA.

Sinha, K, C. "Transportation"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000





This aerial view of the Eurotunnel is located near Folkestone in England. The Eurotunnel, also called the

Channel Tunnel, stretches from the UK under the English Channel to Calais, France. The overall tunnel length is 31.35 miles (50.45 km); 24 miles (38 km) are under the sea. The tunnel lies an average of 148 ft (45 m) below the seabed, with a maximum depth of 246 ft (75 m) below the seabed.

The tunnel carries four types of traffic. At full capacity a shuttle carries 800 passengers, 120 cars, and 12 coaches (which could be replaced by an additional 60 cars). Up to 28 heavy-goods vehicles (HGVs) with a gross weight of 44 tonnes can be carried in a freight shuttle.

At peak times, trains thunder along the tunnel every three minutes at up to 100 mph (160 km/h). Thousands of vehicles and tens of thousands of people pass through the tunnel every day. The tunnel is in operation 24 hours a day, seven days a week, 52 weeks a year in virtually all types of weather. See pages 852 and 853 for additional information on the Eurotunnel. (Copyright Eurotunnel 1994. Photo by QA Photos, Hythe. Used with permission.)

# XII

## Transportation

---

**Kumares C. Sinha**

*Purdue University*

**79 Transportation Planning** *M. D. Meyer*

Basic Framework of Transportation Planning • Transportation Modeling

**80 Design of Transportation Facilities** *J. Leonard II and M. D. Meyer*

Components of the Project Development Process • Basic Concepts of Project Design • Intermodal Transportation Terminals or Transfer Facilities • Advanced Technology Projects

**81 Operations and Environmental Impacts** *P. W. Shuldiner and K. B. Black*

Fundamental Equations • Flow, Speed, and Density Relationships • Traffic Measurements • Level of Service (LOS) • Highway Capacity • Intersection Capacity • Traffic Control Devices • Stop Sign Warrants • Traffic Signal Warrants • Air Quality Effects

**82 Transportation Systems** *P. Schonfeld*

Transportation System Components • Evaluation Measures • Air Transportation • Railroad Transportation • Highway Transportation • Water Transportation • Public Transportation

**83 Safety Analysis** *T. B. Khalil*

Mathematical Models • Summary

AS TRANSPORTATION IS CONCERNED with the movement of people and goods, it has a far-ranging impact on society. It plays a vital role in a country's economy. In the U.S. transportation makes up about 20% of GDP. Transportation engineering cuts through a broad array of disciplines to deal with planning design, construction, maintenance, and operation of various transportation modes. While it has drawn strong contributions from civil, mechanical, and electrical engineering, social, political, and management sciences are also becoming important areas in influencing the transportation engineering field. With the immense growth of the transportation sector in the 20th century, the environmental aspects are fast becoming the limiting constraint for transportation systems.

This section presents an overview of transportation engineering with emphasis on planning, facility design, operations and environmental impacts, characteristics of various transport modes, and safety. Transportation planning is concerned with meeting system demand in the most cost-effective manner. Through transportation planning, the formulation of optimal transportation policies can be achieved. The design of transportation facilities implements transportation policies, particularly in the supply side of transportation infrastructure development. Facility design must consider the traditional engineering aspects of design, such as geotechnical, structural, and geometric design as well as legal, environmental, and land use concerns. Once completed, operational aspects of transportation facilities must be continually monitored to provide for the safest and most efficient means of travel without causing environmental degradation. The



fundamental principles of highway traffic flow are presented, the understanding of which is essential to guide, advise, and regulate vehicle operators. The aspect of highway safety is emphasized from an impact-crashworthiness point of view in a separate chapter. Because transportation of people and goods uses various modes, another chapter presents a discussion of different subsystems. Modern-day challenges to the transportation system exist in the form of increasing system efficiency to minimize adverse environmental effects and to maintain economic viability. To meet such challenges, the synthesis of traditional modes of transportation with emerging technologies is required. Alternative forms of energy (e.g., electric vehicles) and support mechanisms (e.g., magnetic levitation) can provide meaningful solutions.

This section presents the key elements involved in the planning, design, and operation of a safe and efficient transportation system. The text applies basic laws of engineering to illustrate transportation-related concepts in a manner understandable to a wide range of readers.

Meyer, M.D. "Transportation Planning"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Transportation Planning

---

## 79.1 Basic Framework of Transportation Planning

Inventory of Facilities • Collect and Maintain Socioeconomic and Land Use Data • Define Goals and Objectives • Identify System Deficiencies or Opportunities • Develop and Analyze Alternatives • Evaluate Alternatives • Implement Plan • Monitor System Performance

## 79.2 Transportation Modeling

**Michael D. Meyer**

*Georgia Institute of Technology*

Transportation planning is the process of producing the information necessary to determine the most cost-effective strategy for improving the performance of the transportation system.

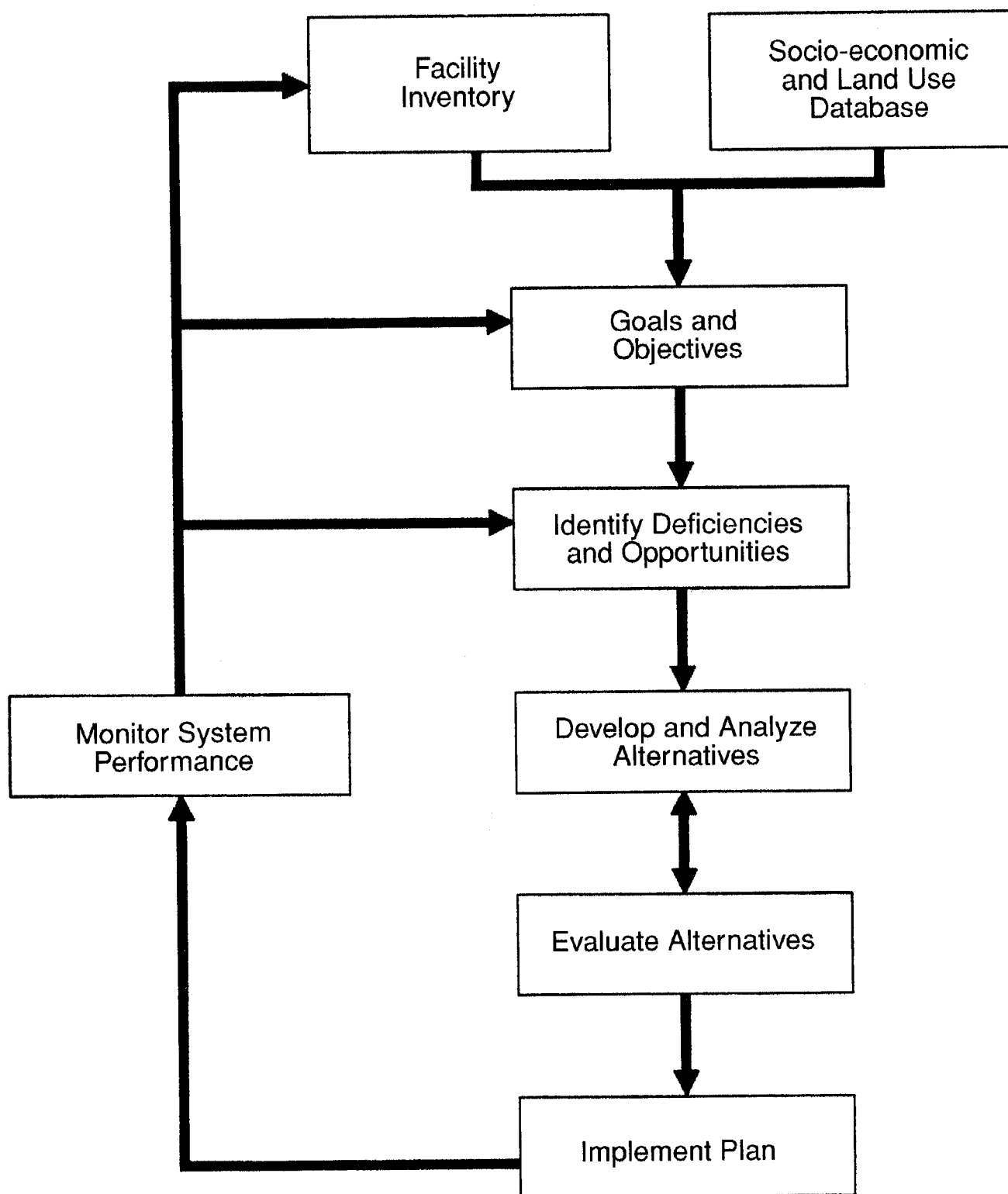
Transportation planning can occur on different scales of analysis—at the national, multistate, state, metropolitan, subarea or corridor, and local levels. At each level the planning methodologies used, databases developed, and information produced are often heavily influenced by government regulations. No matter what modeling approaches are used in transportation planning, a fundamental concept that influences the way transportation demand is estimated is called **derived demand**. Derived demand means that a trip is taken to accomplish some activity at the destination, and that the trip itself is simply a means of reaching this activity. Transportation planning must therefore relate trip making to the types of activities that occur in a study area and also to the characteristics of the trip maker that will influence the way these trips are made. This end is reached by combining similar uses of land into a land use category that can then be used in transportation planning to estimate how many trips are attracted to each type of land use (e.g., the number of trips to schools, shopping centers, residential units, office complexes, etc.).

## 79.1 Basic Framework of Transportation Planning

---

The basic framework for transportation planning at any scale of application is shown in [Fig. 79.1](#). The steps shown in this framework are discussed in the sections that follow.

**Figure 79.1** Basic elements of transportation planning.



## Inventory of Facilities

Knowing what a **transportation network** consists of and the condition and performance of these facilities is an important starting point for transportation planning. Transportation investment is usually aimed at upgrading the physical *condition* of a facility (e.g., repaving a road or building a new bridge) or at improving its *performance* (e.g., providing new person-carrying capacity by providing preferential treatment for high-occupancy vehicles or building a new road to serve existing demand).

## Collect and Maintain Socioeconomic and Land Use Data

Given the concept of derived demand, a basic starting point for planning is characterizing the variables that will influence this demand. Current land use is readily attained through land use inventories. The methods of estimating future land use range from trends analysis to large-scale land use models that predict household and employment sites decades into the future.

Socioeconomic characteristics often include the level of income, number of members in the household, number of autos in the household, number of children, age of the head of household, and highest level of education achieved.

## Define Goals and Objectives

*Goals* are generalized statements that indicate the desired ultimate achievement of a transportation plan. *Objectives* are more specific statements that indicate the means by which these goals will be achieved. The identification of goals and objectives is critical in that they define the evaluation criteria that will be used later in the planning process to assess the relative impacts of alternative projects and strategies. These criteria are often called *measures of effectiveness*.

## Identify System Deficiencies or Opportunities

The methods used to identify transportation deficiencies and opportunities can vary widely. In some cases large-scale transportation network models are used to estimate the future traffic volumes and then to compare them to the capacity of the road network to handle such volumes. This volume-to-capacity (V/C) ratio has served as the major means of identifying the location of system deficiencies. However, other *performance measures* can also be used in the transportation planning process that are much broader than the traditional V/C ratios, such as level of accessibility to economic activities, average travel times, and so on.

## Develop and Analyze Alternatives

Various types of strategies can surface from the planning process: (1) those focused on improvements to highways—for example, adding new lanes, improving traffic control through signals or signing, or improving traffic flow through channelization; and (2) other strategies, such as reducing the demand for transportation through flexible working hours, increasing average vehicle occupancy through such measures as carpools or transit use, or raising the "price" of travel

through the use of tolls. More recently, the application of advanced transportation technologies to the operation of a road system, known as *intelligent transportation systems*, has become an important type of strategy in many cities.

## Evaluate Alternatives

Evaluation brings together all of the information gathered on individual alternatives and provides a framework to compare the relative worth of the alternatives. This evaluation process most often relies on the various measures of effectiveness that relate to the goals and objectives defined at the beginning of the planning process.

## Implement Plan

The major products of the transportation planning process are the *transportation plan* and, at the state and metropolitan levels, the *transportation improvement program (TIP)*, which lists all of the transportation projects that will be implemented over the next few years in the region.

## Monitor System Performance

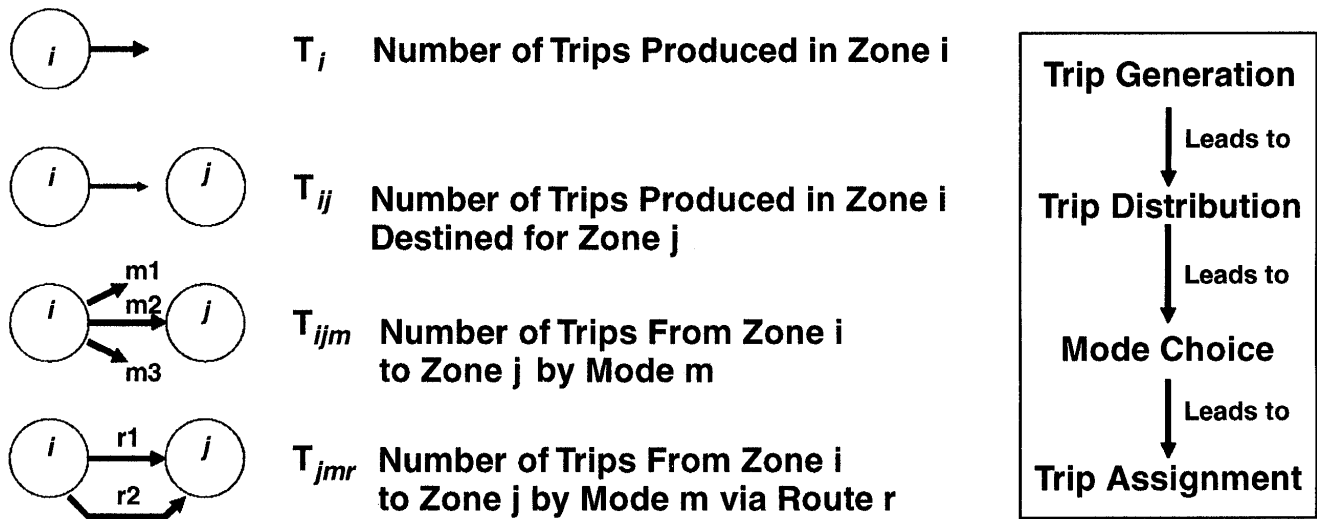
Transportation planning continually examines the performance and condition of the transportation system to identify where improvements can be made. Some means of monitoring system performance is necessary to systematically identify areas where these improvements might occur. This system monitoring has traditionally been based on traffic volumes obtained from traffic counters placed at strategic locations in the network (e.g., increasing traffic volumes indicate congestion).

## 79.2 Transportation Modeling

---

The level of transportation analysis can vary according to the level of complexity and scale of application. In most cases, however, the modeling process consists of four major steps—trip generation, trip distribution, mode split, and trip assignment. Each study area (whether a nation, state, metropolitan area, or neighborhood) is divided into zones of homogeneous characteristics (e.g., similar household incomes) that can then be used as the basic foundation for estimating trips from, or attracted to, that zone. In most cases, transportation planners try to create a **zonal system** by defining zones that are related to other data collection activities (e.g., the U.S. Census tracts). The transportation system is represented in models as a network of *links* and *nodes*. Links represent line-haul facilities, such as roads or transit lines, and nodes represent points of connection, such as intersections. Given the complex nature of transportation systems, the typical transportation network consists of links representing only highly used facilities. The transportation modeling process can thus be represented as shown in [Fig. 79.2](#).

**Figure 79.2** Transportation modeling framework.



*Trip generation* is the process of analytically deriving the number of trips that will be generated from a location or zone based on socioeconomic information of the travelers or, in the case of freight movement, zonal economic characteristics. The trip generation stage also includes predicting the number of trips that will be attracted to each zone in the study area:

Number of trips produced in a zone =  $f$  (Socioeconomic characteristics, land use, transportation mode availability)

Number of trips attracted to a zone =  $f$  (Attractiveness of zone)

There are two approaches for estimating trips generated. The first approach uses *trip rate models*. Based on existing data, trip rates can be calculated that relate trip-making to variables known to influence travel behavior. For an example, see [Table 79.1](#).

**Table 79.1** Cross-Classification Analysis

	Number of People in Households		
	1	2	3+
Low Income	2.4	3.3	4.5
Medium Income	3.5	3.8	4.8
High Income	3.9	4.2	5.4

The other approach is to use *regression models*. Based on existing data, multiple regression equations are estimated relating trip generation to variables found to be significant in travel behavior. For example,

$$\text{Zone estimate: } T_{iz} = 184.2 + 120.6(X1_i) + 34.5(X2_i)$$

$$\text{Household estimate: } T_{ih} = 0.64 + 2.3(X1_{ih}) + 1.5(X2_{ih})$$

$$\text{Zonal attractions: } T_j = 54.2 + 0.23(X3) + 0.43(X4)$$

where

- $T_{iz}$  = Total trips generated in zone  $i$
- $T_{ih}$  = Total trips generated per household in zone  $i$
- $T_j$  = Total trips attracted to zone  $j$
- $X1_i$  = Total number of workers in zone  $i$
- $X2_i$  = Total number of automobiles in zone  $i$
- $X1_{ih}$  = Number of employees per household in zone  $i$
- $X2_{ih}$  = Number of automobiles per household in zone  $i$
- $X3$  = Total number of office employees in zone  $j$
- $X4$  = Total number of retail employees in zone  $j$

*Trip distribution* is the process of estimating the number of trips that originate in each zone in the study area and their corresponding destinations. The result of the trip distribution process is a matrix called a *trip table*, which shows the number of trips originating in each study zone and their corresponding destinations for the time period being examined. The most commonly used method for distributing trips in a zonal system is the gravity model, which is of the following form:

$$T_{ij} = P_i \frac{A_j \times F_{ij} \times K_{ij}}{\sum_j (A_j \times F_{ij} \times K_{ij})}$$

where

- $T_{ij}$  = Number of trips originating in zone  $i$  and destined to zone  $j$
- $P_i$  = Number of trips produced in zone  $i$
- $A_j$  = Level of attractiveness of zone  $j$  (e.g., number of employees)
- $F_{ij}$  = Friction or impedance factor between zones  $i$  and  $j$  (a value usually a function of travel time)
- $K_{ij}$  = Socioeconomic adjustment factors for trips between zones  $i$  and  $j$  (a value that represents variables that influence trip making not accounted for by other variables)

*Mode choice* is the process of estimating the percentage of individuals who will use one mode of transportation versus the others available for a given trip. The basic concept in making this estimation is that each mode of transportation has associated with it some empirically known characteristics that, when combined in a mathematical equation, can define that mode's *utility*. Variables such as travel time, travel cost, modal reliability, and so on are often incorporated into a mode's **utility function**, along with socioeconomic characteristics of the traveler. Freight models use a similar concept in estimating commodity flows by mode. One of the most familiar forms of mode choice models is the logit model, which predicts mode shares based on the following equation:



$$P_{ik} = \frac{e^{U_k}}{\sum_{m=1}^n e^{U_m}}$$

where

- $P_{ik}$  = Probability of individual  $i$  choosing mode  $k$
- $U_k$  = Utility of mode  $k$
- $U_m$  = Utility of mode  $m$
- $n$  = Number of modes available for trip

The utility of each mode is often represented as a linear function of those variables found to influence an individual's choice of mode. For example, a utility function for the automobile mode might be of the form

$$U_a = 6.3 - 0.21(X_1) - 0.43(X_2) - 0.005(X_3)$$

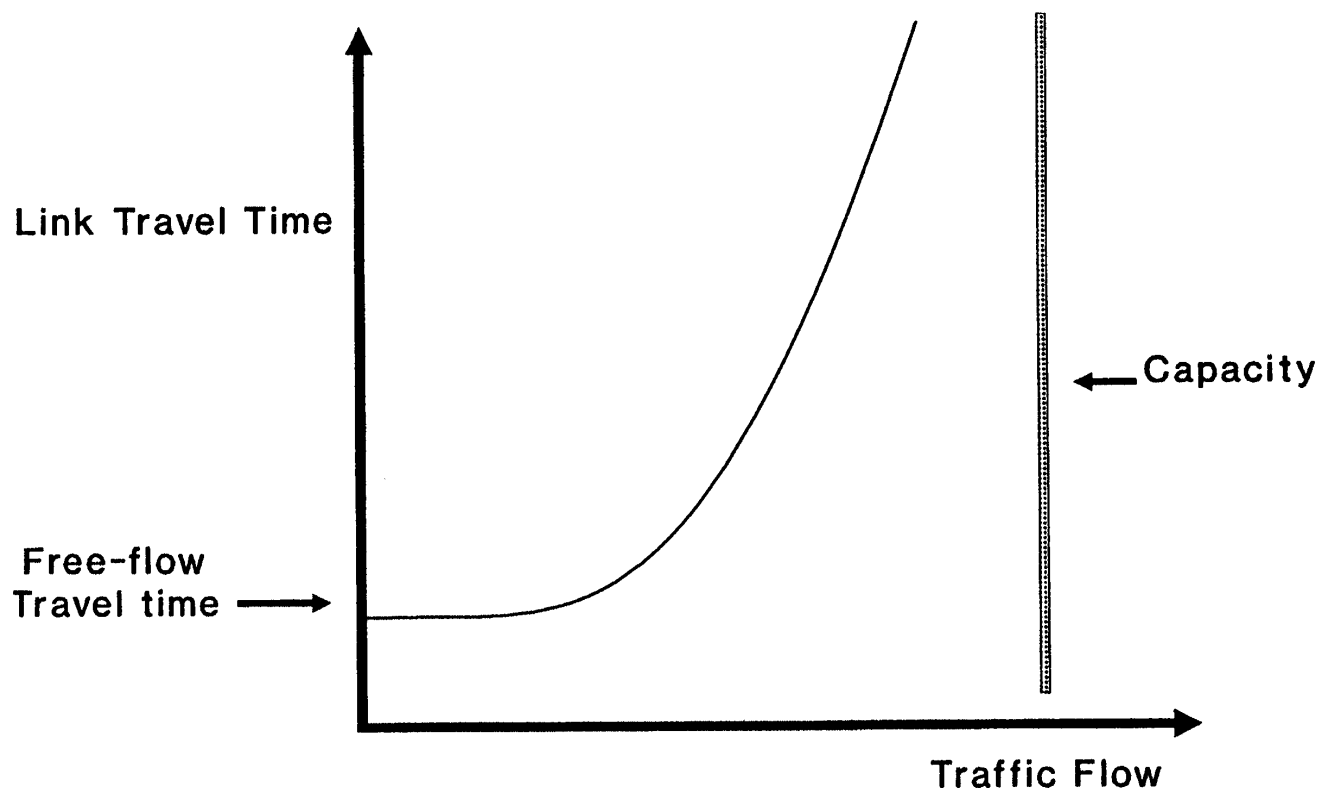
where

- $U_a$  = Utility of automobile
- $X_1$  = Access and egress time when automobile is chosen
- $X_2$  = Line-haul time
- $X_3$  = Travel cost

The utility functions of other modes available for a specific trip would be similarly specified. The probabilities would then be multiplied by the total number of trips to obtain the number of trips made by mode.

*Trip assignment* is the process of estimating the travel flows through a transportation network based on a trip table (which is produced in trip distribution). The basic concept found in all trip assignment models is that travelers choose routes that will minimize travel time—that is, they will choose the shortest path through the network. Link performance functions that relate travel time to number of vehicles on the link (see [Fig. 79.3](#)) are used to iteratively update estimated link travel times so that minimum-path travel times reflect the effect of congestion. Recent research in trip assignment methods has introduced the concept of stochastic assignment, which means that some subset of trip routes will in fact have associated with them some characteristics that attract certain types of travelers, even if the travel time is longer.

**Figure 79.3** Link performance function.



## Defining Terms

**Derived demand:** An assumption that travelers make a trip to accomplish some objective at the destination and that the trip itself is simply a means of reaching this activity.

**Transportation network:** A transportation system is represented in models as a network of *links* and *nodes*. Links represent line-haul facilities, such as roads or transit lines, and nodes represent points of connection, such as intersections. Given the complex nature of transportation systems, the typical transportation network consists of links representing only higher functionally classified facilities, or those links that represent known highly traveled paths.

**Utility function:** A mathematical formulation that assigns a numerical value to the attractiveness of individual modes of transportation based primarily on that mode's characteristics.

**Zonal system:** Each study area (whether a nation, state, metropolitan area, or neighborhood) is divided into zones of homogeneous characteristics (e.g., similar household incomes) that can then be used as the basic foundation for estimating trips from, or attracted to, that zone. In most cases transportation planners try to define zones that are related to other data collection activities (e.g., the U.S. Census tracts).

## References

Institute of Transportation Engineers. 1993. *Trip Generation Handbook*, 5th ed. ITE, Washington,

DC.

- Johnson, E. 1993. *Avoiding the Collision of Cities and Cars*. Report on a Study Project sponsored by the American Academy of Arts and Sciences, September 1993.
- Mannering, F. and Kilareski, W. 1990. *Principles of Highway Engineering and Traffic Analysis*. John Wiley & Sons, New York.
- Meyer, M. and Miller, E. 1984. *Urban Transportation Planning: A Decision-Oriented Approach*. McGraw-Hill, New York.
- Newell, G. 1980. *Traffic Flow on Transportation Networks*. MIT Press, Cambridge, MA.
- Ogden, K. 1992. *Urban Goods Movement*. Ashgate, Brookfield, VT.
- Ortuzar, J. and Willumsen, L. 1994. *Modelling Transport*, 2nd ed. John Wiley & Sons, New York.
- Papacostas, C. S. 1992. *Fundamentals of Transportation Engineering*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.
- Stover, V. and Koepke, F. 1988. *Transportation and Land Development*. Institute of Transportation Engineers, Washington, DC.

## **Further Information**

Transportation Research Board, National Academy of Sciences  
2101 Constitution Ave., N.W.  
Washington, DC 20418

American Association of State Highway and Transportation Officials  
444 N. Capitol St., N.W.  
Suite 225  
Washington, DC 20001

Institute of Transportation Engineers  
525 School St., S.W.  
Suite 410  
Washington, DC 20024

Leonard II, J. & Meyer, M.D. "Design of Transportation Facilities"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Design of Transportation Facilities

---

## 80.1 Components of the Project Development Process

Identify Project Need • Establish Project Limits and Context • Establish Environmental Impact Requirements • Develop Strategy for Interagency Coordination and Public Involvement • Initiate Project Design and Preliminary Engineering • Project Engineering • Final Engineering

## 80.2 Basic Concepts of Project Design

Human Factors • Vehicle or User Performance Factors • Classification Schemes • Capacity and Level of Service • Design Standards

## 80.3 Intermodal Transportation Terminals or Transfer Facilities

## 80.4 Advanced Technology Projects

Define Problems and Needs • Define System • Define Users • Establish Institutional Framework and Partnerships • Develop User Service Plan • Define System Architecture • Evaluate Alternative Technologies

### John Leonard II

*Georgia Institute of Technology*

### Michael D. Meyer

*Georgia Institute of Technology*

The efficient movement of people and goods requires transportation systems and facilities that are designed to provide sufficient capacity for the demands they face in as safe a manner as possible. In addition, in most modern societies, the design of transportation facilities must explicitly minimize harm to the natural and human-made environment while providing for mitigation measures that relate to those impacts that are unavoidable. In many ways the critical challenge to today's designers of transportation projects is successfully designing a facility that minimally harms the environment.

The design of a transportation facility almost always takes place within the context of a much broader **project development process**. This process can vary in complexity with the type of project under design and with the scale of implementation. The importance of the project development process to the designer is that it

- Establishes the key characteristics of the project that must be considered in the design
- Indicates the time frame that will be followed for project design
- Establishes which agencies and groups will be involved in the process and when this involvement will likely occur
- Links the specific elements of the project design with other tasks that must be accomplished

for the project to be constructed

- Satisfies legal requirements for a design process that is open for review and comment
- Indicates the specific products that must be produced by the designers to complete the project design process

In most cases the project development process consists of a well-defined set of tasks that must be accomplished before the next task can occur. These tasks include both technical activities and public involvement activities that are necessary for successful project development.

## **80.1 Components of the Project Development Process**

---

### **Identify Project Need**

A project need can be identified through a formal planning process or from a variety of other sources, including elected officials, agency managers, transportation system users, and citizens. Important in this early portion of project development is an indication of what type of improvement is likely to be initiated. For example, a project could relate to one or more of the following types of improvement strategies:

- *New construction.* A transportation facility constructed at a new location.
- *Major reconstruction.* Addition of new capacity or significant changes to the existing design of a facility, but usually occurring within the area where the current facility is located.
- *Rehabilitation/restoration.* Improvements to a facility usually as it is currently designed and focusing on improving the physical condition of the facility or making minor improvements to enhance safety.
- *Resurfacing.* Providing new pavement surface to a transportation facility that prolongs its useful life.
- *Spot improvements.* Correction of a problem or hazard at an isolated or specific location.

### **Establish Project Limits and Context**

One of the very first steps in the design process is to define the boundaries or limits of the project. This implies establishing how far the project will extend beyond the area being targeted for improvement and the necessary steps to ensure smooth connections to the existing transportation system.

### **Establish Environmental Impact Requirements**

The design of a project will very much be affected by environmental laws or regulations that require design compliance with environmental mandates. These environmental mandates could relate to such things as wetlands, historic properties, use of public park lands, water quality, navigable waterways, fish and wildlife, air quality, noise, and archaeological resources. One of the first steps in project development is to determine whether the likely project impacts are significant

enough to require environmental study.

## **Develop Strategy for Interagency Coordination and Public Involvement**

Depending on the complexity and potential impact of a project, the project designer could spend a great deal of time interacting with agencies having some role in or jurisdictional control over areas directly related to the project. These agencies could have jurisdiction by law (e.g., wetlands) or have special expertise that is important to project design (e.g., historic preservation). In addition to interagency coordination, transportation project development is often subject to requirements for public outreach and/or public hearings.

## **Initiate Project Design and Preliminary Engineering**

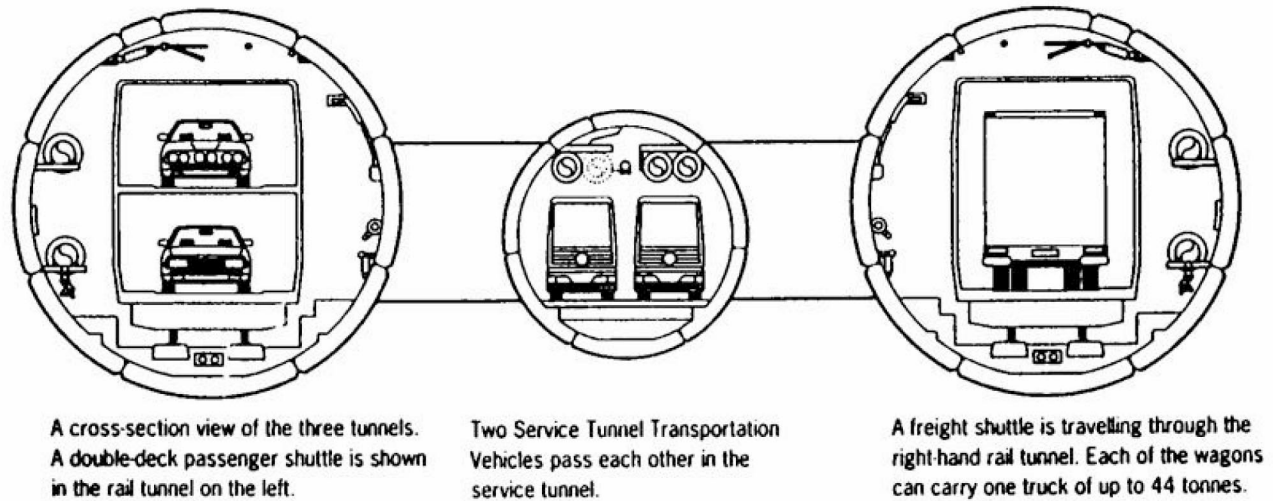
Topographic data of the study area and forecasted vehicular volumes expected to use the facility in the design year are used as input into the preliminary design of the horizontal and vertical alignment of the facility, that is, the physical space the facility will occupy once finished. This preliminary engineering step also includes the preparation of initial right-of-way (ROW) plans, which indicate the amount of land that must be available to construct the facility.

## **Project Engineering**

Once preliminary engineering has provided the basic engineering information for the project, the more detailed project design begins. This entails specific layouts of horizontal and vertical geometry, soils/subsurface examination and design, design of utility location, drainage design, more detailed ROW plans, and initial construction drawings. Concurrent with this design process, the environmental process continues with updated information on project changes that might cause additional environmental harm, the initiation of any permitting process that might be needed to construct the project (e.g., environmental agency permission to affect wetlands), and public hearings/meetings to keep the public involved with project development.

## **Final Engineering**

The final engineering step is the culmination of the design process, which basically completes the previous design plans to the greatest level of detail. This step includes finalizing ROW plans, cost estimates, construction plans, utility relocation plans, and any agreements with other agencies or jurisdictions that might be necessary to complete the project. Environmental permits are received and final project review for environmental impacts is completed.



A cross-section view of the three parallel tunnels of the Eurotunnel system. A double-deck passenger shuttle is shown in the rail tunnel on the left. The service tunnel is in the middle, and a freight shuttle is traveling in the rail tunnel on the right.

## EUROTUNNEL DESIGN

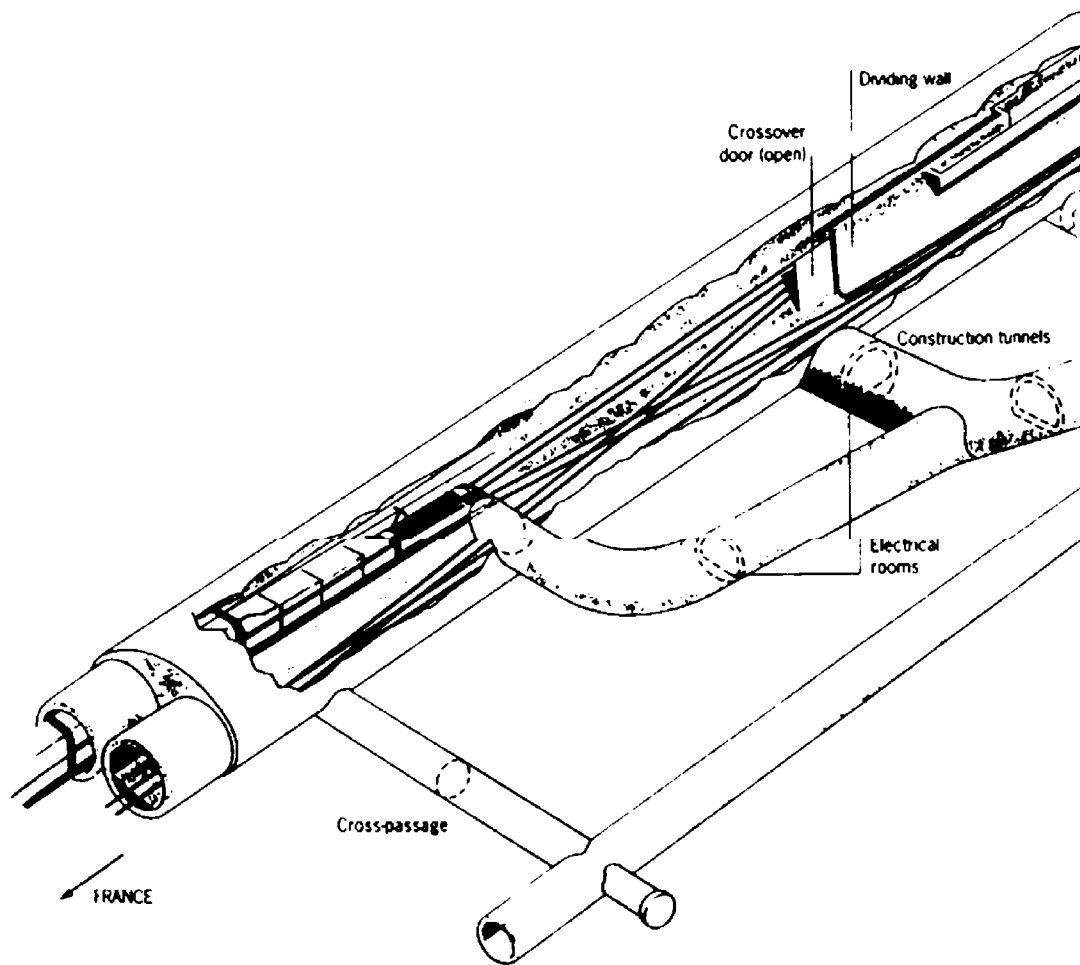
---

The Channel Tunnel or Eurotunnel system is not one tunnel, but actually three parallel tunnels. Two of these are one-way rail tunnels. Trains traveling from Britain to France use the northern tunnel, and those traveling from France to Britain use the southern tunnel. The third tunnel is a smaller service tunnel. It allows engineers to travel along the system without stopping shuttle traffic and is used primarily for routine maintenance and for evacuation in case of an emergency.

The three tunnels are connected by cross-passages every 1230 ft (375 m). These cross-passages give engineers access from the service tunnel into the rail tunnels for necessary maintenance work. In addition to the cross-passages, the two rail tunnels are also joined by piston relief ducts. Open valves in these ducts allow the air pushed down the rail tunnel in front of the speeding trains to discharge harmlessly into the other rail tunnel. The valves are closed only during maintenance work. Air pressure within the service tunnel is kept higher than in the rail tunnels. This allows the atmosphere in the service tunnel to remain clear in the unlikely event of smoke being present in the rail tunnels.

There are two huge "crossover caverns" located at about one-third and two-thirds of the total undersea distance within the tunnel. These crossover caverns allow trains to change from one rail tunnel to the other. When one of the tunnels is closed for maintenance or repairs, the other is used as a single-track railway. As these crossovers divide each rail tunnel into three sections, only one-sixth of the entire system needs to be out of service at any time. This prevents the considerable delays that would occur if one entire tunnel had to be closed. Huge sliding doors separate the two tunnels when the crossovers are not in use. (Copyright Eurotunnel 1994/QA Photos. Used with permission.)





This illustrates one of the two crossover caverns within the Eurotunnel system.

---

## 80.2 Basic Concepts of Project Design

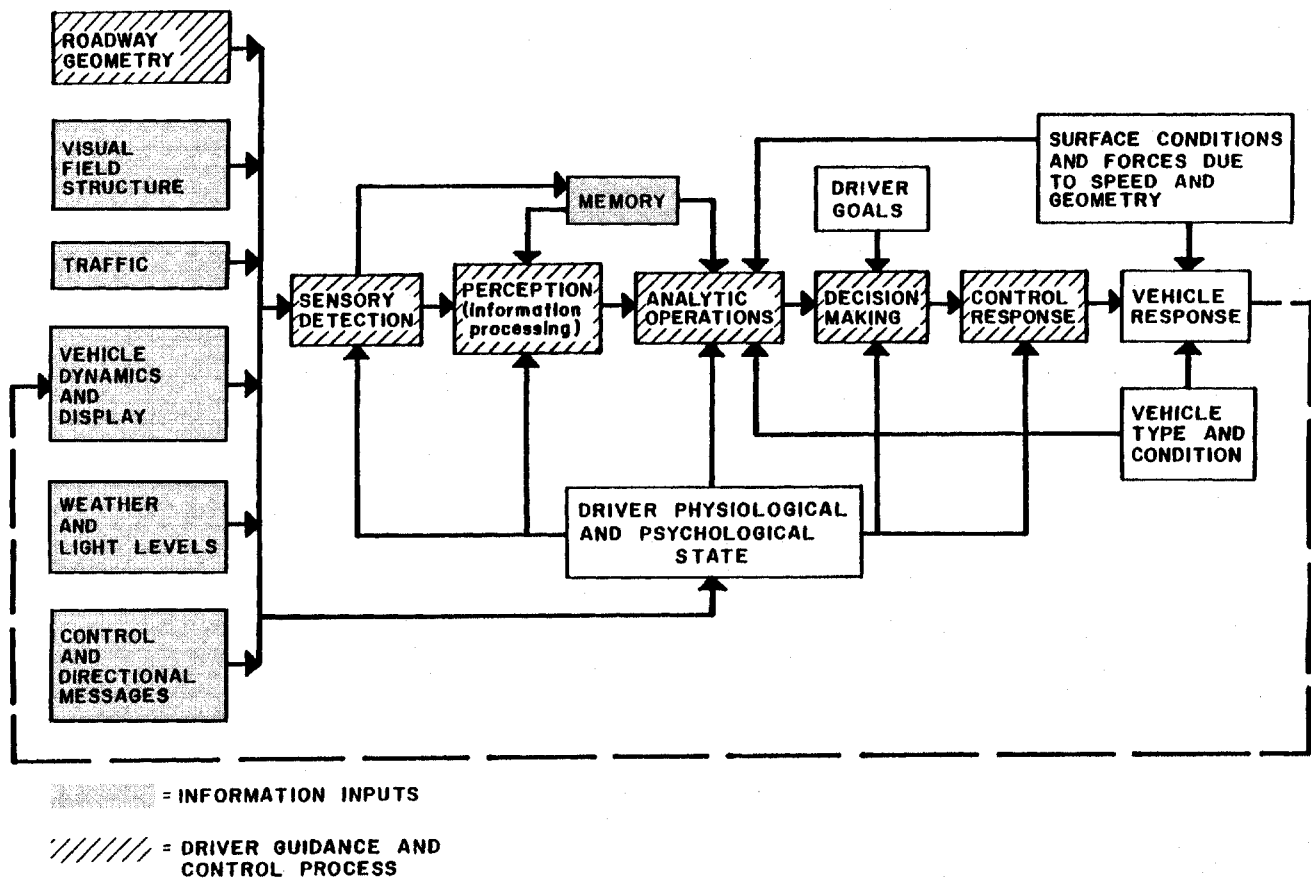
---

### Human Factors

Human factors have a great deal of influence on the design of transportation facilities in such things as width of facility, length and location of access/egress points, braking distance, location of information/guidance aids such as signs, and geometric characteristics of the facility's alignment.

The driver-vehicle-roadway interface is shown in Fig. 80.1.

**Figure 80.1** Driver-vehicle-roadway interface.



## Vehicle or User Performance Factors

The dynamics of vehicle motion play an important role in determining what is effective and safe design. The key vehicle characteristics that become important in design criteria include:

- *Vehicle size.* Influences vertical and horizontal clearances, turning radii, alignment width, and width of vehicle storage berths.
- *Vehicle weight.* Influences strength of material needed to support vehicle operations.
- *Vehicle or user performance.* Influences specifications for horizontal and vertical geometry, braking distances, and needed capacity to allow passing and successful maneuvering (e.g., assumed walking speed of pedestrians crossing a road).

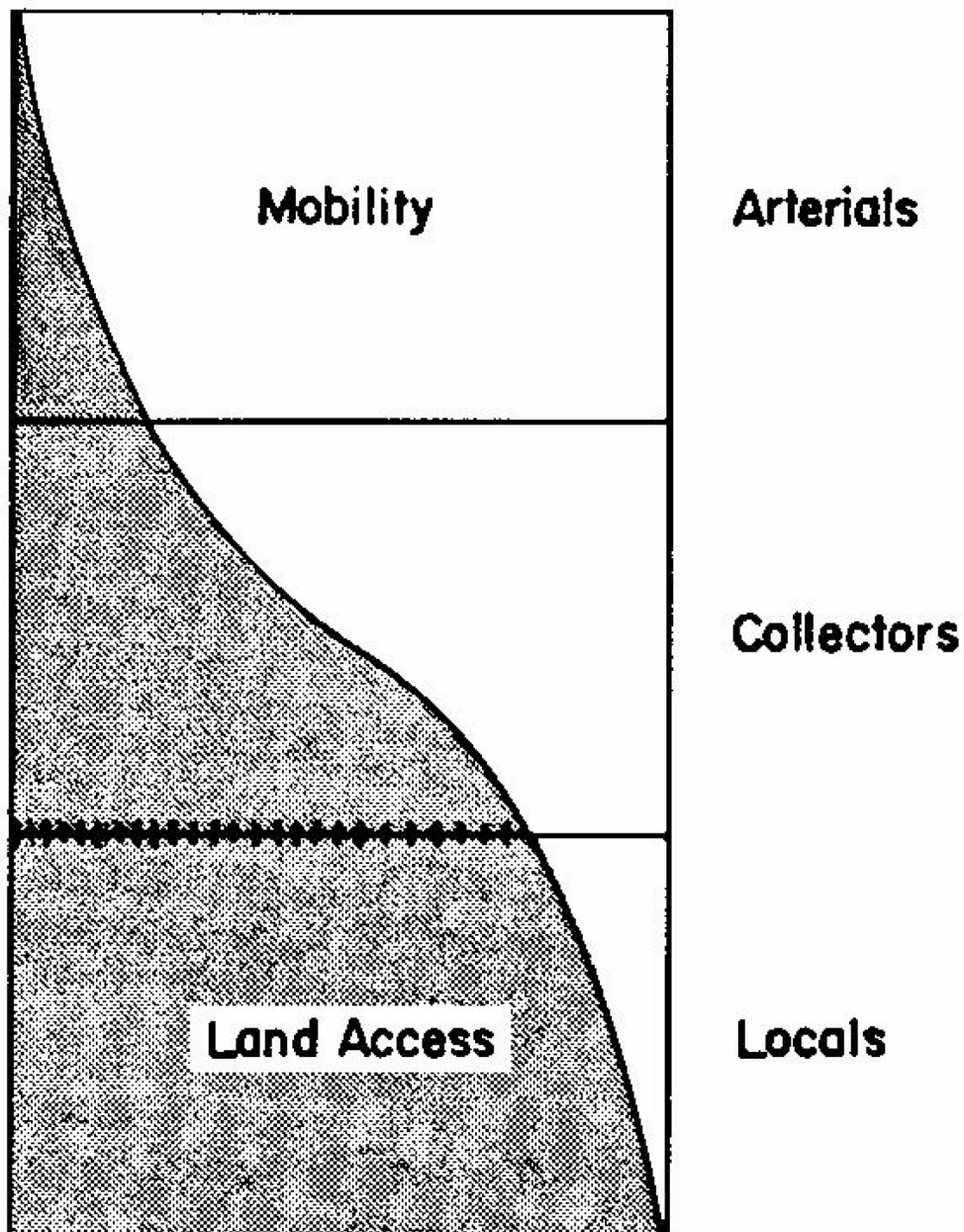
## Classification Schemes

The transportation system serves many functions, ranging from providing access to specific

locations to providing high-speed, high-capacity movement over longer distances. Classification schemes are used to represent these various roles and influence the design criteria that are associated with the facilities in each classification category. A common **functional classification** scheme for highways is shown in [Fig. 80.2](#).

**Figure 80.2** Relationship of functionally classified systems in service traffic mobility and land access. (Source: AASHTO, 1990, Figure I-5.)

## PROPORTION OF SERVICE



## Capacity and Level of Service

Every design usually begins with some estimation of the demand for the transportation facility that will likely occur if the facility is built. The key design question then becomes, what facility capacity (e.g., number of road lanes, runways, transit lines, or vehicle departures) is necessary if a certain level of service is desired? The definition of **level of service (LOS)** thus becomes a critical element in establishing this important design factor (see Fig. 80.3).

**Figure 80.3** Example level of service characteristics (*Source: Transportation Research Board. 1994. Highway Capacity Manual. Washington, DC.*)

FREEWAYS				
Level of Service	Maximum Density (pc/mi/ln)	Minimum Speed (mph)	Max Service Flow Rate (pcphpl)	Maximum v/c Ratio
Free-Flow Speed = 70 mph				
A	10.0	70.0	700	0.318/0.304
B	16.0	70.0	1120	0.509/0.487
C	24.0	68.5	1644	0.747/0.715
D	32.0	63.0	2015	0.916/0.876
E	36.7/39.7	60.0/58.0	2200/2300	1.000
F	var	var	var	var
Free-Flow Speed = 65 mph				
A	10.0	65.0	650	0.295/0.283
B	16.0	65.0	1040	0.473/0.452
C	24.0	64.5	1548	0.704/0.673
D	32.0	61.0	1952	0.887/0.849
E	39.3/43.4	56.0/53.0	2200/2300	1.000
F	var	var	var	var
Free-Flow Speed = 60 mph				
A	10.0	60.0	600	0.272/0.261
B	16.0	60.0	960	0.436/0.417
C	24.0	60.0	1440	0.655/0.626
D	32.0	57.0	1824	0.829/0.793
E	41.5/46.0	53.0/50.0	2200/2300	1.000
F	var	var	var	var
Free-Flow Speed = 55 mph				
A	10.0	55.0	550	0.250/0.239
B	16.0	55.0	880	0.400/0.383
C	24.0	55.0	1320	0.600/0.574
D	32.0	54.8	1760	0.800/0.765
E	44.0/47.9	50.0/48.0	2200/2300	1.000
F	var	var	var	var

Note: In table entries with split values, the first value is for four-lane freeways, and the second is for six- and eight-lane freeways.

PEDESTRIAN WALKWAYS				
Level of Service	Space (sq ft/ped)	Expected Flows and Speeds		
		Ave. Speed, S (ft/min)	Flow Rate, v (ped/min/ft)	Vol/Cap Ratio, v/c
A	≥ 130	≥ 260	≤ 2	≤ 0.08
B	≥ 40	≥ 250	≤ 7	≤ 0.28
C	≥ 24	≥ 240	≤ 10	≤ 0.40
D	≥ 15	≥ 225	≤ 15	≤ 0.60
E	≥ 6	≥ 150	≤ 25	≤ 1.000
F	< 6	< 150	—Variable—	

\*Average conditions for 15 min.

### SIGNALIZED INTERSECTIONS

Level of Service	Stopped Delay per Vehicle (sec)
A	≤ 5.0
B	5.1 to 15.0
C	15.1 to 25.0
D	25.1 to 40.0
E	40.1 to 60.0
F	> 60.0

### ARTERIAL ROADS

Arterial Class	I	II	III
Range of Free Flow Speeds (mph)	45 to 35	35 to 30	35 to 25
Typical Free Flow Speed (mph)	40 mph	33 mph	27 mph
Level of Service	Average Travel Speed (mph)		
A	≥ 35	≥ 30	≥ 25
B	≥ 28	≥ 24	≥ 19
C	≥ 22	≥ 18	≥ 13
D	≥ 17	≥ 14	≥ 9
E	≥ 13	≥ 10	≥ 7
F	< 13	< 10	< 7

## Design Standards

**Design standards** dictate minimum or maximum values of project characteristics that are associated with a particular facility type. Design standards usually result from extensive study of the relationship between various design characteristics and vehicle performance and the safe handling of the vehicles by human operators. Design standards often vary by the "design speed" of the facility (and thus the importance of the facility classification) and by the "design vehicle." Design standards are often the basis for developing typical cross sections (see Figs. 80.4 and 80.5).

**Figure 80.4** Example design criteria. (Source: Massachusetts Department of Public Works. 1988. *Highway Design Manual*. Boston, MA.)

**RECOMMENDED ROADWAY SECTION WIDTHS**

Functional Class	U/R	Number of Lanes	Travel Lane	Shoulder	
				Right	Left <sup>1</sup>
Freeway	Urban	4–8	12	10	4 <sup>2</sup>
	Rural	4–8	12	10	4 <sup>2</sup>
Arterial	Urban	Multilane with median	12	10	4
	Urban	Multilane without median	11–12	8–10 <sup>3</sup>	N/A
	Rural	2 lane	12	See Table 5.2	N/A
	Rural	Multilane with median	12	8–10	4

**WIDTH OF USABLE SHOULDER—EACH SIDE OF TRAVEL WAY  
RURAL TWO-LANE ARTERIAL**

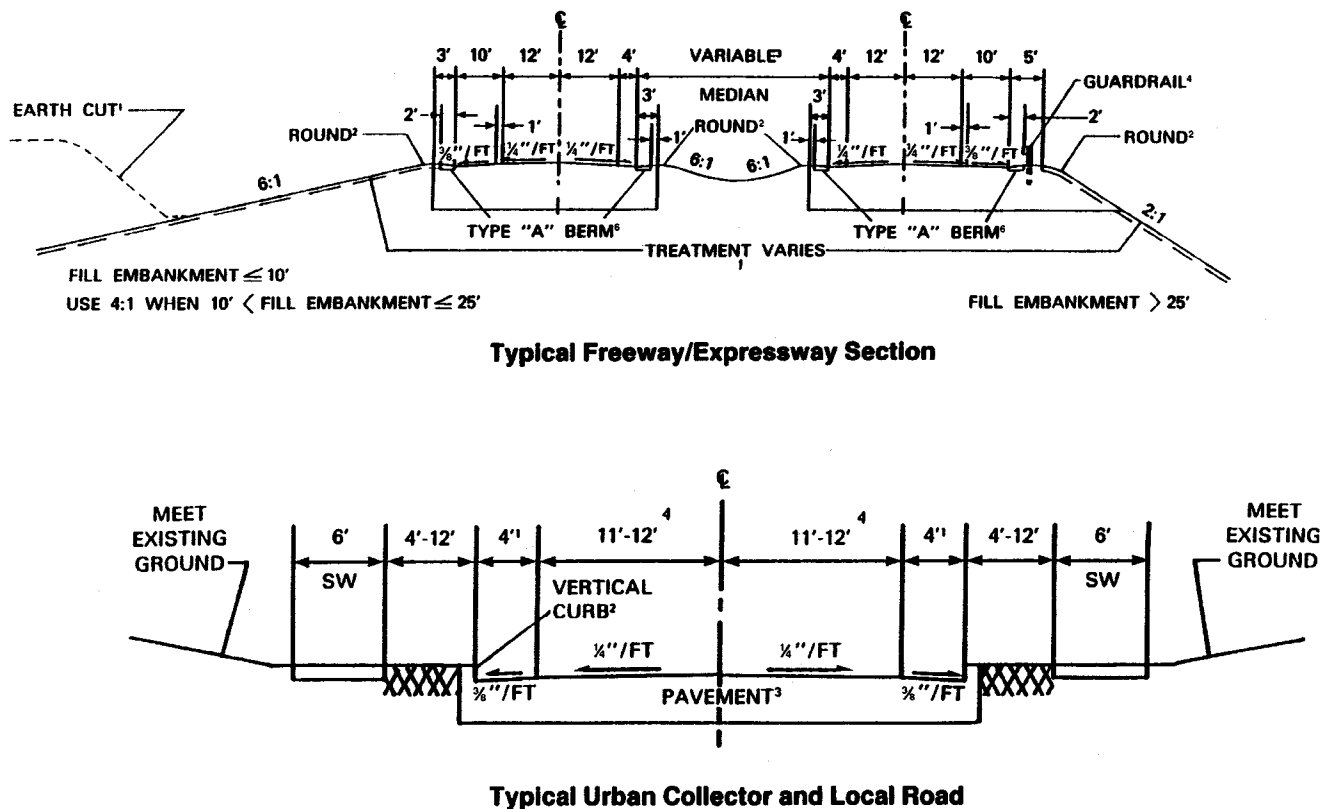
Design Traffic Volume					
	Current ADT Under 400	Current ADT Over 400	DHV 100–200	DHV 200–400	DHV Over 400
All design speeds	4 ft	6 ft	6 ft	8 ft	10 ft

**RECOMMENDED WIDTH OF TRAVEL WAY AND GRADED SHOULDER  
RURAL COLLECTOR**

Design Traffic Volume					
Design Speed (mph)	Current ADT Under 400	Current ADT Over 400	DHV 100–200	DHV 200–400	DHV Over 400
30	20 ft	20 ft	20 ft	22 ft	24 ft
40	20 ft	22 ft	22 ft	22 ft	24 ft
50	20 ft	22 ft	22 ft	24 ft	24 ft
60	22 ft	22 ft	22 ft	24 ft	24 ft
Graded Shoulder (Each Side)*					
All speeds	2 ft	4 ft	6 ft	8 ft	8 ft

\*If right-of-way permits

**Figure 80.5** Example cross sections.



## 80.3 Intermodal Transportation Terminals or Transfer Facilities

Terminals or transfer facilities are locations where users of the transportation system change from one mode of travel to another. The effective design of such facilities is a critical element of successful transportation system performance, given the potential bottlenecks they represent if not designed appropriately. The design of terminals and transfer facilities must pay special attention to the needs of the users of the facility, in that they serve to establish the effective capacity of the facility, for example:

1. Internal pedestrian movement facilities and areas (stairs, ramps, escalators, elevators, corridors, etc.)
2. Line-haul transit access area (entry control, fare collection, loading, and unloading)
3. Components that facilitate movements between access modes and the station (ramps or electric doors)
4. Communications (public address systems and signage)
5. Special provisions for disabled patrons (elevators and ramps)

The criteria that could relate to the design of such a facility include threshold values for pedestrian level of service, delay at access points, connectivity from one area of the facility to another, and low-cost maintenance. For the vehicle side of such terminals, special consideration must be given to the performance of the design vehicle (e.g., turning radii of buses or semitrailer

trucks) and the vehicle storage requirements (e.g., the number, size, and orientation of loading/unloading berths).

## **80.4 Advanced Technology Projects**

---

One of the characteristics of transportation system development in recent years has been the increased application of advanced (usually electronic) technologies to improve system performance. The following steps apply to designing advanced technology projects.

### **Define Problems and Needs**

Typical problems or needs might relate to congestion, excessively high accident rates, or improving current system capabilities and levels of service.

### **Define System**

A system definition should include a mission statement, listing of physical components (e.g., roads, travelers, buses, rolling stock, existing rail lines, control centers, and communication links), and the physical relationship between those components.

### **Define Users**

*System users* is a rather broad description of all individuals, organizations, and other systems that might interact or have a stake in the fully implemented transportation system under study.

### **Establish Institutional Framework and Partnerships**

Various organizations possess differing missions, priorities, and policies—sometimes in conflict. Strong emphasis on coalition building during the early stages of project planning and engineering can help diffuse potential project-stopping disagreements later in the process.

### **Develop User Service Plan**

The development of a user service plan consists of the following steps: (1) establish user services, (2) identify technology areas, and (3) map user services to technology areas. User services might include:

- Traveler information services
- Freight and fleet management services
- Emergency vehicle management services
- Traffic management services
- Public transport services

Available technologies fall within one of the following functional areas: (1) surveillance, (2)

communications, (3) traveler interface, (4) control strategies, (5) navigation/guidance, (6) data processing, and (7) in-vehicle sensors.

## Define System Architecture

A logical architecture consists of a *block diagram* identifying the major systems and subsystems, the participating agencies, and users. Through the use of arrows, the flow of information between these elements is identified. A logical architecture also shows the allocation of responsibilities throughout the transportation system.

## Evaluate Alternative Technologies

Some of the factors to be considered in this evaluation include (1) cost, (2) performance, (3) reliability, (4) compatibility, (5) environmental impacts, and (6) compliance to standards.

## Defining Terms

**Design standard:** Physical characteristics of a proposed facility that are professionally accepted and often based on safety considerations.

**Functional classification:** Classifying a transportation facility based on the function it serves in the transportation system. Such classification becomes important in that design standards are often directly related to the functional classification of a facility.

**Level of service:** An assessment of the performance of a transportation facility based on measurable physical characteristics (e.g., vehicular speed, average delay, density, flow rate). Level of service is usually subjectively defined as ranging from level of service A (good performance) to level of service F (bad or heavily congested performance).

**Project development process:** The steps that are followed to take a project from initial concept to final engineering. This process includes not only the detailed engineering associated with a project design but also the interaction with the general public and with agencies having jurisdiction over some aspect of project design.

## References

- American Association of State Highway and Transportation Officials. 1990. *A Policy on the Geometric Design of Highways and Streets*. AASHTO, Washington, DC.
- Mannering, F. and Kilareski, W. 1990. *Principles of Highway Engineering and Traffic Analysis*. John Wiley & Sons, New York.
- Massachusetts Department of Public Works. 1988. *Highway Design Manual*. MDPW, Boston, MA.
- Transportation Research Board. 1994. *Highway Capacity Manual*. Special Report 209, TRB, Washington, DC.
- Wright, P. and Ashford, N. 1989. *Transportation Engineering, Planning and Design*. John Wiley & Sons, New York.



## **Further Information**

Transportation Research Board, National Academy of Sciences, 2101 Constitution Ave., N.W.,  
Washington, DC 20418

American Association of State Highway and Transportation Officials, 444 N. Capitol St., N.W.,  
Suite 225, Washington, DC 20001

Institute of Transportation Engineers, 525 School St., S.W., Suite 410, Washington, DC 20024

Shuldiner, P.W. & Black, K.B. "Operations and Environmental Impacts  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

This chapter is not available because of  
copyright issues

Schonfeld, P. "Transportation Systems"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Transportation Systems

---

- 82.1 Transportation System Components
- 82.2 Evaluation Measures
- 82.3 Air Transportation
- 82.4 Railroad Transportation
- 82.5 Highway Transportation
- 82.6 Water Transportation
- 82.7 Public Transportation

## Paul Schonfeld

*University of Maryland*

The various forms of transportation that have been developed over time are called **modes**. The classification of modes may be very broad (e.g., highway transportation or air transportation) or more restrictive (e.g., chartered helicopter service). The major distinctions among transportation modes that help to classify them include:

1. Medium (e.g., air, space, surface, underground, water, underwater)
2. Users (e.g., passengers vs. cargo, general-purpose vs. special trips or commodities, common vs. private carrier)
3. Service type (scheduled vs. **demand responsive**, fixed vs. variable route, nonstop vs. express or local, mass vs. personal)
4. Right-of-way type (exclusive, semi-exclusive, shared)
5. Technology:
  - a. Propulsion (e.g., electric motors, diesel engines, gas turbines, linear induction motors, powered cables)
  - b. Energy sources (e.g., petroleum fuels, natural gas, electric batteries, electric power from conducting cables)
  - c. Support (e.g., aerodynamic lift, flotation on water, steel wheels on two steel rails, monorails, air cushions, magnetic levitation, suspension from cables)
  - d. Local control (e.g., lateral control by steering wheels, wheel flanges on railroad vehicles, rudders, longitudinal control by humans or automatic devices)
  - e. Network guidance and control systems (with various degrees of automation and optimization)

A mode may be defined by its combination of such features. The number of conceivable combinations greatly exceeds the number of modes that have been actually tried, which, in turn,

exceeds the number of successful modes. Success may be limited to relatively narrow markets and applications (e.g., for helicopters or aerial cablecars) or may be quite general. Thus, automobiles are successful in a very broad range of applications and have become the basis for distinct transportation modes such as taxis, carpools, or ambulances.

The relative success of various transportation modes depends on available technology and socio-economic conditions at any particular time, as well as on geographic factors. As technology or socio-economic conditions change, new transportation modes appear, develop, and may later decline as more effective competitors appear. For many centuries water transportation was considerably cheaper than overland transportation. Access to waterways was quite influential in the location of economic activities and cities. Access to good transportation is still very important to industries and communities. Technological developments have so drastically improved the relative effectiveness of air transportation that within a short period (approximately 1950 to 1965) aircraft almost totally replaced ships for transporting passengers across oceans. It is also notable that as economic prosperity grows, personal transportation tends to shift from the walking mode to bicycles, motorcycles, and then automobiles. Geography can significantly affect the relative attractiveness of transportation modes. Thus, natural waterways are highly valuable where they exist. Hilly terrain decreases the economic competitiveness of artificial waterways or conventional railroads while favoring highway modes. In very mountainous terrain even highways may become uncompetitive compared to alternatives such as helicopters, pipelines, and aerial cablecars.

The relative shares of U.S. intercity passenger and freight traffic are shown in [Table 82.1](#). The table shows the relative growth since 1929 of airlines, private automobiles, and trucks and the relative decline of railroad traffic.

**Table 82.1** Volume of U.S. Intercity Freight and Passenger Traffic

**Table 82.1** Volume of U.S. Intercity Freight and Passenger Traffic

Millions of Revenue Freight Ton-Miles and Percentage of Total													
Year	Railroads	%	Trucks	%	Great Lakes	%	Rivers and Canals	%	Pipelines	%	Air	%	Total
1929	454,800	74.9	19,689	3.2	97,322	16.0	8,661	1.4	26,900	4.4	3	0.0	607,375
1939	338,850	62.3	52,821	9.7	76,312	14.0	19,937	3.7	55,602	10.2	12	0.0	543,534
1944	746,912	68.6	58,264	5.4	118,769	10.9	31,386	2.9	132,864	12.2	71	0.0	1,088,266
1950	596,940	56.2	172,860	16.3	111,687	10.5	51,657	4.9	129,175	12.2	318	0.0	1,062,637
1960	579,130	44.1	285,483	21.7	99,468	7.6	120,785	9.2	228,626	17.4	778	0.1	1,314,270
1970	771,168	39.8	412,000	21.3	114,475	5.9	204,085	10.5	431,000	22.3	3,295	0.2	1,936,023
1980	932,000	37.5	555,000	22.3	96,000	3.9	311,000	12.5	588,000	23.6	4,840	0.2	2,486,840
1986	889,000	35.5	634,000	25.3	68,000	2.7	325,000	13.0	578,000	23.1	7,340	0.3	2,501,340
1987	968,000	36.3	668,000	25.1	78,000	2.9	358,000	13.4	585,000	22.0	8,720	0.3	2,665,720
Millions of Revenue Passenger-Miles and Percentage of Total (Except Private)													
Year	Railroads	%	Buses	%	Air Carriers	%	Inland Waterways	%	Total (except private)	Private Automobiles	Private Airplanes	Total (including private)	
1929	33,965	77.1	6,800	15.4	—	—	3,300	7.5	44,065	175,000	—	219,065	
1939	23,669	67.7	9,100	26.0	683	2.0	1,486	4.3	34,938	275,000	—	309,938	
1944	97,705	75.7	26,920	20.9	2,177	1.7	2,187	1.7	128,989	181,000	1	309,990	
1950	32,481	47.2	26,436	38.4	8,773	12.7	1,190	1.7	68,880	438,293	1,299	508,472	
1960	21,574	28.6	19,327	25.7	31,730	42.1	2,688	3.6	75,319	706,079	2,228	783,626	
1970	10,903	7.3	25,300	16.9	109,499	73.1	4,000	2.7	149,702	1,026,000	9,101	1,184,803	
1980	11,000	4.5	27,400	11.3	204,400	84.2	NA	—	242,800	1,300,400	14,700	1,557,900	
1986	11,800	3.4	23,700	6.9	307,900	89.7	NA	—	343,400	1,450,100	12,400	1,805,900	
1987	12,300	3.4	22,800	6.2	329,100	90.4	NA	—	364,200	1,494,900	12,400	1,871,500	

Note: Railroads includes all classes, including electric railways, Amtrak and Auto-Train.

Source: Transportation Policy Associates

Source: Railroad facts, 1988 Edition, Association of American Railroads

## 82.1 Transportation System Components

The major components of transportation systems are:

1. Links
2. Terminals
3. Vehicles
4. Control systems

Certain "continuous-flow" transportation systems such as pipelines, conveyor belts, and escalators have no discrete vehicles and, in effect, combine the vehicles with the link.

Transportation systems may be developed into extensive networks. The networks may have a hierarchical structure. Thus, highway networks may include freeways, arterials, collector streets, local streets, and driveways. Links and networks may be shared by several transportation modes (e.g., cars, buses, trucks, taxis, bicycles, and pedestrians on local streets). Exclusive lanes may be provided for particular modes (e.g., pedestrian or bicycles) or groups of modes (e.g., buses and carpools).

Transportation terminals provide interfaces among modes or among vehicles of the same mode. They may range from marked bus stops or truck loading zones on local streets to huge airports or ports.

## 82.2 Evaluation Measures

---

Transportation systems are evaluated in terms of their effects on their suppliers, users, and environment. Both their costs and benefits may be classified into supplier, user, and external components. Private transportation companies normally seek to maximize their profits (i.e., total revenues minus total supplier costs). Publicly owned transportation agencies should normally maximize net benefits to their jurisdictions, possibly subject to financial constraints.

From the supplier's perspective, the major indicators of performance include measures of **capacity** (maximum throughput), speed, **utilization rate** (i.e., fraction of time in use), **load factor** (i.e., fraction of maximum payload actually used), energy efficiency (e.g., Btu per ton-mile or per passenger mile), and labor productivity (e.g., worker hours per passenger mile or per ton-mile). Measures of environmental impact (e.g., noise decibels or parts of pollutant per million) are also increasingly relevant. To users, price and service quality measures, including travel time, wait time, access time, reliability, safety, comfort (ride quality, roominess), simplicity of use, and privacy are relevant in selecting modes, routes, travel times, and suppliers.

## 82.3 Air Transportation

---

Air transportation is relatively recent, having become practical for transporting mail and passengers in the early 1920s. Until the 1970s its growth was paced primarily by technological developments in propulsion, aerodynamics, materials, structures, and control systems. These developments have improved its speed, load capacity, energy efficiency, labor productivity, reliability, and safety, to the point where it now dominates long-distance mass transportation of passengers overland and practically monopolizes it over oceans. Airliners have put ocean passenger liners out of business because they are much faster and also, remarkably, more fuel efficient and labor efficient. However, despite this fast growth, the cargo share of air transportation is still small.

For cargoes that are perishable, high in value, or urgently needed, air transportation is preferred over long distances. For other cargo types, ships, trucks, and railroads provide more economic alternatives. In the 1990s, the nearest competitors to air cargo are containerships over oceans and trucks over land. For passengers, air transportation competes with private cars, trains, and intercity buses over land, with practically no competitors over oceans. The growth of air transportation has been restricted to some extent by the availability of adequate airports, by environmental concerns (especially noise), and by the fear of flying of some passengers.

There are approximately 10 000 commercial jet airliners in the world, of which the largest (as of 1995) are Boeing B-747 types, of approximately 800 000 lb gross takeoff weight, with a capacity of 550 passengers. The economic cruising speed of these and smaller "conventional" (i.e., **subsonic**) airliners has stayed at around 560 mph since the late 1950s. A few supersonic transports (SSTs) capable of cruising at approximately 1300 mph were built in the 1970s (the Anglo-French Concorde and the Soviet Tu-144) but, due to high capital and fuel costs, were not economically



successful. About eight Concorde SSTs are still operating, with government subsidies.

The distance that an aircraft can fly depends on its payload, according to the following equation:

$$R = \frac{V}{c'} \left( \frac{L}{D} \right) \ln(W_{TO}/W_L) \quad (82.1)$$

where

- $R$  = range (mi)
- $c'$  = specific fuel consumption (lb fuel/lb thrust  $\times$  h)
- $(L/D)$  = lift-to-drag ratio (dimensionless)
- $W_{TO}$  = aircraft takeoff weight (lb) =  $W_L + W_F$
- $W_L$  = aircraft landing weight (lb) =  $W_E + W_R + P$
- $W_E$  = aircraft empty weight (lb)
- $W_R$  = reserve fuel weight (lb)
- $W_F$  = consumed fuel weight (lb)
- $P$  = payload (lb)

This equation assumes that the difference between the takeoff weight and landing weight is the fuel consumed. For example, suppose that for a Boeing B-747 the maximum payload carried (based on internal fuselage volume and structural limits) is 260 000 lb, maximum  $W_{TO}$  is 800 000 lb,  $W_R = 15$  000 lb,  $W_E = 370$  000,  $L/D = 17$ ,  $V = 580$  mph, and  $c' = 0.65$  lb/lb thrust  $\times$  h. The resulting weight ratio [ $W_{TO}/W_L = 800/(370 + 15 + 260)$ ] is 1.24 and the range  $R$  is 3267 mi. Payloads below is 260 000 allow higher ranges.

Most airline companies fly scheduled routes, although charter services are common. U.S. airlines are largely free to fly whatever routes (i.e., origin-destination pairs) they prefer in the U.S. In most of the rest of the world, authority to serve particular routes is regulated or negotiated by international agreements. The major components of airline costs are direct operating costs (e.g., aircraft depreciation or rentals, aircrews, fuel, and aircraft maintenance) and indirect operating costs (e.g., reservations, advertising and other marketing costs, in-flight service, ground processing of passengers and bags, and administration).

The efficiency and competitiveness of airline service is heavily dependent on efficient operational planning. Airline scheduling is a complex problem in which demand at various times and places, route authority, aircraft availability and maintenance schedules, crew availability and flying restrictions, availability of airport gates and other facilities, and various other factors must all be considered. Airline management problems are discussed in [Wells, 1984].

Airports range from small unmarked grass strips to major facilities requiring many thousands of acres and billions of dollars. Strictly speaking, an airport consists of an airfield (or "airside") and terminal (or "landside"). Airports are designed to accommodate specified traffic loads carried by aircraft up to a "design aircraft," which is the most demanding aircraft to be accommodated. The design aircraft might determine such features as runway lengths, pavement strengths, or terminal gate dimensions at an airport. Detailed guidelines for most aspects of airport design (e.g., runway lengths and other airfield dimensions, pavement characteristics, drainage requirements, allowable noise and other environmental impacts, allowable obstruction heights, lighting, markings, and

signing) are specified by the U.S. Federal Aviation Administration (FAA) in a series of circulars.

Airport master plans are prepared to guide airport growth, usually in stages, toward ultimate development. These master plans:

1. Specify the airport's requirements.
2. Indicate a site if a new airport is considered.
3. Provide detailed plans for airport layout, land use around the airport, terminal areas, and access facilities.
4. Provide financial plans, including economic and financial feasibility analysis.

Major new airports tend to be very expensive and very difficult to locate. Desirable airport sites must be reasonably close to the urban areas they serve yet far enough away to ensure affordable land and acceptable noise impacts. Many other factors—including airspace interference with other airports, obstructions (e.g., hills, buildings), topography, soil, winds, visibility, and utilities—must be reconciled. Hence, few major new airports are being built, and most airport engineering and planning work in the U.S. is devoted to improving existing airports. Governments sometimes develop multi-airport system plans for entire regions or countries.

National agencies (such as the FAA in the U.S.) are responsible for traffic control and airspace management. Experienced traffic controllers, computers, and specialized sensors and communication systems are required for this function. Increasingly sophisticated equipment has been developed to maintain safe operations even for crowded airspace and poor visibility conditions. For the future we can expect increasing automation in air traffic control, relying on precise aircraft location with global positioning satellite (GPS) systems and fully automated landings. Improvements in the precision and reliability of control systems are increasing (slowly) the capacity of individual runways as well as the required separation among parallel runways, allowing capacity increases in restricted airport sites.

## **82.4 Railroad Transportation**

---

The main advantages of railroad technology are low frictional resistance and automatic lateral guidance. The low friction reduces energy and power requirements but limits braking and hill-climbing abilities. The lateral guidance provided by wheel flanges allows railroad vehicles to be grouped into very long trains, yielding economies of scale and, with adequate control systems, high capacities per track. The potential energy efficiency and labor productivity of railroads is considerably higher than for highway modes, but is not necessarily realized, due to regulations, managerial decisions, demand characteristics, or terrain.

The main competitors of railroads include automobiles, aircraft, and buses for passenger transportation, and trucks, ships, and pipelines for freight transportation. To take advantage of their scale economies, railroad operators usually seek to combine many shipments into large trains. Service frequency is thus necessarily reduced. Moreover, to concentrate many shipments, rail cars are frequently re-sorted into different trains, rather than moving directly from origin to destination, which results in long periods spent waiting in classification yards, long delivery times, and poor vehicle utilization. An alternative operational concept relying on direct nonstop "unit trains" is

feasible only when demand is sufficiently large between an origin-destination pair.

Substantial traffic is required to cover the relatively high fixed costs of railroad track. Moreover, U.S. railroads, which are privately owned, must pay property taxes on their tracks, unlike their highway competitors. By 1920 highway developments had rendered low-traffic railroad branch lines noncompetitive in the U.S. Abandonment of such lines has greatly reduced the U.S. railroad network, even though the process was retarded by political regulation.

The alignment of railroad track is based on a compromise between initial costs and operating costs. The latter are reduced by a more straight and level alignment, which requires more expensive earthwork, bridges, or tunnels. Hay [1982] provides design guidelines for railroads.

In general, trains are especially sensitive to gradients. Thus, compared to highways, railroad tracks are more likely to go around rather than over terrain obstacles, which increases the **circuity factors** for railroad transportation.

The resistance for railroad vehicles may be computed using the Davis equation [Hay, 1982]:

$$r = 1.3 + 29/w + bV + CAV^2/wn + 20G + 0.8D \quad (82.2)$$

where

$G$  = gradient (%)

$D$  = degree of curvature

$r$  = unit resistance (lb of force per ton of vehicle weight)

$w$  = weight (tons per axle of car or locomotive)

$n$  = number of axles

$b$  = coefficient of flange friction, swaying, and concussion (0.045 for freight cars and motor cars in trains, 0.03 for locomotives and passenger cars, and 0.09 for single-rail cars)

$C$  = drag coefficient of air [0.0025 for locomotives (0.0017 for streamlined locomotives) and single- or head-end-rail cars, 0.0005 for freight cars, and 0.00034 for trailing passenger cars, including rapid transit]

$A$  = cross-sectional area of locomotives and cars (usually 105 to 120 ft<sup>2</sup> for locomotives, 85 to 90 ft<sup>2</sup> for freight cars, 110–120 ft<sup>2</sup> for multiple-unit and passenger cars, and 70 to 110 ft<sup>2</sup> for single- or head-end-rail cars)

$V$  = speed (mph)

The coefficients shown for this equation reflect relatively old railroad technology and can be significantly reduced for modern equipment [Hay, 1982]. The equation provides the unit resistance in pounds of force per ton of vehicle weight. The total resistance of a railroad vehicle (in lb) is

$$R_v = rwn \quad (82.3)$$

The total resistance of a train  $R$  is the sum of resistances for individual cars and locomotives. The rated horsepower (hp) required for a train is:

$$\text{hp} = \frac{RV}{375\eta} \quad (82.4)$$

where  $\eta$  = transmission efficiency (typically about 0.83 for a diesel electric locomotive).

The hourly fuel consumption for a train may be computed by multiplying hp by a specific fuel consumption rate (approximately 0.32 lb/hp  $\times$  h for a diesel electric locomotive).

Diesel electric locomotives with powers up to 5000 hp haul most trains in the U.S. Electric locomotion is widespread in other countries, especially those with low petroleum reserves. It is especially competitive on high-traffic routes (needed to amortize electrification costs) and for high-speed passenger trains. Steam engines have almost disappeared.

The main types of freight rail cars are box cars, flat cars (often used to carry truck trailers or intermodal containers), open-top gondola cars, and tank cars. Passenger trains may include restaurant cars and sleeping cars. Rail cars have tended toward increasing specialization for different commodities carried, a trend that reduces opportunities for back hauls. Recently, many "double-stack" container cars have been built to carry two tiers of containers. Such cars require a vertical clearance of nearly 20 ft, as well as reduced superelevation (banking) on horizontal curves. In the U.S. standard freight rail cars with gross weights up to 315 000 lb are used.

High-speed passenger trains have been developed intensively in Japan, France, Great Britain, Italy, Germany, and Sweden. The most advanced (in 1995) appear to be the latest French TGV versions, with cruising speeds of 186 mph and double-deck cars. At such high speeds, trains can climb long, steep grades (e.g., 3.5%) without slowing down much. Construction costs in hilly terrain can thus be significantly reduced. Even higher speeds are being tested in experimental railroad and magnetic levitation (MAGLEV) trains.

## 82.5 Highway Transportation

---

Highways provide very flexible and ubiquitous transportation for people and freight. A great variety of transportation modes, including automobiles, buses, trucks, motorcycles, bicycles, pedestrians, animal-drawn vehicles, taxis, and carpools, can share the same roads. From unpaved roads to multilane freeways, roads can vary enormously in their cost and performance. Some highway vehicles may even travel off the roads in some circumstances. The vehicles also range widely in cost and performance, and at their lower range (e.g., bicycles) are affordable for private use even in poor societies.

Flexibility, ubiquity, and affordability account for the great success of highway modes. Personal vehicles from bicycles to private automobiles offer their users great freedom and access to many economic and social opportunities. Trucks increase the freedom and opportunities available to farmers and small businesses. Motor vehicles are so desirable and affordable that in the U.S. the number of registered cars and trucks approximates the number of people of driving age. Other developed countries are approaching the same state despite strenuous efforts to discourage motor vehicle use.

The use of motor vehicles brings significant problems and costs. These include:

1. Road capacity and congestion. Motor vehicles require considerable road space, which is

scarce in urban areas and costly elsewhere. Shortage of road capacity results in severe congestion and delays.

2. Parking availability and cost.
3. Fuel consumption. Motor vehicles consume vast amounts of petroleum fuels. Most countries have to import such fuels and are vulnerable to price increases and supply interruptions.
4. Safety. The numbers of people killed and injured and the property damages in motor vehicle accidents are very significant.
5. Air quality. Motor vehicles are major contributors to air pollution.
6. Regional development patterns. Many planners consider the low-density "sprawl" resulting from motor vehicle dominance to be inefficient and inferior to the more concentrated development produced by mass transportation and railroads.

In the U.S. trucks have steadily increased their share of the freight transportation market, mostly at the expense of railroads, as shown in [Table 82.1](#). They can usually provide more flexible, direct, and responsive service than railroads, but at higher unit cost. They are intermediate between rail and air transportation in both cost and service quality. With one driver required per truck, the labor productivity is much lower than for railroads, and there are strong economic incentives to maximize the load capacity for each driver. Hence, the tendency has been to increase the number, dimensions, and weights allowed for trailers in truck-trailer combinations, which requires increased vertical clearances (e.g., bridge overpasses), geometric standards for roads, and pavement costs.

Various aspects of highway flow characteristics, design standards, and safety problems were presented in **Chapters 79–81**. The main reference for highway design is the AASHTO manual [[AASHTO, 1990](#)]. For capacity, the main reference is the Transportation Research Board *Highway Capacity Manual* [[TRB, 1985](#)]. Extensive software packages have been developed for planning, capacity analysis, geometric design, and traffic control.

Currently (1995), major research and development efforts are being devoted to exploiting advances in information technology to improve highway operations. The Intelligent Transportation Systems (ITS) program of the U.S. Department of Transportation includes, among other activities, an Advanced Traffic Management System (ATMS) program to greatly improve the control of vehicles through congested road networks, an Advanced Travelers Information System (ATIS) program to guide users through networks, and, most ambitiously, an Automated Highway System (AHS) program to replace human driving with hardware. Such automation, when it becomes feasible and safe, has the potential to drastically improve lane capacity at high speeds, by greatly reducing spacing between vehicles. Other potential benefits include reduced labor costs for trucks, buses, and taxis; higher and steadier speeds; improved routings through networks; remote self-parking vehicles; and use of vehicles by nondrivers such as children and handicapped persons. However, substantial technological, economical, and political problems will have to be surmounted.

## 82.6 Water Transportation

---

Water transportation may be classified into (1) marine transportation across seas and (2) inland waterway transportation; their characteristics differ very significantly. Inland waterways consist mostly of rivers, which may be substantially altered to aid transportation. Lakes and artificial canals may also be part of inland waterways. Rivers in their natural states are often too shallow, too fast, or too variable in their flows. All these problems may be alleviated by impounding water behind dams at various intervals. (This also helps generate electric power.) Boats can climb or descend across dams by using **locks** or other elevating systems [Hochstein, 1981]. In the U.S. inland waterway network there are well over 100 major lock structures, with chambers up to 1200 ft long and 110 ft wide. Such chambers allow up to 18 large barges ( $35 \times 195$  ft) to be raised or lowered simultaneously.

In typical inland waterway operations, large diesel-powered "towboats" (which actually push barges) handle a rigidly tied group of barges (a "tow"). Tows with up to 48 barges ( $35 \times 195$  ft, or about 1300 tons/barge) are operated on the lower Mississippi, where there are no locks or dams. On other rivers, where locks are encountered at frequent intervals, tow sizes are adjusted to fit through locks. The locks constitute significant bottlenecks in the network, restricting capacity and causing significant delays.

Table 82.1 indicates that the waterway share of U.S. freight transportation has increased substantially in recent years. This is largely attributable to extensive improvements to the inland waterway system undertaken by the responsible agency, the U.S. Army Corps of Engineers.

The main advantage of both inland waterway and marine transportation is low cost. The main disadvantage is relatively low speed. Provided that sufficiently deep water is available, ships and barges can be built in much larger sizes than ground vehicles. Ship costs increase less than proportionally with ship size, for ship construction, crew, and fuel. Energy efficiency is very good at low speeds [e.g., 10–20 knots (nautical mi/h)]. However, at higher speeds the wave resistance of a conventional ship increases with the fourth power of speed ( $V^4$ ). Hence, the fuel consumption increases with  $V^4$  and the power required increases with  $V^5$ . Therefore, conventional-displacement ships almost never exceed 30 knots in commercial operation. Higher practical speed may be obtained by lifting ships out of the water on hydrofoils or air cushions. However, such unconventional marine vehicles have relatively high costs and limited markets at this time. Over time, ships have increased in size and specialization. Crude oil tankers of up to 550 000 tons (of payload) have been built. Tankers carrying fluids are less restricted in size than other ships because they can pump their cargo from deep water offshore without entering harbors. Bulk carriers (e.g., for coal, minerals, or grains) have also been built in sizes exceeding 300 000 tons. They may also be loaded through conveyor belts built over long pier structures to reach deep water. General cargo ships and containerhips are practically always handled at shoreline berths and require much storage space nearby.

The use of intermodal containers has revolutionized the transportation of many cargoes. Such containers greatly reduce the time and cost required to load and unload ships. Up to 4500 standard 20-ft containers ( $20 \times 8 \times 8$  ft) can be carried at a speed of about 24 knots on recently built containerhips.

Port facilities for ships should provide shelter from waves and sufficiently deep water, including approach channels to the ports. In addition, ports should provide adequate terminal facilities, including loading and unloading equipment, storage capacity, and suitable connections to other

transportation networks. Ports often compete strenuously with other ports and strive to have facilities that are at least equal to those of competitors. Since ports generate substantial employment and economic activities, they often receive financial and other support from governments.

Geography limits the availability of inland waterways and the directness of ship paths across oceans. Major expensive canals (e.g., Suez, Panama, Kiel) have been built to provide shortcuts in shipping routes. These canals may be so valuable that ship dimensions are sometimes compromised (i.e., reduced) to fit through these canals. In some parts of the world (e.g., Baltic, North Sea, most U.S. coasts) the waters are too shallow for the largest ships in existence. Less efficient, smaller ships must be used there. The dredging of deeper access channels and ports can increase the allowable ship size, if the costs and environmental impacts are considered acceptable.

## 82.7 Public Transportation

---

Public transportation is the term for ground passenger transportation modes available to the general public. It connotes public availability rather than ownership. "Conventional" public transportation modes have fixed routes *and* fixed schedules and include most bus and rail transit services. "Unconventional" modes (also labeled "paratransit") include taxis, carpools and van pools, rented cars, dial-a-ride services, and subscription services.

The main purposes of public transportation services, especially conventional mass transportation services in developed countries, are to provide mobility for persons without automobiles (e.g., children, poor, nondrivers); to improve the efficiency of transportation in urban areas; to reduce congestion effects, pollution, accidents, and other negative impacts of automobiles; and to foster preferred urban development patterns (e.g., strong downtowns and concentrated rather than sprawled development).

Conventional services (i.e., bus and rail transit networks) are quite sensitive to demand density. Higher densities support higher service frequencies and higher network densities, which decrease user wait times and access times, respectively. Compared to automobile users, bus or rail transit users must spend extra time in access to and from stations and in waiting at stations (including transfer stations). Direct routes are much less likely to be available, and one or more transfers (with possible reliability problems) may be required. Thus, mass transit services tend to be slower than automobiles unless exclusive rights-of-way (e.g., bus lanes, rail tunnels) can favor them. Such exclusive rights-of-way can be quite expensive if placed on elevated structures or in tunnels. Even when unhindered by traffic, average speeds may be limited by frequent stops and allowable acceleration limits for standing passengers. Prices usually favor mass transit, especially if parking for automobiles is scarce and expensive.

The capacity of a transit route can be expressed as:

$$C = FLP \quad (82.5)$$

where

$C$  = one-way capacity (passengers/hour) past a certain point

$F$  = service frequency (e.g., trains/hour)

$L$  = train length (cars/train)

$P$  = passenger capacity of cars (spaces/car)

For rail transit lines where high capacity is needed in peak periods,  $C$  can reach 100000 passengers/hour (i.e., 40 trains/hour  $\times$  10 cars/train  $\times$  250 passenger spaces/car). There are few places in the world where such capacities are required. For a bus line the train length  $L$  would usually be 1.0. If no on-line stops are allowed, an exclusive bus lane also has a large capacity (e.g., 1000 buses/hour  $\times$  90 passenger spaces/bus), but such demand levels for bus lanes have not been observed.

The average wait time of passengers on a rail or bus line depends on the headway, which is the interval between successive buses or trains. This can be approximated by:

$$\overline{W} = \overline{H}/2 + \text{var}(H)/2\overline{H} \quad (82.6)$$

where

$\overline{W}$  = average wait time (e.g., minutes)

$\overline{H}$  = average headway (e.g., minutes)

$\text{var}(H)$  = variance of headway (e.g., minutes<sup>2</sup>)

It should be noted that the headway is the inverse of the service frequency.

The number of vehicles  $N$  required to serve a route is:

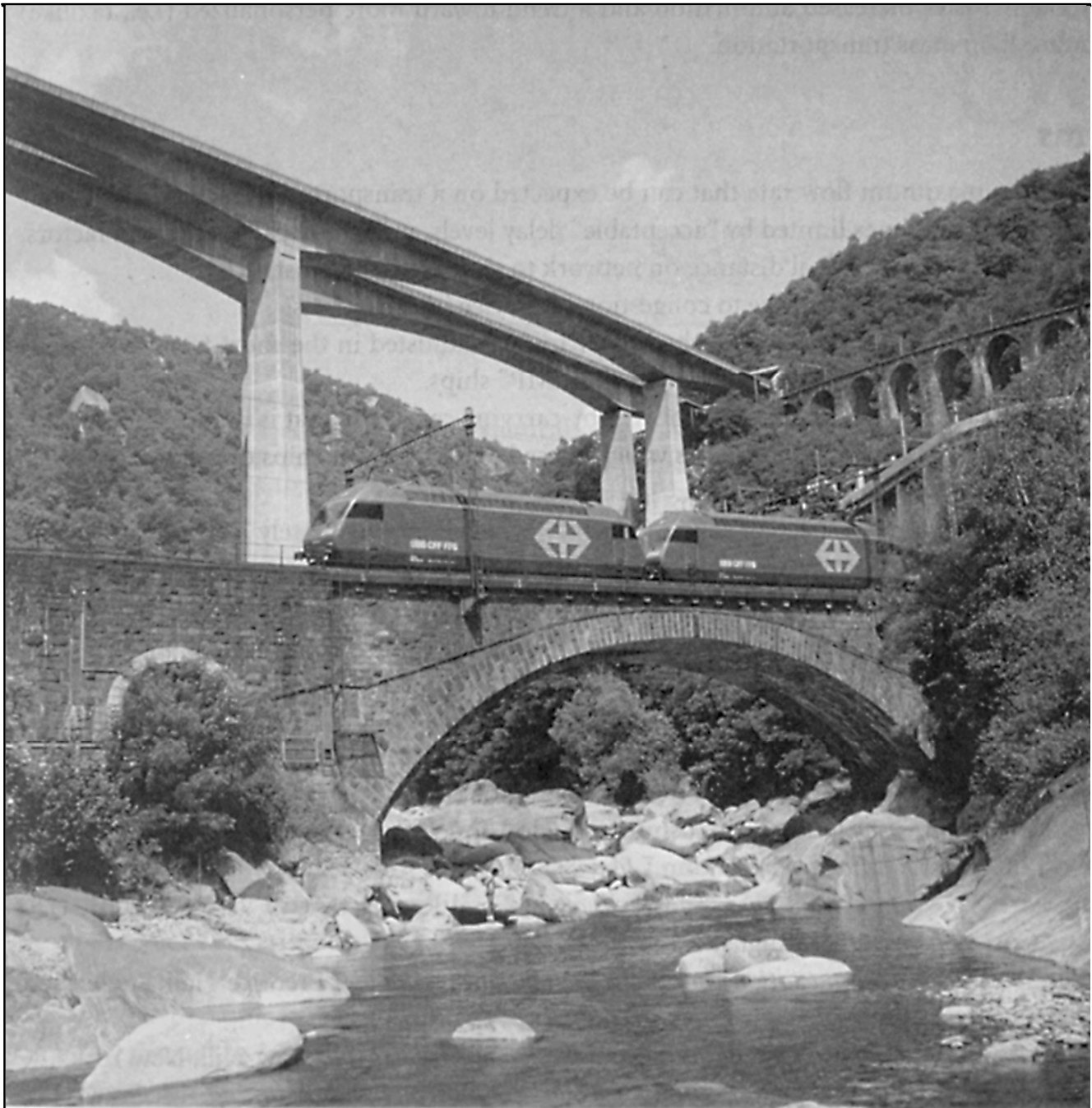
$$N = RFL \quad (82.7)$$

where  $R$  = vehicle round trip time on route (e.g., hours).

The effectiveness of a public transportation system depends on many factors, including demand distribution and density, network configuration, routing and scheduling of vehicles, fleet management, personnel management, pricing policies, and service reliability. Demand and economic viability of services also depend on how good and uncongested the road system is for automobile users.

Engineers can choose from a great variety of options for propulsion, support, guidance and control, vehicle configurations, facility designs, construction methods, and operating concepts. New information and control technology can significantly improve public transportation systems. It will probably foster increased automation and a trend toward more personalized (i.e., taxilike) service rather than mass transportation.





## THE GOTTHARD PASS

*B. van Gelder, Purdue University*

St. Gotthard is a pass in the Central Alps connecting Switzerland and Italy, 2108 m above sea level. Road and railroad bridges and eventually the Gotthard tunnel greatly facilitate traffic between the Swiss canton Uri to the north and the canton of Ticino to the south. On a larger scale, these roads form the Central Traffic Route between Northern Europe and Italy.

The roads and railroads to the north cut through the valley of the Reuss River, where the city of Wassen is situated. Over the St. Gotthard pass or through the St. Gotthard tunnel the roads and railroads finally reach, to the south, the Leventina valley with the Ticino River and the city of Faido. Near the latter city the depicted modern traffic bridge and the older railroad bridge are situated. (Photo courtesy of the Swiss National Tourist Office.)

## Defining Terms

**Capacity:** The maximum flow rate that can be expected on a transportation facility. "Practical" capacity is sometimes limited by "acceptable" delay levels, utilization rates, and load factors.

**Circuitry factor:** Ratio of actual distance on network to shortest airline distance.

**Delay:** Increase in service time due to congestion or service interruptions.

**Demand-responsive:** A mode whose schedule or route is adjusted in the short term as demand varies, such as taxis, charter airlines, and "TRAMP" ships.

**Load factor:** Fraction of available space or weight-carrying capability that is used.

**Lock:** A structure with gates at both ends which is used to lift or lower ships or other vessels.

**Mode:** A distinct form of transportation.

**Subsonic:** Flying below the speed of sound (Mach 1), which is approximately 700 mph at cruising altitudes of approximately 33 000 ft.

**Utilization rate:** Fraction of time that a vehicle, facility, or equipment unit is in productive use.

## References

- AASHTO (American Society of State Highway and Transportation Officials). 1990. *A Policy on Geometric Design of Highways and Streets*. Washington, DC.
- Brun, E. 1981. *Port Engineering*. Gulf Publishing Co., Houston.
- Hay, W. W. 1982. *Railroad Engineering*. John Wiley & Sons, New York.
- Hochstein, A. 1981. *Waterways Science and Technology*, Final Report DACW 72-79-C-0003. U.S. Army Corps of Engineers, August.
- Homburger, W. S. 1982. *Transportation and Traffic Engineering Handbook*. Prentice-Hall, Englewood Cliffs, NJ.
- Horonjeff, R. and McKelvey, F. 1994. *Planning and Design of Airports*. McGraw-Hill, New York.
- Morlok, E. K. 1976. *Introduction to Transportation Engineering and Planning*. McGraw-Hill, New York.
- TRB (Transportation Research Board). 1985. *Highway Capacity Manual*. Special Report 209. TRB, Washington, DC.
- Vuchic, V. 1981. *Urban Public Transportation*. Prentice-Hall, Englewood Cliffs, NJ.
- Wells, A. T. 1984. *Air Transportation*. Wadsworth Publishing Co., Belmont, CA.
- Wright, P. H. and Paquette, R. J. 1987. *Highway Engineering*. John Wiley & Sons, New York.

## Further Information

The ITE Handbook [[Homburger, 1992](#)] and Morlok [[1978](#)] cover most transportation modes. Horonjeff and McKelvey [[1994](#)], Hay [[1982](#)], Wright and Paquette [[1987](#)], Brun [[1981](#)], and Vuchic [[1981](#)] are more specialized textbooks covering airports, railroads, highways, ports, and urban public transportation systems, respectively. Periodicals such as *Aviation Week & Space Technology*, *Railway Age*, *Motor Ship*, and *Mass Transit* cover recent developments in their subject areas.

Khalil, T. B. "Safety Analysis"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

### 83.1 Mathematical Models

Lumped Parameter Models • Hybrid Models • Finite Element Models

### 83.2 Summary

#### **Tawfik B. Khalil**

*General Motors Corp.*

Safety of transportation systems, including land, water, and space vehicles, can be defined as the ability of the vehicle structure to provide sufficient protection to mitigate occupants' harm and to reduce cargo damage in the event of a crash. This goal is typically achieved by a combination of structural **crashworthiness** to manage the crash energy and by a system of restraints within the passenger compartment to minimize the impact forces on the human body during the second collision. Crash energy management is viewed here as absorption of the crash kinetic energy of the vehicle while maintaining sufficient resistance to sustain the passenger compartment integrity.

Safety studies for land motor vehicles, the subject discussed in this chapter, are accomplished by a combination of experimental and analytical techniques. Experimental techniques involve sled tests, in which mechanical surrogates of humans (**anthropomorphic** test devices, or "dummies") are subjected to dynamic loads similar to a vehicle deceleration–time pulse to study occupant response, in either frontal or lateral impact modes. The measured dummy kinematics and associated loads (moments) provide a measure of the impact severity and the effectiveness of the restraint system in reducing loads transferred to the occupant. Another type of test typically run to ensure total vehicle structural integrity (crashworthiness) and compliance with government-mandated regulations is the full-scale frontal vehicle to barrier impact. In this test a fully instrumented vehicle with a dummy in the driver seat is launched to impact a rigid barrier from an initial velocity of 30 mph. Other tests include side impacts, rear impacts, and rollover simulations.

Such experimental studies are not only time consuming but also expensive, particularly at the early stages of design, where only prototype vehicles are available. The need to simulate the crash event by an analytical procedure is obvious. This chapter addresses the use of analytical techniques in design of transportation systems, with particular emphasis on motor vehicles.

## 83.1 Mathematical Models

---

Developing mathematical models for structural crashworthiness and occupant response to impact is conceptually easy; it involves solving a set of partial differential equations that govern the response of structures to dynamic loads, subject to initial and boundary conditions. In practice, however, the problem is complicated due to the following factors:

- The crash process is a dynamic event persisting for a short duration of approximately 100–200 milliseconds.
- Vehicle structures are typically complex in geometry, manufactured from metallic and composite shell components, and assembled by various fastening techniques.
- Biomechanical simulation of human or mechanical surrogates to impact requires extensive knowledge of human anthropometry, biological tissue properties, and human tolerance to impact.
- The governing equations are highly nonlinear due to large deformations, large rotations, buckling, elastic-plastic rate-dependent material response, and contact and folding in the shell structures.

Given the aforementioned factors, it is not surprising that analytical simulations of vehicle collisions and occupant response to impact have been evolving over the last 25 years. Three types of models are used in safety simulations: **lumped parameter models** (LP), **hybrid models** (HM), and *finite element* (FE) models.

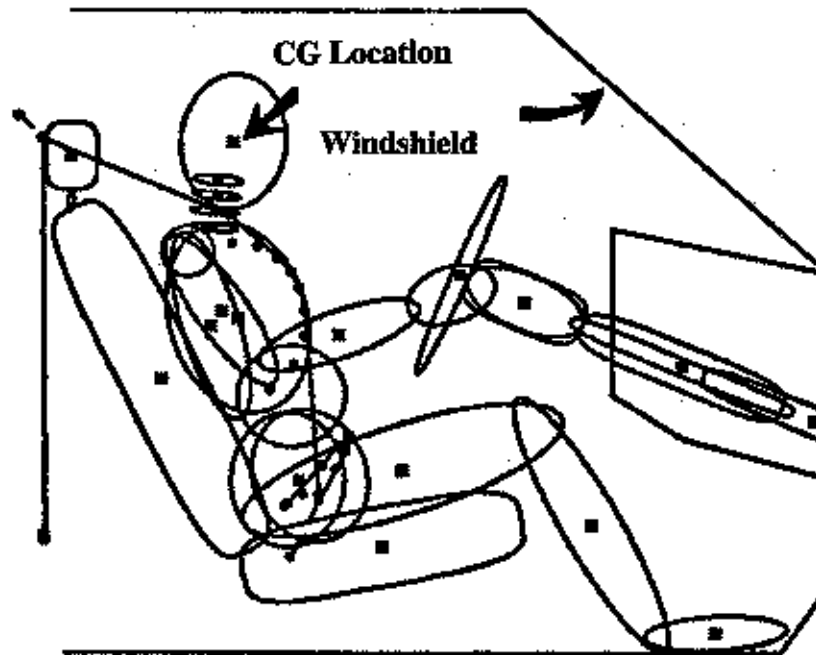
### Lumped Parameter Models

The first vehicle structure LP model was developed [Kamal, 1970] by using lumped spring-mass components. It simulated the vehicle response to frontal impact into a rigid wall by a unidimensional model consisting of three masses, which represented the inertia of the vehicle body, engine, transmission, and drive shafts. The masses are interconnected by nonlinear springs to simulate the compliance of the vehicle structure. The force-deformation characteristics of the springs are determined by quasi-static crush of vehicle components, which incidentally require significant experience on the behavior of thin sheet metal structural components subject to large deformations and various end conditions (e.g., fixed, hinged, or free). This type of model is still widely used by crash engineers because of its simplicity and surprisingly relative accuracy in comparison with test data. In fact, this modeling approach has been extended to simulate side impact collisions between two vehicles. It is important to note, however, that developing such models relies on experimental data and extensive experience on structural behavior in crash environments. Further, translating model parameters into design data is not immediately obvious.

Two-dimensional and three-dimensional LP models [Prasad and Chou, 1989] are also developed to simulate occupant response to deceleration pulses generated by vehicle structures. These models consist of a group of masses that simulate the inertia and CG locations of anthropomorphic dummies used in crash testing. The dummy segments are connected by joints with appropriate moment-rotation and force-deformation characteristics to represent biomechanical human articulation. Interactions between the dummy model and the passenger interior compartment are achieved by specifying force-deformation curves between the dummy segment and the potential

contact target. Figure 83.1 shows an example of a three-dimensional average size dummy model used in crash simulations. Similar to LP models of vehicle structure, the occupant models are relatively simple to develop and not computationally demanding. In fact, all LP models can be run in minutes on an engineering workstation or a personal computer.

**Figure 83.1** Three-dimensional LP model of a seated dummy, represented by ellipsoidal rigid bodies interconnected by appropriate joints.



## Hybrid Models

Hybrid models were developed to remedy the limitations inherent in the LP spring-mass models. The modeling technique, simply, recommended calculation of the force-deformation component response from structural mechanics equations, instead of testing, which would subsequently be used as the spring property in the LP model. The recommended components initially were generic S-shaped hollow beams built from thin sheet metal [Ni, 1981]. Two beams typically represent the lower or mid-rails (also referred to as *torque boxes*) of the vehicle and represent the main longitudinal load-path carriers from the bumper to the vehicle body. The analysis was accomplished by a finite difference solution, which treated the shell structure by a series of beams with appropriate geometry and material models. Plastic behavior at the hinges was accomplished



by a moment-rotation curve, derived experimentally. Although the component response can be calculated in three dimensions, due to the inability of the LP technology to superimpose path-dependent plastic deformations, the analysis allowed only uniaxial deformations. Further, the boundary conditions at the spring ends cannot be accurately represented. These limitations rendered the technique approximate. Yet, it is commonly used in vehicle design due to its simplicity and to its low requirements of computer resources.

## Finite Element Models

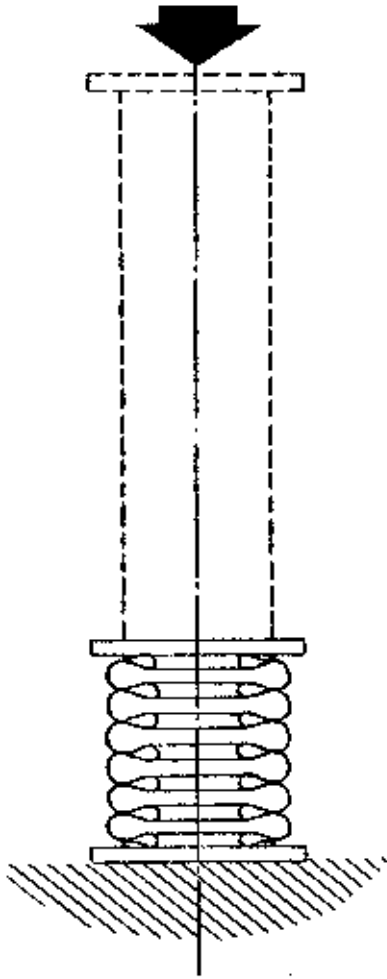
There are two types of FE models, discussed in the following sections, that are used in structural crashworthiness:

### Heuristic Beam Models

Heuristic models (semianalytical models) are formulated by complementing the equations of mechanics with experimental data and empirical information. These models are developed to provide design guidelines for vehicle structure at the early stages in vehicle conception. Four types of models are constructed and analyzed in parallel to investigate the synergy between the major collision modes, namely, front, rear, side, and rollover impacts. At the early stage in vehicle design, crashworthiness is considered in parallel with other design requirements, such as packaging, vehicle dynamics, noise and vibrations, and so on. Accordingly, a computationally efficient scheme along with a fast process to build models is necessary. This led to the development of FE beam models [[Mahmood and Paluzeny, 1986](#)], which define all major components of the vehicle skeleton by means of beam elements. With skill and experience, the influence of connecting panels can be included in the analysis.

The basic building blocks of these models are structural members that are referred to as *columns* when subjected to uniaxial compression and as *beams* when subjected to bending deformations. In either case these components are manufactured from stamped thin sheet metal. Column members are typically exposed to axial or slightly off-angle loads, which can produce progressive (accordion) crush, as shown in [Fig. 83.2](#). This type of progressively stable collapse, highly desired in energy absorption, requires a compressive load smaller than the Euler buckling load and larger than the plastic yield of the column plates. Unstable column collapse, which can include some folding, is less efficient in absorbing energy.

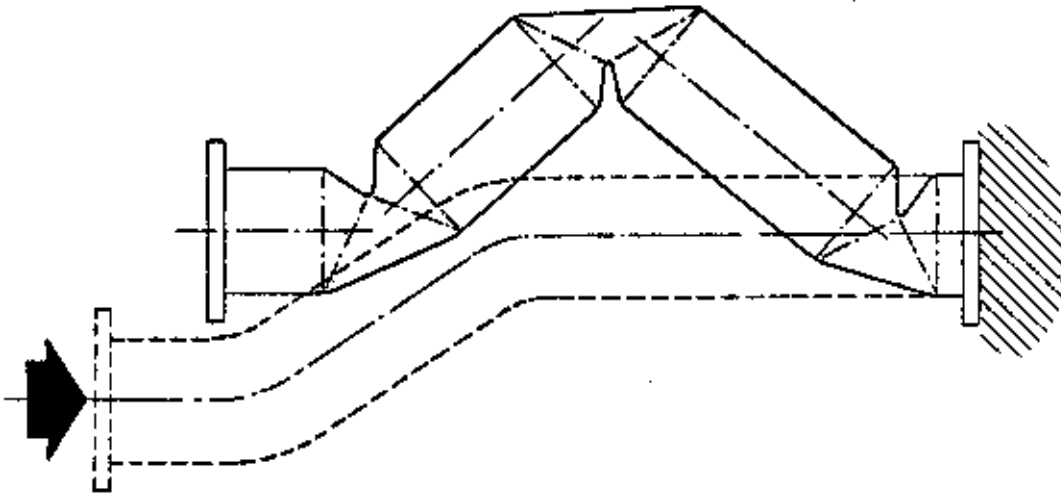
**Figure 83.2** Schematic of axial compression of a straight column, showing progressive formation of accordion folds.



Beam bending is the dominant mode of collapse in many vehicle structures, due to its need for the least amount of energy. Collapse by pure bending is rare. In most vehicle crashes, structural deformations involve a combination of axial compression, bending, and torsion. Component collapse will initiate where the compressive stress exceeds the material yield/local-buckling strength by forming a plastic hinge. The structure cannot continue to support an increasing load at the hinge and stress redistribution occurs, followed by the formation of more plastic hinges. This continues until eventually the structure evolves into a kinematically movable framework, as shown in [Fig. 83.3](#). Therefore, it is important that the model captures the plastic hinge formations and subsequent linkage kinematics effects.



**Figure 83.3** Schematic of S-rail bending deformations, showing plastic hinge formations and subsequent linkage-like kinematics.



Approximate formulas, based on a force method or a displacement method, are derived [Mahmood and Paluszny, 1986] to determine the peak and average crush loads on the basis of local buckling and plastic yielding of thin-walled columns and beams. The component geometry is subdivided into plate elements, which are joined by nodes. A computer program is developed to determine the maximum load-carrying capacity of vehicle structures and subsequent energy absorption following large deformation collapse of the structure. The computation can account for compression and biaxial bending deformations.

### Analytically Based FE Models

All previously cited models require some prior knowledge of the potential failure mechanism of the structure, in addition to experimental data as input to the model. It has always been the desire of safety engineers to develop analytically based models for vehicle crash and occupant dynamics simulations. These models should be based on the physical process involved in the crash event—geometry of the structure, basic stress-strain response of the material, initial conditions of impact and boundary/constraint conditions. This type of analysis—known in mechanics as *initial-boundary value problem*—requires the solution of a nonlinear, coupled system of partial differential equations that can only be applied to extremely simple geometries. Accordingly, classical closed form solutions are nearly inapplicable to real-world structural mechanics problems.

FE technology was introduced in the early 1960s for linear structural analysis, in which the geometry of the structure can be discretized into a set of idealized substructures, called *elements*. Several Fortran computer codes were developed, for both research and commercial applications. Application of the FE technology to crashworthiness analysis did not gain serious momentum until the mid-1980s, due to accelerated advances in explicit time-integration nonlinear FE technology

[Liu *et al.*, 1986] and due to the development of FE codes with reduced integrated elements for spatial discretization and versatile contact algorithms [Goudreau and Hallquist, 1982]. In addition, the introduction of supercomputers provided the necessary impetus for applying the technology to practical problems.

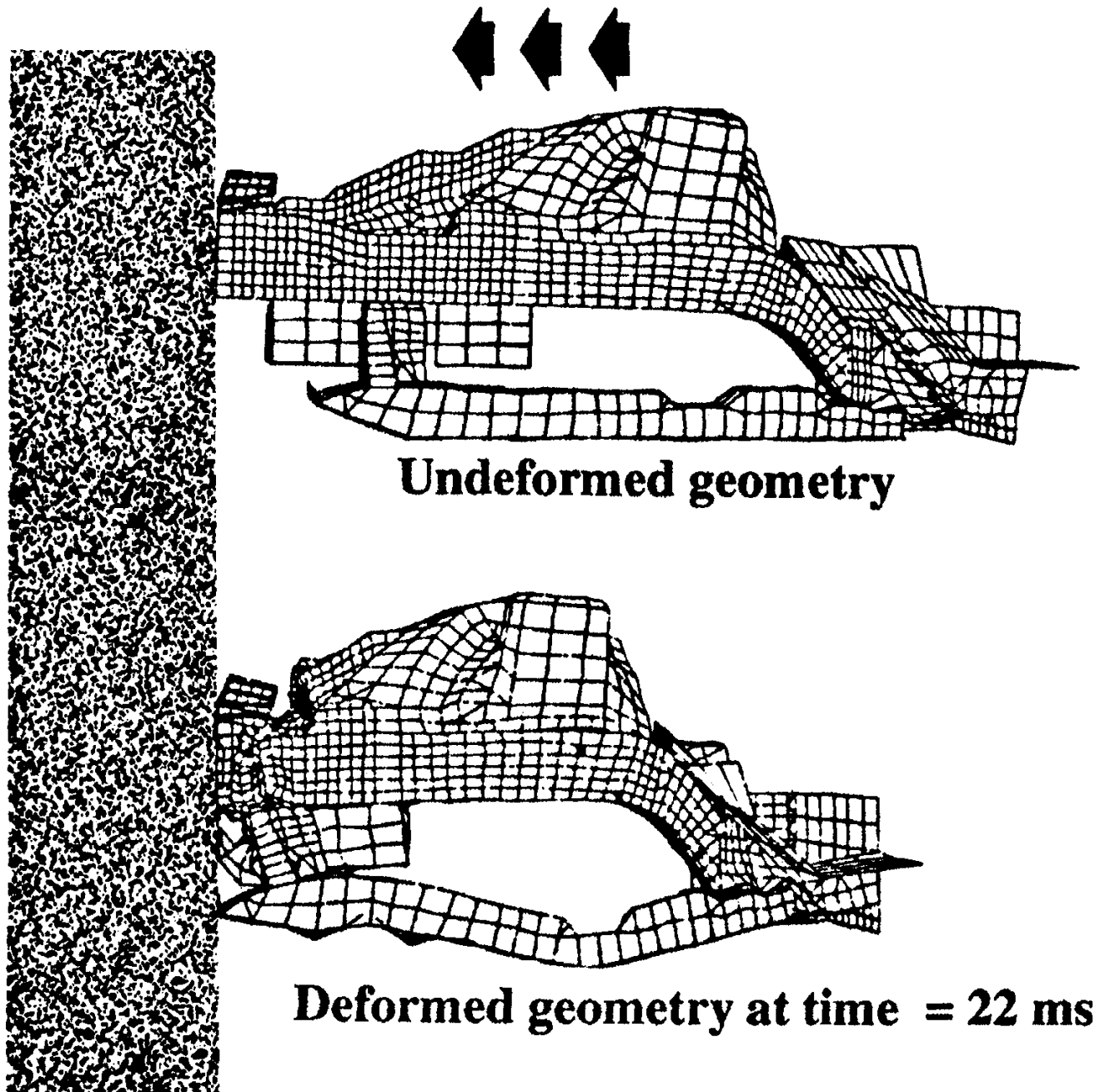
FE crash codes are based on updated Lagrangian mechanics. The equations of motion are obtained from stating the balance of linear momentum in an integral form and introducing spatial discretization by linear isoparametric elements. The semidiscretized second-order set of equations of motion can be written as

$$\mathbf{M}\mathbf{a} = \mathbf{P}(x, t) - \mathbf{Q}(x, t)$$

where  $\mathbf{M}$  is the diagonal mass matrix,  $\mathbf{a}$  is the acceleration vector,  $\mathbf{P}$  is the external force vector,  $\mathbf{Q}$  is the nodal internal force vector,  $x$  is a spatial coordinate, and  $t$  is time. The solution of the previous set of equations in time is accomplished by the explicit central difference technique. The integration scheme, though conditionally stable, has the advantage of avoiding implicit integration and iterative solution of the stiffness matrix.

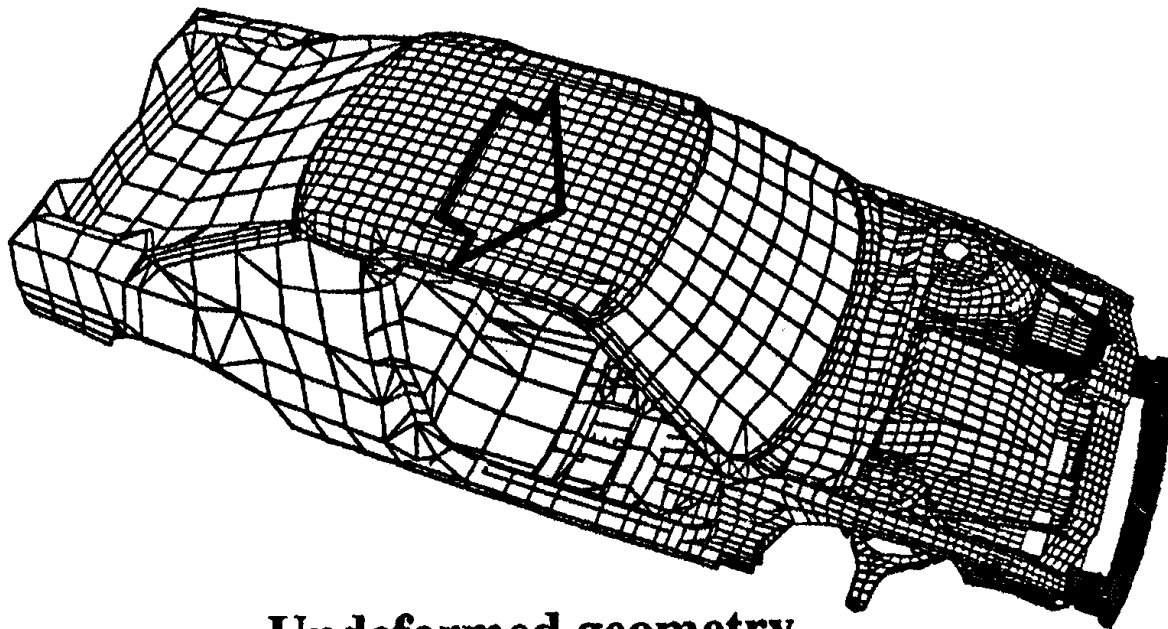
Initially, this technology was applied to analyze generic components (columns and S-rails) with emphasis on analytically capturing the plastic hinge formation, peak load, sustained collapse load, and associated energy absorption. Following this, a number of simulations modeled structural components manufactured from thin sheet metal and assembled by spot welding [Khalil and Vander Lugt, 1989]. Figure 83.4 shows the initial and deformed configurations of the front-end structure of an experimental vehicle launched to impact a rigid wall from an initial velocity of 50 km/h. The model simulated the structure by predominantly quadrilateral shell elements, which allows for both bending and membrane deformations. Elastic-plastic material properties with appropriate strain hardening and rate effects were assigned to the shell elements. Constraint conditions were used to tie the shell nodes where spot welds were used. A single-surface contact definition was specified for the frontal part of the structure to allow for sheet metal stacking without penetration. The predicted peak force from this impact was 250 kN, which agreed quite well with test data.

**Figure 83.4** FE model of a vehicle structure front end, launched to impact a rigid barrier from an initial velocity = 50 km/h.

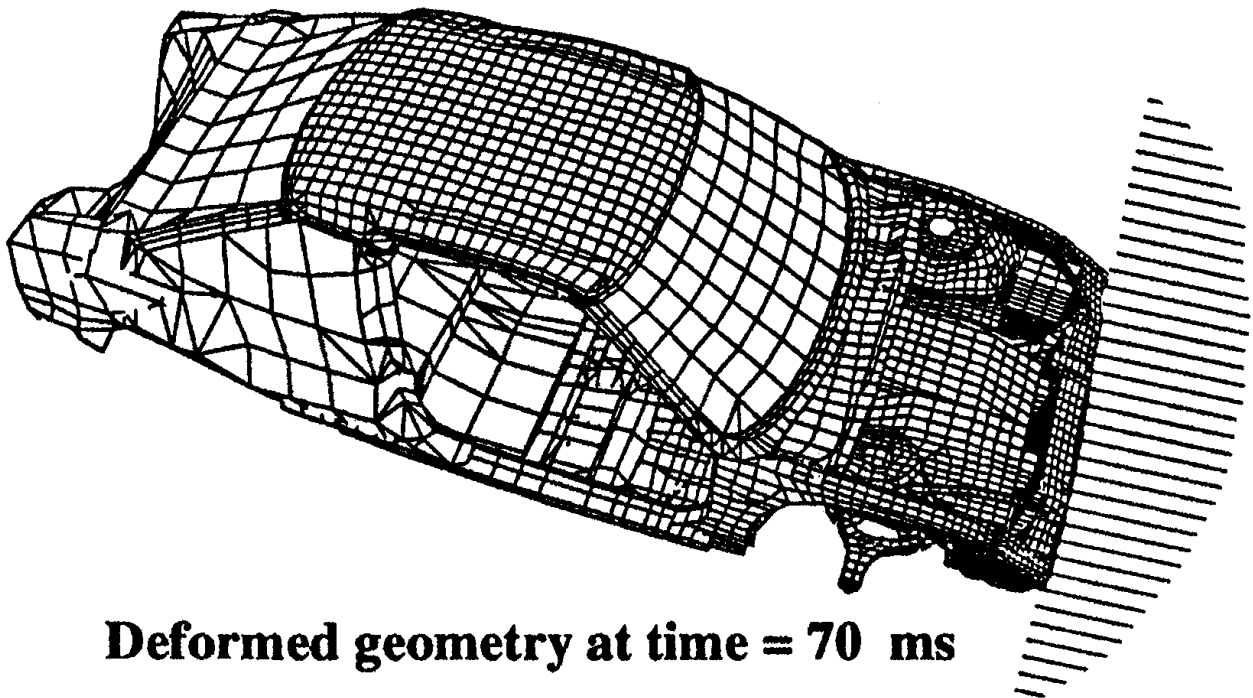


After advances in the ability of FE technology to simulate subsystem impact response, the methodology was extended to simulate full-scale vehicle collisions. Current simulations include frontal vehicle collision with a rigid barrier, commonly conducted for compliance with federal safety standards. [Figure 83.5](#) shows an FE model of a vehicle structure before and after impact with a rigid barrier from an initial velocity of 50 km/h [[Johnson and Skynar, 1989](#)]. Other models of vehicle structures simulate a movable deformable barrier impacting the side of a stationary vehicle. Also, simulations of vehicle-to-vehicle frontal impact as well as rear impact have been successfully attempted.

**Figure 83.5** FE model of frontal vehicle collision with a rigid barrier, initial velocity = 50 km/h.



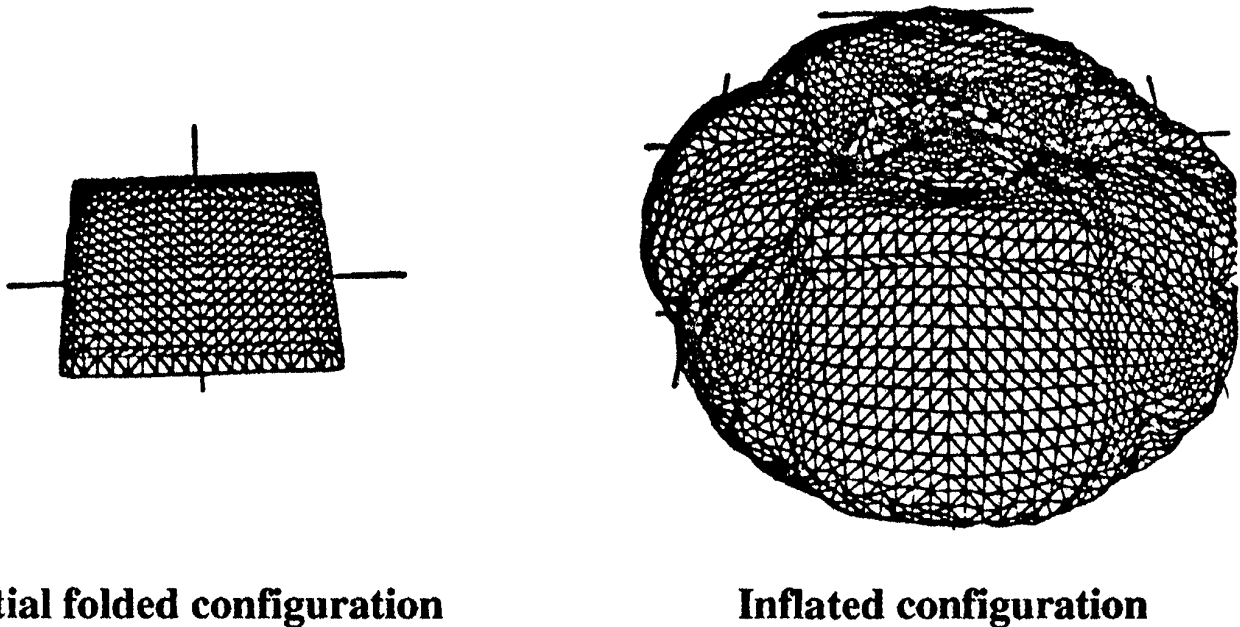
**Undeformed geometry**



**Deformed geometry at time = 70 ms**

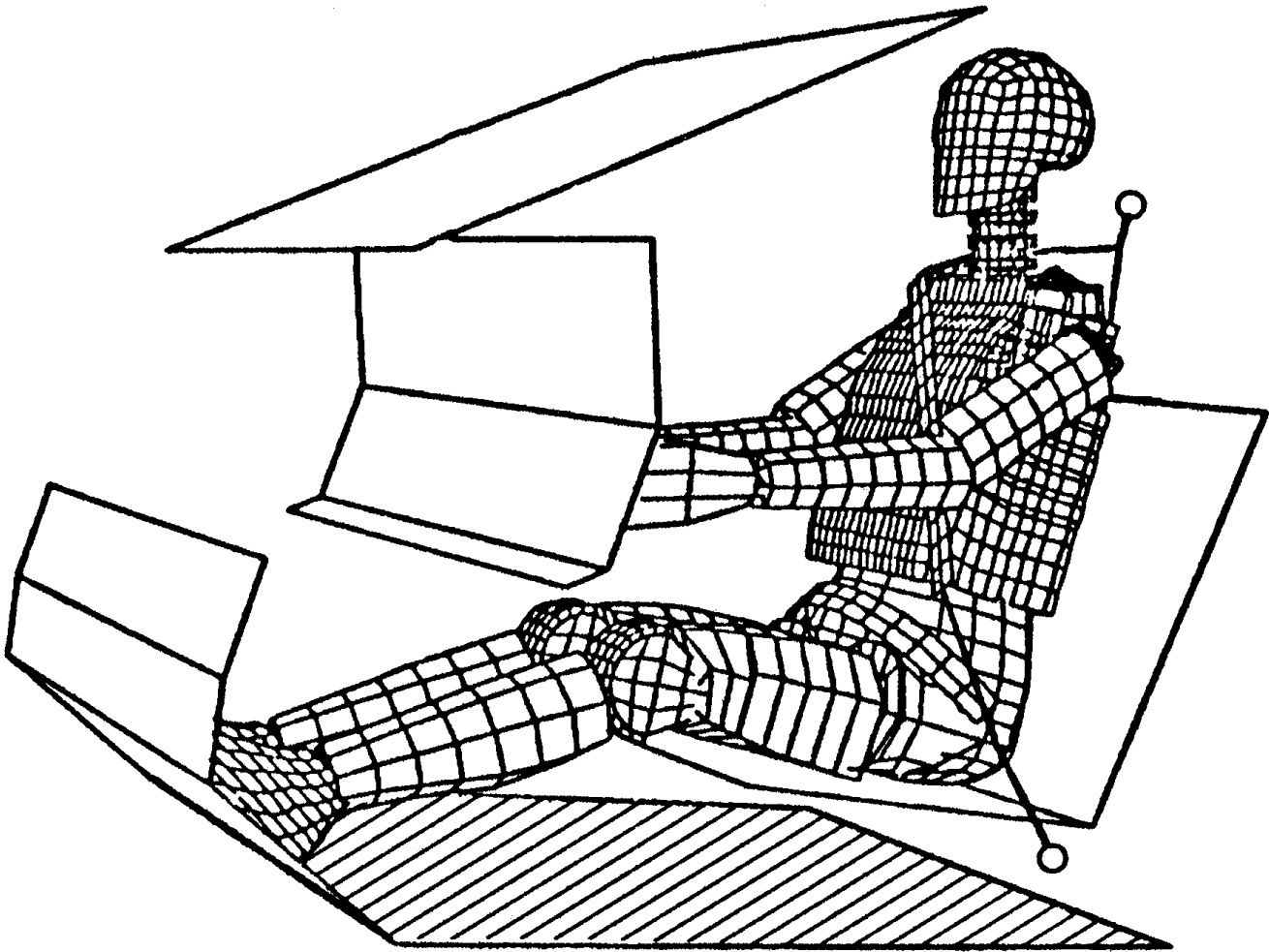
Recently, the technology has been extended to simulate air bag inflation and deployment of driver-side and passenger-side air bags. [Figure 83.6](#) shows an isolated, folded driver-side air bag in its initial configuration and its shape subsequent to inflation. Of particular interest here, from the FE simulation point of view, is the technology's ability to simulate the bag fabric material, which can sustain tension and no compression; the gas dynamics and their interaction with the bag to allow for pressurization and controlled leakage; and, finally, the contact and interactions among the bag layers, which should allow for deployment without penetration.

**Figure 83.6** FE simulation of driver-side air bag inflation.



The limitation of using LP models in simulating occupants has been recognized for some time. With the success demonstrated in simulating vehicle structures, analysts were encouraged to extend FE analysis to simulate occupant interactions with interior passenger compartments [[Khalil and Lin, 1991](#)]. [Figure 83.7](#) shows an FE model of a dummy used to simulate an occupant in crash testing.

**Figure 83.7** FE representation of a seated dummy with three-point belt harness.



## 83.2 Summary

---

Several analytical techniques are used in safety analysis to simulate vehicle structural response and occupant behavior in a crash environment. Early models include lumped-parameter, hybrid, and FE heuristic beam models. These models are characterized by gross geometric approximations, and, consequently, they are quick to develop and require minimum computer resources that can be provided by a PC. In the past seven years, detailed representations of vehicle structures by FE models have evolved in size and complexity from geometries represented by 2800 shell elements with one or two material models to current models simulated by over 50 000 shell, solid, and beam elements. These models also include several material representations for metallic and nonmetallic components. Currently, vehicle models exist in the open literature for frontal, side, and rear-impact simulations. Also, modeling of occupant interactions with inflatable restraint systems has recently been published and discussed. It is anticipated that in the near future (within five years) system models representing vehicle structures, occupants, and restraint systems will be in the neighborhood of 100 000 elements and will become a routine design tool in the transportation

industry. However, this increase in model size—coupled with demands for lighter vehicles manufactured from materials such as aluminum and composites—will require new developments in FE technology and hardware architecture to allow for reducing model development effort and computation time.

## Defining Terms

**Anthropomorphic:** Describes a mechanical manikin that possesses geometric, inertial, and material characteristics similar to a human's.

**Crashworthiness:** The ability of a vehicle structure to absorb mechanically energy resulting from collision with another object while maintaining integrity of the passenger compartment.

**Heuristic model:** A model formulated from discrete deformable elements with built-in empirical knowledge.

**Hybrid model:** A lumped parameter model in which the discrete springs are replaced by deformable components.

**Lumped parameter model:** A mechanical system model that represents a continuum structure by discrete masses, springs, and dampers.

## References

- Goudreau, G. L. and Hallquist, J. O. 1982. Recent developments in large-scale finite element Lagrangian hydrocode technology. *Comp. Methods Appl. Mech. Eng.* 33: 725–757.
- Johnson, J. P. and Skynar, M. J. 1989. Automotive crash analysis. In *Crashworthiness and Occupant Protection in Transportation Systems*, ed. T. B. Khalil and A. I. King, p. 27–33. AMD-Vol. 106, BED-Vol. 13. ASME, New York.
- Kamal, M. M. 1970. Analysis and simulation of vehicle to barrier impact. *SAE*. 700414:1498–1503.
- Khalil, T. B. and Lin, K. H. 1991. Hybrid III thoracic impact of self-aligning steering wheel by finite element analysis and mini-sled tests. In *35th Stapp Car Crash Conference Proceedings*. Paper No. 912894. SAE, Warrendale, PA.
- Khalil, T. B. and Vander Lugt, D. A. 1989. Identification of vehicle front structure crashworthiness by experiments and finite element analysis. In *Crashworthiness and Occupant Protection in Transportation Systems*, ed. T. B. Khalil and A. I. King, p. 41–53. AMD-Vol. 106, BED-Vol. 13. ASME, New York.
- Liu, W. K., Belytschko, T., and Chang, H. 1986. An arbitrary Lagrangian-Eulerian finite element method for path-dependent materials. *Comp. Methods Appl. Mech. Eng.* 58: 227–245.
- Mahmood, H. F. and Paluzeny, A. 1986. Analytical technique for simulating crash response of vehicle structures composed of beam elements. In *Sixth International Conference on Vehicle Structural Mechanics*. Paper No. 860820. SAE, Warrendale, PA.
- Ni, C. M. 1981. A general purpose technique for nonlinear dynamic response of integrated structures. In *Fourth International Conference on Vehicle Structural Mechanics*. SAE, Warrendale, PA.
- Prasad, P. and Chou, C. C. 1989. A review of mathematical occupant simulation models. In

*Crashworthiness and Occupant Protection in Transportation Systems*, ed. T. B. Khalil and A. I. King, p. 95–113. AMD-Vol. 106. ASME, New York.

### **Further Information**

ASME Winter Annual Meeting Proceedings: published annually by the Applied Mechanics Division.

Vehicle Structures Mechanics Conference: published biannually by the Society of Automotive Engineers (SAE).

U.S. Department of Transportation, International Conference on Experimental Safety Vehicles: published biannually by the National Highway Traffic Safety Administration.

Stapp Car Crash Conference: published annually by the Society of Automotive Engineers (SAE).



Wood, W. L. "Ocean and Coastal Engineering"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



Bullwinkle, the world's tallest fixed offshore platform, is located approximately 150 miles southwest of New Orleans. With the platform and drilling rigs in place, the Bullwinkle structure stands at close to 1615 feet, which is 161 feet taller than the world's tallest building, the Sears Tower. Even more impressive than the sheer size of the structure is the fact that it was built to withstand a hurricane current of 4 ft/s, wave heights of 72 ft, and a wind velocity of 140 mph.

Twenty-eight piles were used to secure the jacket to the ocean floor. Each pile is 84 inches in diameter and up to 541 feet long. The piles were driven to penetrations of up to 437 feet with underwater hammers.

Since 1993 Bullwinkle has been producing more than 54 000 barrels of oil and 95 million cubic feet of natural gas per day. (Photo courtesy of Shell Oil Company.)

# XIII

## Ocean and Coastal Engineering

---

**William L. Wood**

*Purdue University*

### 84 Shallow Water and Deep Water Engineering *J. B. Herbich*

Wave Phenomena • Sediment Processes • Beach Profile • Longshore Sediment Transport • Coastal Structures • Navigational Channels • Marine Foundations • Oil Spills • Offshore Structures

IN THE VIEW OF TRADITIONAL ENGINEERING, ocean engineering may be seen as a relatively new field. In contrast, it may be argued that ocean engineering originated when the first human launched a craft in pursuit of what lay beyond the ocean's horizon. Regardless of view, the field of ocean engineering involves the application of literally all of the disciplines of science and engineering to the oceans. The physics of ocean waves on structures, the chemistry of desalination, and the biology and electrical engineering of marine bioacoustics are but a few of the rapidly emerging areas of ocean engineering. What is particularly exciting is that the fundamentals have been around for a long time, yet their application to the ocean continues to evolve new areas of engineering interest.

The concept of ocean engineering encompasses all of the geographic extent of the ocean, but usage has led to the connotation of deep open ocean engineering. "Coastal" is used to refer to shallow water engineering, but this usage suffers from ambiguity because shallow water is defined relative to properties of ocean waves.

Recognizing that ocean wave behavior in deep and shallow water is fundamental to understanding the principal constructs of ocean and coastal engineering, the chapter in this section presents the basic concepts of ocean surface wave behavior. Important wave-related engineering problems such as sediment transport, beach profile adjustment, and forces on seawalls are presented to illustrate useful coastal engineering applications. Forces on offshore structures are used to exemplify wave and current effects in deep water. This chapter provides a useful base of information from which to expand into the multitude of ocean and coastal engineering disciplines.

Herbich, J. B. "Shallow Water and Deep Water Engineering"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 84

## Shallow Water and Deep Water Engineering

---

### 84.1 Wave Phenomena

Airy (Low Amplitude) • Cnoidal (Shallow Water, Long Waves) • Stream Function • Stokian (Third Order)

### 84.2 Sediment Processes

### 84.3 Beach Profile

### 84.4 Longshore Sediment Transport

General Energy Flux Equation • Threshold of Sand Movement by Waves

### 84.5 Coastal Structures

Seawalls • Breakwaters

### 84.6 Navigational Channels

### 84.7 Marine Foundations

### 84.8 Oil Spills

### 84.9 Offshore Structures

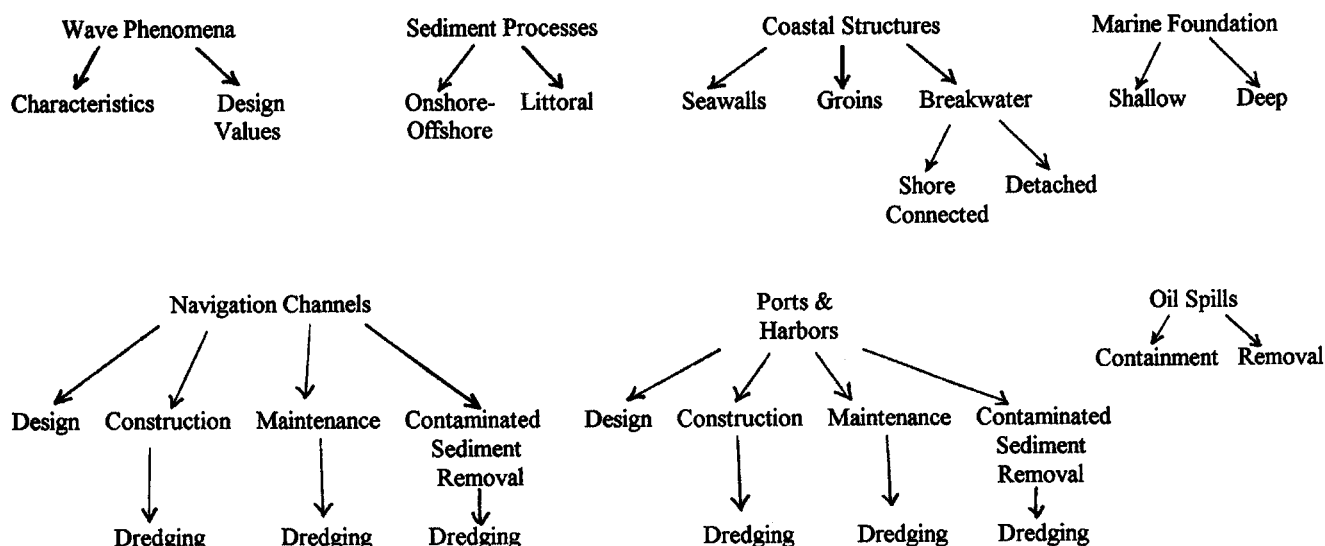
### John B. Herbich

*Texas A&M University Consulting & Research Services, Inc.*

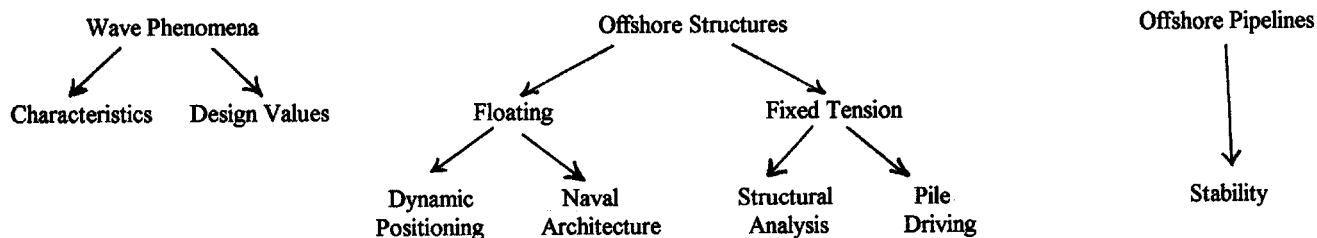
Ocean engineering is a relatively new branch of engineering. The need for this new specialty was recognized in the 1960s. Several universities, including Texas A&M, MIT, Florida Atlantic, the U.S. Coast Guard Academy, and the U.S. Naval Academy, established undergraduate degree programs in ocean engineering. Several universities have also developed programs at the graduate level specializing in ocean engineering.

Ocean and coastal engineering covers many topics, generally divided between shallow water (coastal engineering) and deep water (ocean engineering), shown in [Figs. 84.1](#) and [84.2](#).

**Figure 84.1** Coastal engineering (shallow water).



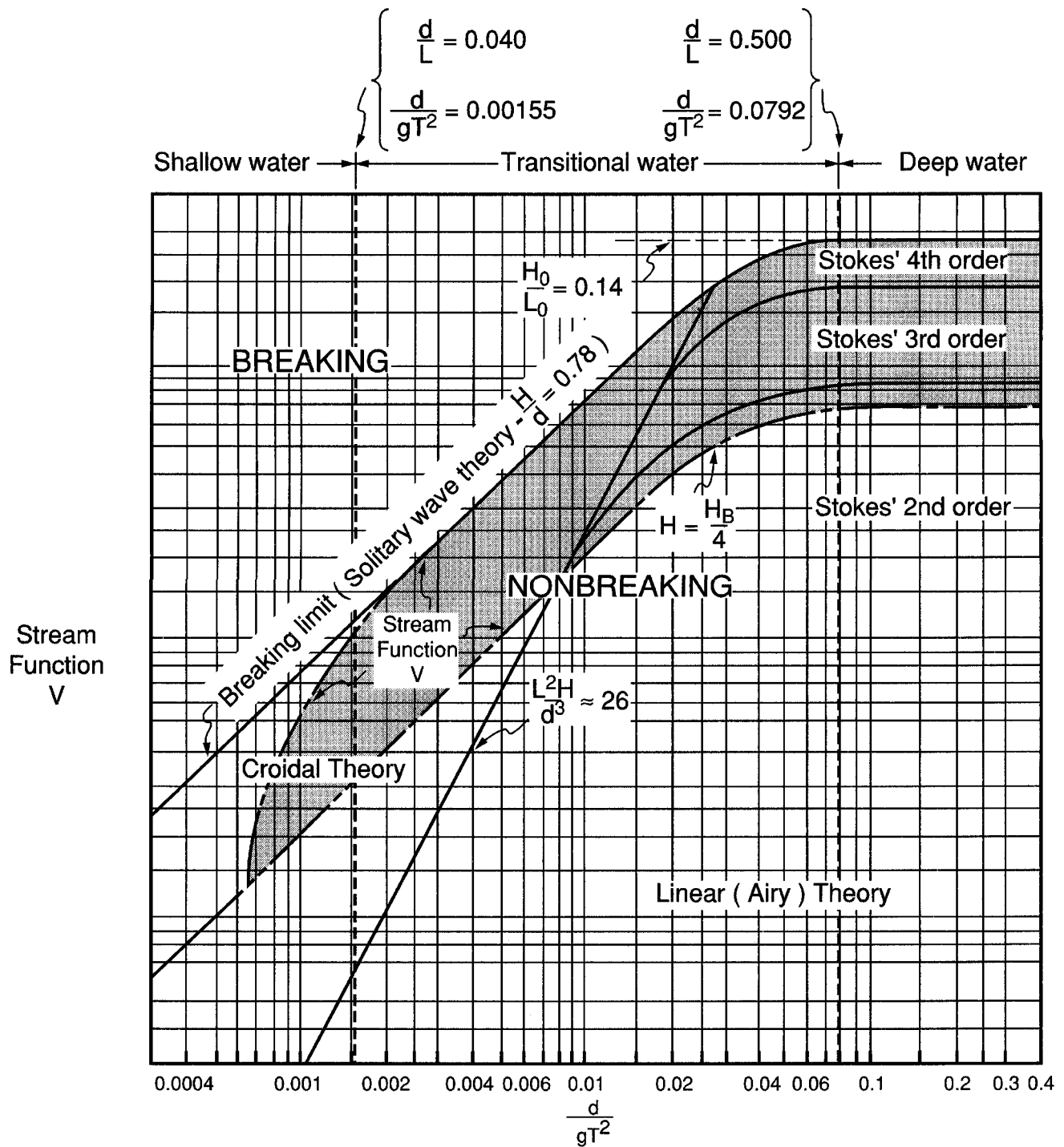
**Figure 84.2** Ocean engineering (deep water).



## 84.1 Wave Phenomena

Wave phenomena are of great importance in coastal and ocean engineering. Waves determine the composition and geometry of beaches. Since waves interact with human-made shore structures or offshore structures, safe design of these structures depends to a large extent on the selected wave characteristics. The structural stability criteria are often stated in terms of extreme environmental conditions (wave heights, periods, water levels, astronomical tides, storm surges, tsunamis, and winds). Waves in the ocean constantly change and are irregular in shape, particularly when under the influence of wind; such waves are called *seas*. When waves are no longer under the influence of wind and are out of the generating area, they are referred to as *swells*. Many wave theories have been developed, including the Airy, cnoidal, solitary, stream function, Stokian, and so forth. [Figure 84.3](#) describes the regions of validity for various wave theories. Cnoidal and stream function theories apply principally to shallow and transitional water, whereas Airy and Stokian theories apply to transitional and deep water (Airy applies to low amplitude waves).

**Figure 84.3** Regions of validity for various wave theories. (Source: Le Méhauté, 1969.)



## Airy (Low Amplitude)

Wavelength is given by the following equations.



$$\text{Shallow water} \quad L = T\sqrt{gh} = CT \quad (84.1)$$

$$\text{Transitional water} \quad L = \frac{gT^2}{2\pi} \tanh \left( \frac{2\pi h}{L} \right) \quad (84.2)$$

$$\text{Deep water} \quad L_o = \frac{gT^2}{2\pi} = C_o T \quad (84.3)$$

where

$T$  = wave period

$g$  = acceleration due to gravity

$h$  = water depth

$C$  = wave celerity

Subscript  $o$  denotes deep water conditions.

## Cnoidal (Shallow Water, Long Waves)

The theory originally developed by Boussinesq [1877], has been studied and presented in more usable form by several researchers. Wavelength is given by

$$L = \sqrt{\frac{16d^3}{3H}} kK(k) \quad (84.4)$$

and wave period by

$$T\sqrt{\frac{g}{h}} = \sqrt{\frac{16y_t}{3H}} \frac{h}{y_t} \left[ \frac{kK(k)}{1 + \frac{H}{y_t k^2} \left( \frac{1}{2} - \frac{E(k)}{K(k)} \right)} \right] \quad (84.5)$$

where

$y_t$  = distance from the bottom to the wave trough

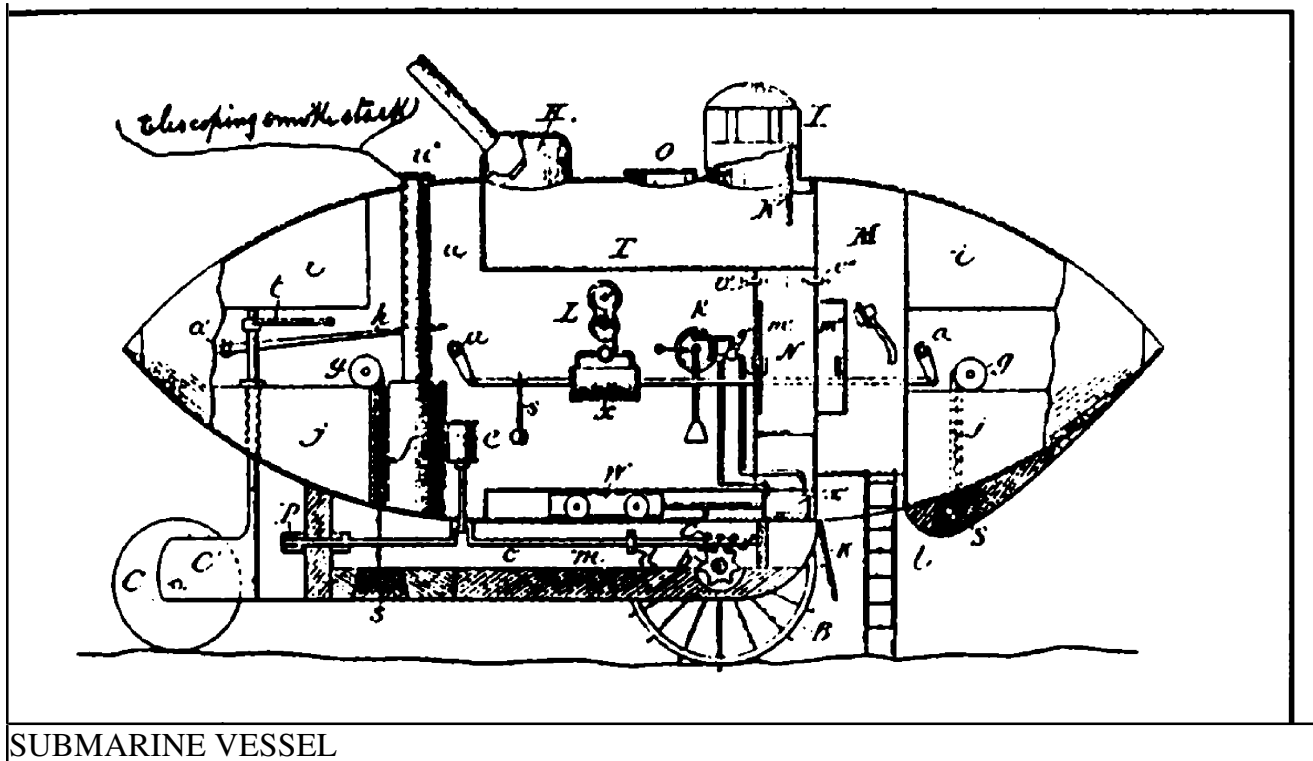
$k$  = modulus of the elliptic integrals

$K(k)$  = complete elliptic integral of the first kind

$E(k)$  = complete elliptic integral of the second kind

Cnoidal waves are periodic and of permanent form; thus  $L = CT$ .





SUBMARINE VESSEL

*Simon Lake*

*Patented April 20, 1897*

**#581,213**

Up to this time, many submarines had been described and some even built; including the Turtle, basically a water-tight barrel with a propeller, used once to plant a bomb under an enemy vessel during the American Revolution.

Lake's patent was the first to describe many of the features used in modern submarines such as compressed-air ballast for vertical submerging, an air lock for entering and leaving the vessel while it was submerged and an automatic control for maintaining the vessel at a selected depth while under power. The problem of locomotion while under water was addressed by reference to "chemical engines, storage batteries, and the like" which are means commonly used in today's small research submarines. Even described was the forerunner of the periscope for viewing the surface while the vessel was submerged. (© 1994, DewRay Products, Inc. Used with permission.)

## Stream Function

Stream function was developed by Dean [1977] and is of analytical form with the wavelength  $L$ ,

coefficients  $X(n)$ , and the value of stream function on the free surface  $\psi_\eta$  determined numerically. The expression for the stream function,  $\psi$ , for a wave system rendered stationary by a reference frame moving with the speed of the wave,  $C$ , is

$$\psi = \left( \frac{L}{T} - U \right) z + \sum_{n=1}^{NN} X(n) \sinh \left[ \frac{2\pi n}{L} (h + z) \right] \cos \left( \frac{2\pi n x}{L} \right) \quad (84.6)$$

with the coordinate  $z$  referenced to the mean water level;  $U$  is a uniform current.

Stream function (Table 84.1) provides values of wavelength  $L' = L/L_o$ ,  $\eta'_c = \eta_c/H$  (water surface elevation above mean water),  $\eta'_t = \eta_t/H$  (wave surface elevation below mean water),  $u'_c$  (horizontal dimensionless velocity at the crest),  $w'_m$  (maximum dimensionless vertical velocity),  $(F'_D)_m$  (maximum dimensionless drag force), and  $(F'_I)_m$  (maximum dimensionless inertia force).

**Table 84.1** Selected Summary of Tabulated Dimensionless Stream Function Quantities

Case	$h/L_o$	$H/L_o$	$L'$	$\eta'_c$	$\eta'_t$	$u'_c$	$w'_m$	$\theta(w'_m)^{\circ}$	$(F'_D)_m$	$(F'_I)_m$	$\theta(F'_I)_m^{\circ}$	$p'_{Dc}$ (Bottom)
1-A	0.002	0.00039	0.120	0.910	-0.090	49.68	13.31	10°	2574.0	815.6	10°	1.57
1-B	0.002	0.00078	0.128	0.938	-0.062	47.32	15.57	10°	2774.6	1027.0	10°	1.45
1-C	0.002	0.00117	0.137	0.951	-0.049	43.64	14.98	10°	2861.0	1043.5	10°	1.35
1-D	0.002	0.00156	0.146	0.959	-0.041	40.02	13.63	10°	2985.6	1001.7	10°	1.29
2-A	0.005	0.00097	0.187	0.857	-0.143	29.82	8.70	20°	907.0	327.1	20°	1.46
2-B	0.005	0.00195	0.199	0.904	-0.096	29.08	9.29	10°	1007.9	407.1	10°	1.36
2-C	0.005	0.00293	0.211	0.927	-0.073	26.71	9.85	10°	1060.7	465.7	10°	1.23
2-D	0.005	0.00388	0.223	0.944	-0.056	23.98	9.47	10°	1128.4	465.2	10°	1.11
3-A	0.01	0.00195	0.260	0.799	-0.201	19.83	6.22	30°	390.3	162.1	30°	1.34
3-B	0.01	0.00389	0.276	0.865	-0.135	19.87	7.34	20°	457.3	209.0	20°	1.28
3-C	0.01	0.00582	0.292	0.898	-0.102	18.47	6.98	20°	494.7	225.6	10°	1.16
3-D	0.01	0.00775	0.308	0.922	-0.078	16.46	6.22	10°	535.4	242.4	10°	1.04
4-A	0.02	0.00390	0.359	0.722	-0.278	12.82	4.50	30°	156.3	82.2	30°	1.18
4-B	0.02	0.00777	0.380	0.810	-0.190	13.35	5.38	30°	197.6	103.4	20°	1.16

4-C	0.02	0.01168	0.401	0.858	-0.142	12.58	5.29	20°	222.9	116.1	20°	1.06
4-D	0.02	0.01555	0.422	0.889	-0.111	11.29	4.99	20°	242.4	113.5	20°	0.97
5-A	0.05	0.00975	0.541	0.623	-0.377	7.20	3.44	50°	44.3	37.6	50°	0.93
5-B	0.05	0.01951	0.566	0.716	-0.284	7.66	3.69	50°	59.1	38.5	50°	0.94
5-C	0.05	0.02916	0.597	0.784	-0.216	7.41	3.63	30°	72.0	47.1	30°	0.88
5-D	0.05	0.03900	0.627	0.839	-0.161	6.47	3.16	30°	85.5	45.1	20°	0.76
6-A	0.10	0.0183	0.718	0.571	-0.429	4.88	3.16	75°	17.12	22.62	75°	0.73
6-B	0.10	0.0366	0.744	0.642	-0.358	5.09	3.07	50°	22.37	23.67	50°	0.73
6-C	0.10	0.0549	0.783	0.713	-0.287	5.00	2.98	50°	28.79	23.64	30°	0.70
6-D	0.10	0.0730	0.824	0.782	-0.218	4.43	2.44	50°	36.48	22.43	30°	0.62
7-A	0.20	0.0313	0.899	0.544	-0.456	3.63	3.05	75°	6.69	13.86	75°	0.46
7-B	0.20	0.0625	0.931	0.593	-0.407	3.64	2.93	75°	8.60	13.61	75°	0.47
7-C	0.20	0.0938	0.981	0.653	-0.347	3.54	2.49	50°	11.31	13.31	50°	0.47
7-D	0.20	0.1245	1.035	0.724	-0.276	3.16	2.14	50°	15.16	11.68	50°	0.44
8-A	0.50	0.0420	1.013	0.534	-0.466	3.11	2.99	75°	2.09	6.20	75°	0.090
8-B	0.50	0.0840	1.059	0.570	-0.430	3.01	2.85	75°	2.71	6.21	75°	0.101
8-C	0.50	0.1260	1.125	0.611	-0.389	2.86	2.62	75°	3.53	5.96	75°	0.116
8-D	0.50	0.1681	1.194	0.677	-0.323	2.57	1.94	50°	4.96	5.36	50°	0.120
9-A	1.00	0.0427	1.017	0.534	-0.466	3.09	2.99	75°	1.025	3.116	75°	0.004
9-B	1.00	0.0852	1.065	0.569	-0.431	2.98	2.85	75°	1.329	3.126	75°	0.005
9-C	1.00	0.1280	1.133	0.609	-0.391	2.83	2.62	75°	1.720	3.011	75°	0.008
9-D	1.00	0.1697	1.211	0.661	-0.339	2.60	1.99	75°	2.303	2.836	50°	0.009
10-A	2.00	0.0426	1.018	0.533	-0.467	3.09	2.99	75°	0.513	1.558	75°	-0.001
10-B	2.00	0.0852	1.065	0.569	-0.431	2.98	2.85	75°	0.664	1.563	75°	0.000
10-C	2.00	0.1275	1.134	0.608	-0.392	2.83	2.63	75°	0.860	1.510	75°	-0.001
10-D	2.00	0.1704	1.222	0.657	-0.343	2.62	2.04	75°	1.137	1.479	50°	0.000

\*Notes: (1) Except where obvious or noted otherwise, dimensionless quantities are presented for mean water elevation. (2) The maximum dimensionless drag and inertial forces apply for a piling extending through the entire water column. (3) Subscripts *m*, *c*, and *t* denote "maximum," "crest," and "trough," respectively.

Source: Dean, R. G. 1991. Beach profiles. In *Handbook of Coastal and Ocean Engineering, Volume 2*, ed. J. B. Herbich. Gulf, Houston, TX. Copyright ©1990 by Gulf Publishing Company, Houston, TX. Used with permission. All rights reserved.

## Stokian (Third Order)

Wavelength is given by

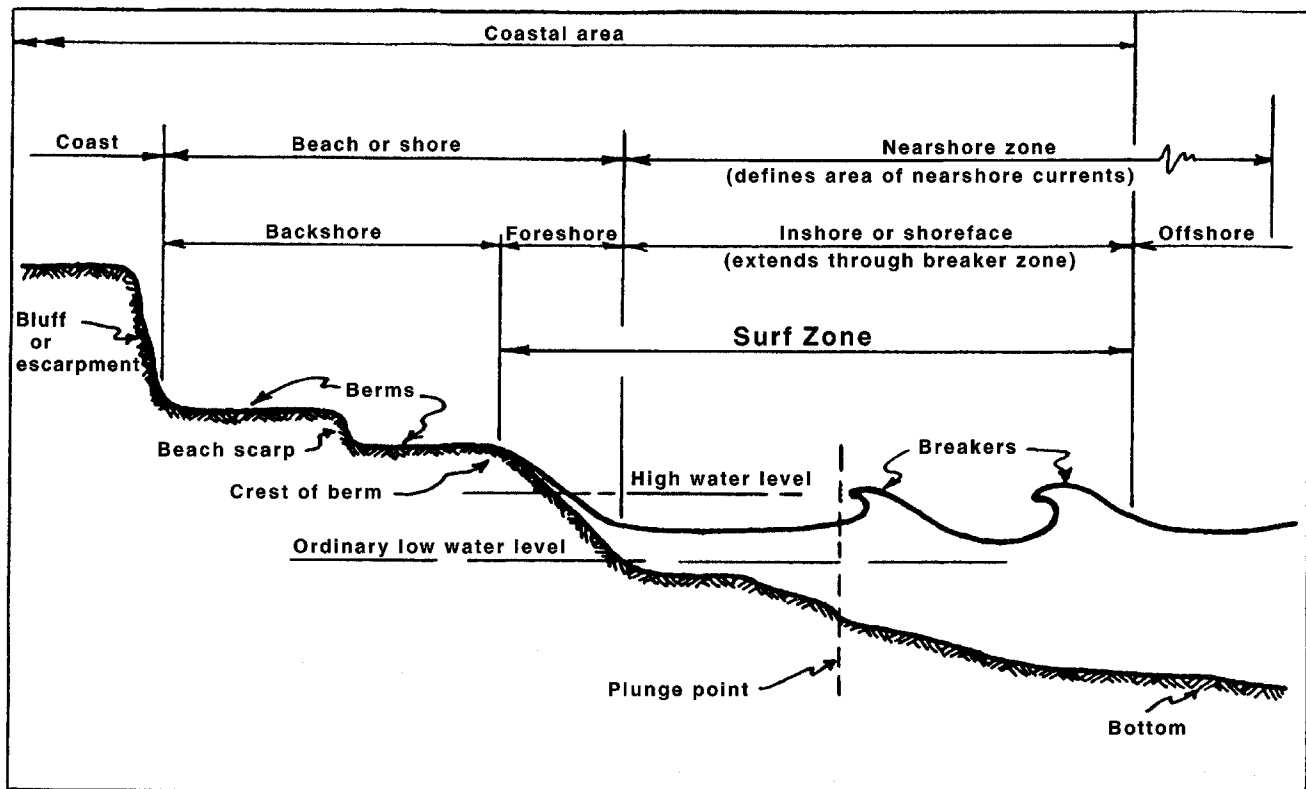
$$L = \frac{gT^2}{2\pi} \tanh\left(\frac{2\pi h}{L}\right) \left\{ 1 + \left(\frac{\pi H}{L}\right)^2 \left[ \frac{5 + 2 \cosh(4\pi h/L) + 2 \cosh^2(4\pi h/L)}{8 \sinh^4(2\pi d/L)} \right] \right\} \quad (84.7)$$

## 84.2 Sediment Processes

Along the coasts the ocean meets land. Waves, currents, tsunamis, and storms have been shaping the beaches for many thousands of years. Beaches form the first defense against the waves and are constantly moving on, off, and along the shore (littoral drift). [Figure 84.4](#) provides a definition for terms describing a typical beach profile. The shoreline behavior is very complex and difficult to understand; it cannot be expressed by equations because many of the processes are site-specific. Researchers have, however, developed equations that should be summarized. There are two basic sediment movements:

1. On- and offshore
2. Parallel to the shore and at an angle to the shore.

**Figure 84.4** Visual definition of terms describing a typical beach profile. (Source: Department of the Army. 1984. *Shore Protection Manual*, vols. I and II. Department of the Army, Corps of Engineers, Coastal Engineering Research Center, Waterways Experiment Station, Vicksburg, MS.)



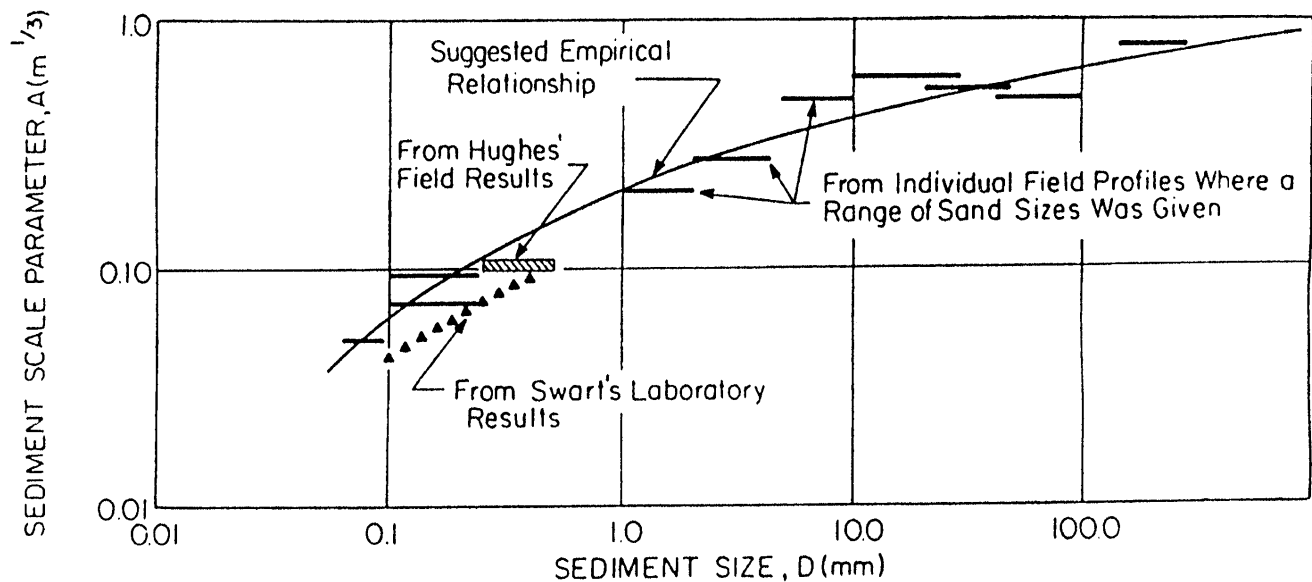
## 84.3 Beach Profile

Information on beach profiles is essential in designing structural modifications (such as seawalls, revetments, and breakwaters, both connected and detached), pipeline crossings, and beach replenishment. Bruun [1954] indicated that many beach profiles (Fig. 84.5) can be represented by

$$h(x) = Ax^{2/3}$$

where  $h$  is the water depth at a distance  $x$  offshore, and  $A$  is a dimensional scale parameter.

**Figure 84.5** Beach profile scale factor,  $A$ , versus sediment diameter,  $D$ , in relationship  $h = Ax^{2/3}$ . (Source: Dean, R. G. 1991. Beach profiles. In *Handbook of Coastal and Ocean Engineering*, Volume 2, ed. J. B. Herbich. Gulf, Houston, TX. Copyright 1990 by Gulf Publishing Company, Houston, TX. Used with permission. All rights reserved.)



Dean [1977] showed that  $H_b/wT$  is an important parameter distinguishing **barred** profiles from nonbarred profiles (where  $H_b$  is breaking wave height,  $w$  is fall velocity of sediment in water, and  $T$  is wave period). This parameter is consistent with the following beach profiles in nature:

$$\begin{array}{lcl}
\text{Milder slope profiles} & \left\{ \begin{array}{l} \text{High waves} \\ \text{Short periods} \\ \text{Small sediment diameter} \end{array} \right. & \\
\text{Steeper profiles} & \left\{ \begin{array}{l} \text{Low waves} \\ \text{Long periods} \\ \text{Large sediment diameter} \end{array} \right. & 
\end{array}$$

$$\text{When } \frac{H_b}{wT} > 0.85, \text{ one can expect } \mathbf{\bar{b}ar} \text{ formation.} \quad (84.8a)$$

$$\text{When } \frac{H_b}{wT} < 0.85, \text{ a monotonic profile can be expected.} \quad (84.8b)$$

Later, on the basis of large laboratory data, Kriebel *et al.* [1986] found the value of 2.3 rather than 0.85 in Eqs. (84.8a) and (84.8b).

## 84.4 Longshore Sediment Transport

---

The longshore transport ( $Q$ ) is the volumetric rate of sand movement parallel to the shoreline. Much longshore transport occurs in the surf zone and is caused by the approach of waves at an angle to the shoreline.

Longshore transport rate ( $Q$ , given in unit volume per second) is assumed to depend upon the longshore component of wave energy flux,  $P_{ls}$  (Department of the Army, 1984):

$$Q = \frac{K}{(\rho_s - \rho)g\alpha} P_{ls} \quad (84.9)$$

where

$K$  = dimensionless empirical coefficient (based on field measurements) = 0.39

$\rho_s$  = density of sand

$\rho$  = density of water

$g$  = acceleration due to gravity

$\alpha$  = ratio of the volume of solids to total volume, accounting for sand porosity = 0.6

## General Energy Flux Equation

The energy flux per unit length of wave crest or, equivalently, the rate at which wave energy is transmitted across a plane of unit width perpendicular to the direction of wave advance,

is

$$P = EC_g \quad (84.10)$$

where  $E$  is wave energy density and  $C_g$  is wave group speed. The wave energy density is calculated by

$$E = \frac{\rho g H^2}{8} \quad (84.11)$$

where  $\rho$  is mass density of water,  $g$  is acceleration of gravity, and  $H$  is wave height.

If the wave crests make an angle  $\alpha$  with the shoreline, the energy flux in the direction of wave advance per unit length of beach is

$$P \cos \alpha = \frac{\rho g H^2}{8} C_g \cos \alpha \quad (84.12)$$

The longshore component of wave energy flux is

$$P_l = P \cos \alpha \sin \alpha = \frac{\rho g H^2}{8} C_g \cos \alpha \sin \alpha \quad (84.13)$$

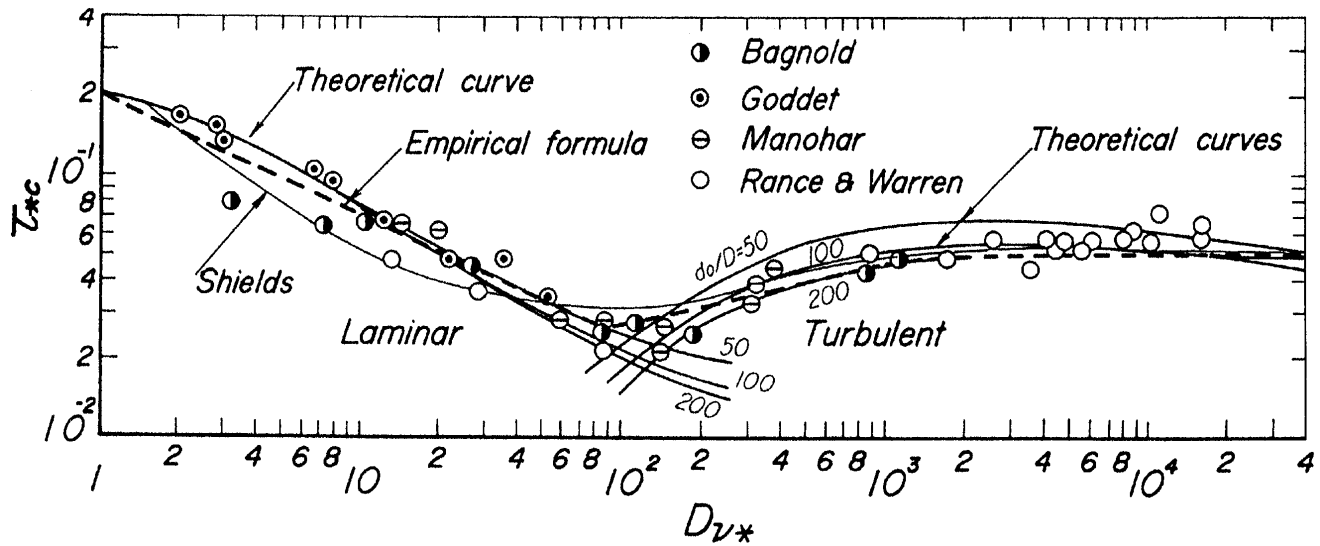
or,

$$P_l = \frac{\rho g}{16} H^2 C_g \sin 2\alpha \quad (84.14)$$

## Threshold of Sand Movement by Waves

The threshold of sand movement by wave action has been investigated by a number of researchers [e.g., [Tsuchiya, 1991](#)]. [Figure 84.6](#) shows the modified Shields diagram, where  $\tau_{*c} = 1/\varepsilon \psi_i(D_{\nu*})$ , and  $\psi_i(D_{\nu*})$  is a function of sediment-fluid number only, plotted as a function of  $D_{\nu*}$ .

**Figure 84.6** Threshold of sand movement by waves with Shields, Sleath, and Tsuchiya empirical curves, as well as the theoretical curve. (Source: Tsuchiya, Y. 1991. Threshold of sand movement. In *Handbook of Coastal and Ocean Engineering, Volume 2*, ed. J. B. Herbich. Gulf, Houston, TX. Copyright 1990 by Gulf Publishing Company, Houston, TX. Used with permission. All rights reserved.)



The empirical formula shown by dashed lines is as follows:

$$\begin{aligned}
 \tau_{*c} &= 0.20 & \text{for } D_{v*} \leq 1 \\
 &= 0.20 D_{v*}^{-2/3} & \text{for } 1 \leq D_{v*} \leq 20 \\
 &= 0.010 D_{v*}^{1/3} & \text{for } 20 \leq D_{v*} \leq 125 \\
 &= 0.050 & \text{for } 125 \leq D_{v*}
 \end{aligned} \tag{84.15}$$

## 84.5 Coastal Structures

Wave forces act on coastal and offshore structures; the forces may be classified as due to nonbreaking, breaking, and broken waves. Fixed coastal structures include (a) wall-type structures such as **seawalls**, bulkheads, revetments, and certain types of breakwaters, (b) pile-supported structures such as piers and offshore platforms, and (c) rubble structures such as breakwaters, **groins**, and revetments.

### Seawalls

Forces due to nonbreaking waves may be calculated using Sainflou or Miche-Rundgren formulas. Employing the Miche-Rundgren formula, the pressure distribution is

$$p_1 = \left( \frac{1 + \chi}{2} \right) \frac{\gamma H_i}{\cosh(2\pi h/L)} \tag{84.16}$$

$\chi$  = wave reflection coefficient

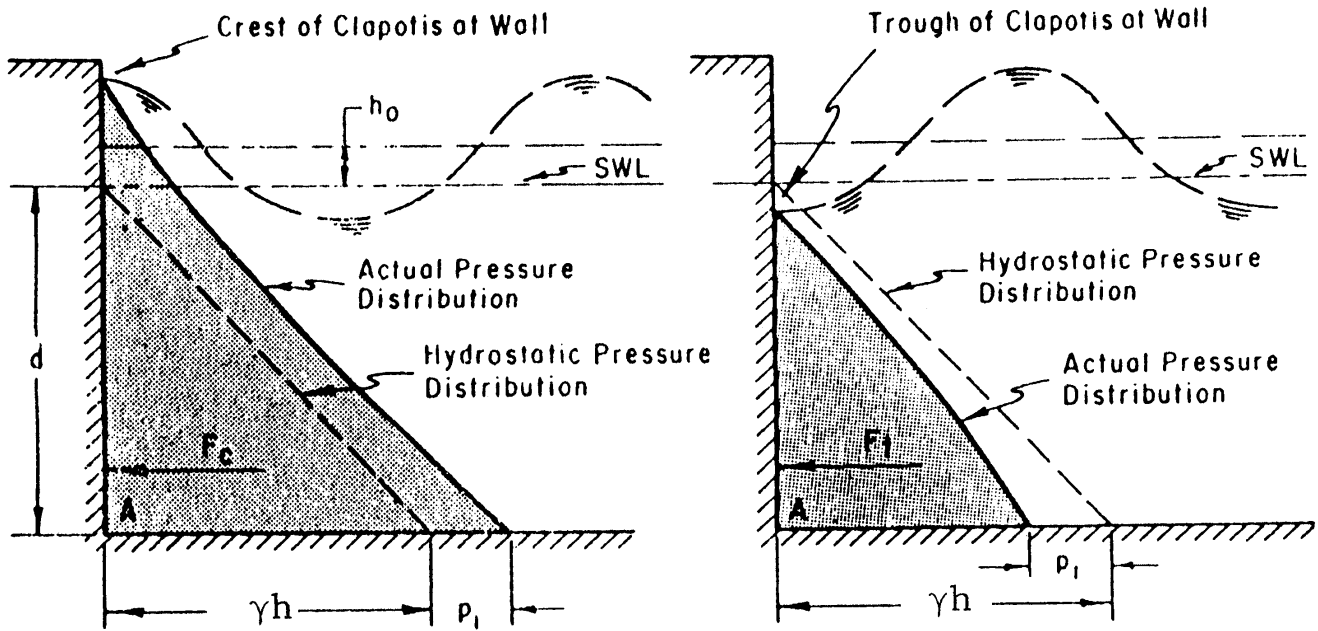
$\gamma$  = unit weight of water



$H_i$  = incident wave height  
 $h$  = water depth  
 $L$  = wavelength

Figure 84.7 shows the pressure distribution at a vertical wall at the crest and trough of a clapotis.

**Figure 84.7** Pressure distributions for nonbreaking waves. (Source: Department of the Army. 1984. *Shore Protection Manual*, vols. I and II. Department of the Army, Corps of Engineers, Coastal Engineering Research Center, Waterways Experiment Station, Vicksburg, MS.)



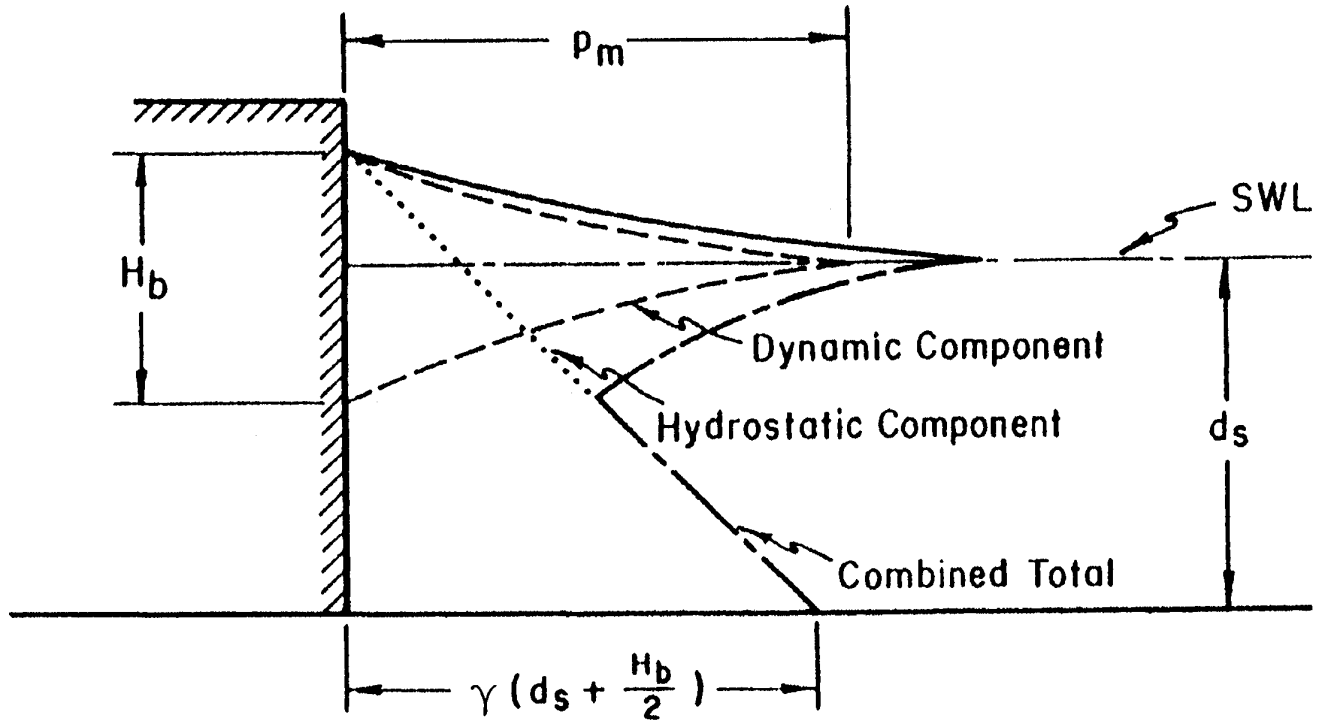
Forces due to breaking waves may be estimated by Minikin and Goda methods. The Minikin method described by the Department of the Army [1984] estimates the maximum pressure (assumed to act on the SWL) to be:

$$p_m = 101\gamma \frac{H_b}{L_D} \frac{d_s}{D} (D + d_s) \quad (84.17)$$

where  $p_m$  is the maximum dynamic pressure,  $H_b$  is the breaker height,  $d_s$  is the depth at the toe of the wall,  $D$  is the depth one wavelength in front of the wall, and  $L_D$  is the wavelength in water depth  $D$ . The distribution of dynamic pressure is shown in Fig. 84.8. The pressure decreases parabolically from  $p_m$  at the WL to zero at a distance of  $H_b/2$  above and below the SWL. The force represented by the area under the dynamic pressure distribution is

$$R_m = \frac{p_m H_b}{3} \quad (84.18)$$

**Figure 84.8** Minikin wave pressure diagram. (Source: Department of the Army. 1984. *Shore Protection Manual*, vols. I and II. Department of the Army, Corps of Engineers, Coastal Engineering Research Center, Waterways Experiment Station, Vicksburg, MS.)

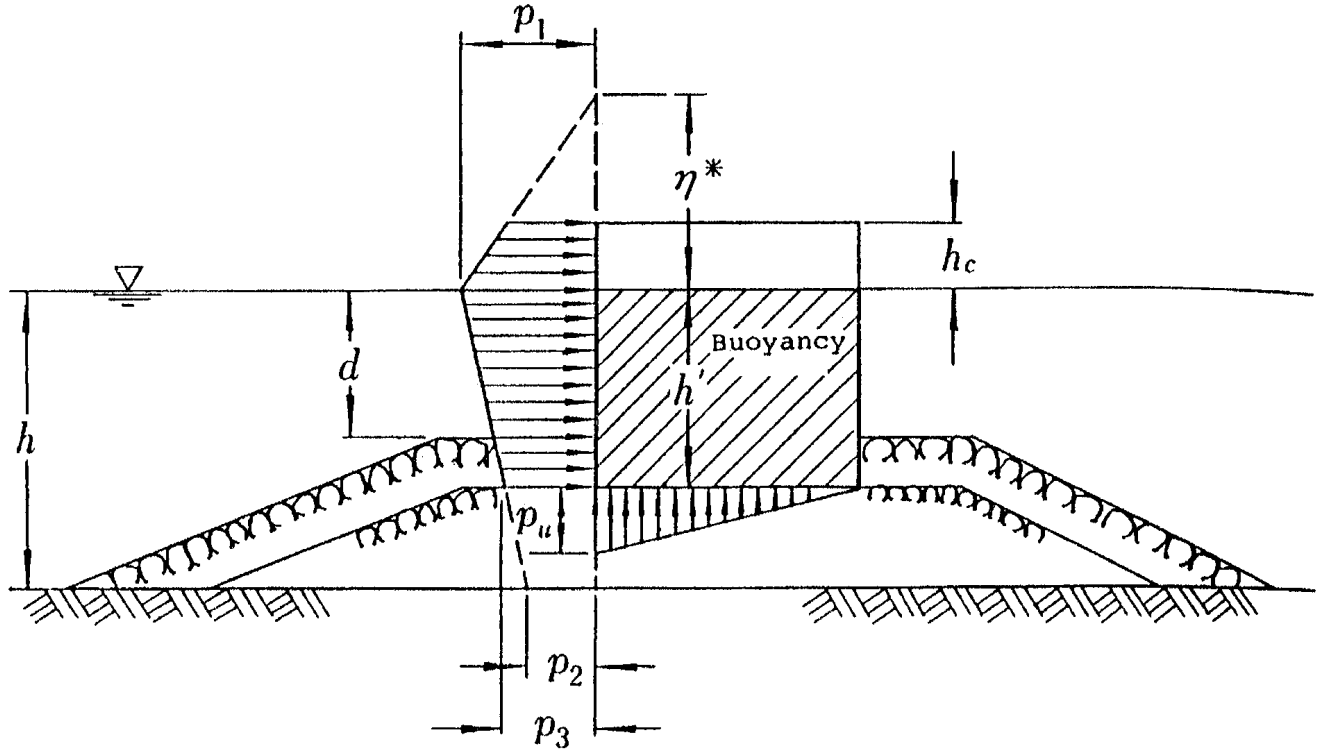


Goda's method [1985] assumes a trapezoidal pressure distribution (Fig. 84.9). The pressure extends to a point measured from SWL at a distance given by  $\eta^*$ :

$$\eta^* = 0.75(1 + \cos \beta) H_{\max} \quad (84.19)$$

in which  $\beta$  denotes the angle between the direction of wave approach and a line normal to the breakwater.

**Figure 84.9** Distribution of wave pressure on an upright section of a vertical breakwater. (Source: Goda, Y. 1990. Random wave interaction with structures. In *Handbook of Coastal and Ocean Engineering, Volume 1*, ed. J. B. Herbich. Gulf, Houston, TX. Copyright 1990 by Gulf Publishing Company, Houston, TX. Used with permission. All rights reserved.)



The wave pressure at the wall is given by

$$p_1 = \frac{1}{2}(1 + \cos \beta)(\alpha_1 + \alpha_2 \cos^2 \beta)\gamma H_{\max} \quad (84.20)$$

$$p_2 = \frac{p_1}{\cosh(2\pi h/L)} \quad (84.21)$$

$$P_3 = \alpha_3 p_1 \quad (84.22)$$

in which

$$\alpha_1 = 0.6 + 0.5 \left[ \frac{4\pi h/L}{\sinh(4\pi h/L)} \right]^2 \quad (84.23)$$

$$\alpha_2 = \min \left[ \frac{h_b - d}{3h_b} \left( \frac{H_{\max}}{d} \right)^2, \frac{2d}{H_{\max}} \right] \quad (84.24)$$

$$\alpha_3 = 1 - \frac{h'}{h} \left[ 1 - \frac{1}{\cosh(2\pi h/L)} \right] \quad (84.25)$$

## Breakwaters

Rubble-mound breakwaters are the oldest form of breakwaters, dating back to Roman times. The rubble mound is protected by larger rocks or artificial concrete units. This protective layer is usually referred to as **armor** or cover layer.

$$W = \frac{\gamma_r H^3}{K_D (S_r - 1)^3 \cot \theta} \quad (84.26)$$

where

$W$  = weight in newtons or pounds of an individual armor unit in the primary cover layer

$\gamma_r$  = unit weight (saturated surface dry) of armor unit in  $\text{N/m}^3$  or  $\text{lb/ft}^3$

$S_r$  = specific gravity of armor unit, relative to the water at the structure ( $S_r = w_r/w_\omega$ )

$\gamma_\omega$  = unit weight of water: freshwater =  $9800 \text{ N/m}^3$  ( $62.4 \text{ lb/ft}^3$ );

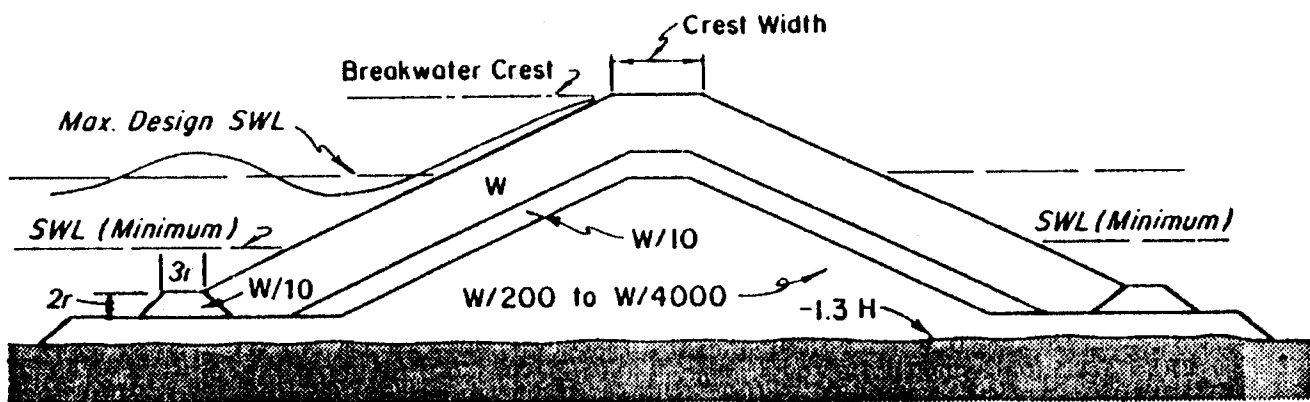
seawater =  $10\,047 \text{ N/m}^3$  ( $64.0 \text{ lb/ft}^3$ )

$\theta$  = angle of structure slope measured from horizontal in degrees

$K_D$  = stability coefficient that varies primarily with the shape of the armor units, roughness of the armor unit surface, sharpness of edges, and degree of interlocking obtained in placement

Figure 84.10 presents the recommended three-layer section of a rubble-mound breakwater. Note that underlayer units are given in terms of  $W$ , the weight of armor units.

**Figure 84.10** Rubble-mound section for wave exposure on both sides with moderate overtopping conditions. (Source: Department of the Army. 1984. *Shore Protection Manual*, vols. I and II. Department of the Army, Corps of Engineers, Coastal Engineering Research Center, Waterways Experiment Station, Vicksburg, MS.)



**Recommended Three-layer Section**

Automated coastal engineering system (ACES) describes the computer programs available for the design of breakwaters using Hudson and related equations.

Van der Meer [1987] developed stability formulas for plunging (breaking) waves and for surging (nonbreaking) waves. For plunging waves,

$$H_s / \Delta D_{n50} * \sqrt{\xi_z} = 6.2 P^{0.18} (S / \sqrt{N^{0.2}}) \quad (84.27)$$

For surging waves,

$$H_s / \Delta D_{n50} = 1.0 P^{-0.13} (S / \sqrt{N^{0.2}}) \sqrt{\cot \alpha} \xi_z^p \quad (84.28)$$

$H_s$  = **significant wave height** at the toe of the structure

$\xi_z$  = surf similarity parameter,  $\xi_z = \frac{\tan \alpha}{\sqrt{2\pi H_s / g T_z^2}}$

$T_z$  = zero up-crossing wave period

$\alpha$  = slope angle

$\Delta$  = relative mass density of the stone,  $\Delta = \rho_a / (\rho - 1)$

$\rho_a$  = mass density of the stone

$\rho$  = mass density of water

$D_{n50}$  = nominal diameter of the stone,  $D_{n50} = (W_{50} / \rho_a)^{1/3}$

$W_{50}$  = 50% value (median) of the mass distribution curve

$P$  = permeability coefficient of the structure

$S$  = damage level,  $S = A / D_{n50}^2$

$A$  = erosion area in a cross section

$N$  = number of waves (storm duration)

Influence of breakwater slope angle is depicted in Fig. 84.11.

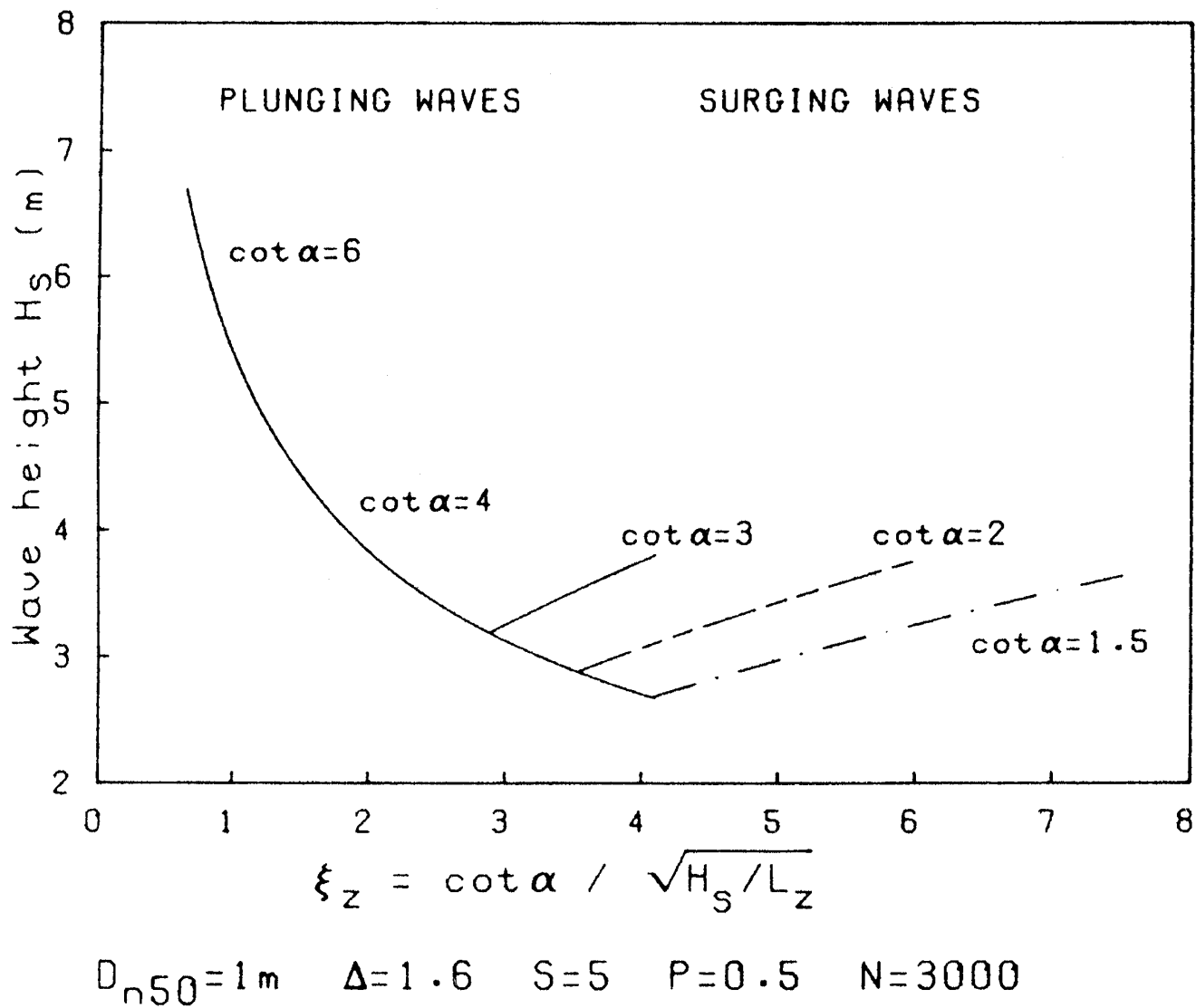
## 84.6 Navigational Channels

The development of very large commercial craft (VLCC) and ultralarge commercial craft (ULCC) forced many government planners and port managers to evaluate existing channels. Navigational channels allow large vessels to reach harbors. Of paramount design consideration is the safety of vessels in a channel, particularly when passing [Herbich, 1992].

Vessel behavior in channels is a function of bottom suction, bank suction, interference of passing ships, waves, winds, and currents. Most major maritime countries have criteria regarding the depth and width of channels. The international commission ICORELS (sponsored by the Permanent International Association of Navigation Congresses\xd1 PIANC) recommends that general criteria for gross underkeel clearances can be given for drawing up preliminary plans:

- *Open sea area.* When exposed to strong and long stern or quarter swells where speed may be

**Figure 84.11** Influence of slope angle. (Source: Van der Meer, J. W. 1990. Rubble mounds—Recent modifications. In *Handbook of Coastal and Ocean Engineering, Volume 1*, ed. J. B. Herbich. Gulf, Houston, TX. Copyright 1990 by Gulf Publishing Company, Houston, TX. Used with permission. All rights reserved.)



high, the gross underkeel clearance should be about 20% of the maximum draft of the large ships to be received.

- *Waiting area.* When exposed to strong or long swells, the gross underkeel clearance should be about 15% of the draft.
- *Channel.* For sections exposed to long swells, the gross underkeel clearance should be about 15% of the draft.

The *Engineering Manual* [U.S. Army Corps of Engineers, 1983] provides guidance for the layout and design of deep-draft navigation channels. Table 84.2 provides the general criteria for channel widths.

**Table 84.2** General Criteria for Channel Widths

Location	Minimum Channel Width in Percent of Beam			
	Vessel Controllability			Channels with Yawing Forces
	Very Good	Good	Poor	
Maneuvering lane, straight channel	160	180	200	Judgment*
Bend, 26° turn	325	370	415	Judgment*
Bend, 40° turn	385	440	490	Judgment*
Ship clearance	80	80	80	100 but not less than 100 ft
Bank clearance	60	60 plus	60 plus	150

\*Judgment will have to be based on local conditions at each project.

Source: U.S. Army Corps of Engineers. 1983. *Engineering Manual: Hydraulic Design of Deep Draft Navigation Projects*. EM 1110-2-1613. U.S. Army Corps of Engineers, Washington, DC.

## 84.7 Marine Foundations

Design of marine foundations is an integral part of any design of marine structures. The design criteria require a thorough understanding of marine geology; geotechnical properties of sediments at a given location; and wind, wave, currents, tides, and surges during maximum storm conditions. In the arctic areas information on fast ice and pack ice is required for the design of offshore structures (on artificial islands) and offshore pipelines.

A number of soil engineering parameters are required, as shown in Table 84.3. Many of the properties may be obtained employing standard geotechnical methods. Geotechnical surveys and mapping of seabed characteristics have reached a high degree of sophistication. High-resolution geophysical surveys determine water depth, seafloor imagery, and vertical profiles. Bottom-mapping systems include multibeam bathymetry, sea beam, side-scan sonars, and subbottom profilers (including shallow, medium, and deep penetration types).

**Table 84.3** Soil Engineering Parameters Normally Required for Categories of Geotechnical Engineering Applications

Application	Soil Classification	Grain Size	Atterberg Limits	Strength Properties				Common Properties			Subbottom Depth of Survey
				Clay		Sand		Clay		Sand	
				$S_u, S_r$	$\bar{c}, \phi'$	$\phi'$	$\phi$ or $S_u$	$C_v, k$	$C_c$	$C_c$	
Shallow foundation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1.5 to 2 × foundation width
Deadweight anchors	Yes	No	No	Yes	Yes	Yes	No	No	No	No	1.5 to 2 × anchor width
Deep pile foundations	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	1 to 1.5 × pile group width, below individual pile tips
Pile anchors	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	To depth of pile anchor
Direct-embedment anchors	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	No	To expected penetration of anchor, maximum 33 to 50 ft clay; 13 to 33 ft sand
Drag anchors	Yes	Yes	No	Yes	No	No	No	No	No	No	33 to 50 ft clay; 10 to 16½ ft sand for large anchors
Penetration	Yes	Yes	No	Yes	No	Yes	Yes	No	No	No	33 to 50 ft clay; 13 to 33 ft sand
Breakout	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	1 × object width plus embedment depth
Scour	Yes	Yes	No	Yes	No	No	No	No	No	No	3.3 to 16½ ft; related to object size and water motion
Slope stability	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	33 to 100 ft; more on rare occasions

Symbols:

$S_u$  = undrained shear strength

$S_r$  = sensitivity

$\bar{c}$  = drained cohesion intercept

$\phi'$  = drained friction angle

$\phi$  = undrained friction angle for sands rapidly sheared

$C_v$  = coefficient of consolidation

$k$  = permeability

$C_c$  = compression index

Source: Marine Board, National Research Council. 1989. *Our Seabed Frontier—Challenges and Choices*. National Academy Press, Washington, DC.

The geotechnical investigation is designed to include sediment stratigraphy; sediment types; and sediment properties, including density, strength, and deformational characteristics. Deployment systems employed for sampling in situ include self-contained units, drilling rigs, and submersibles. (Figure 84.12 shows the deployment systems.)

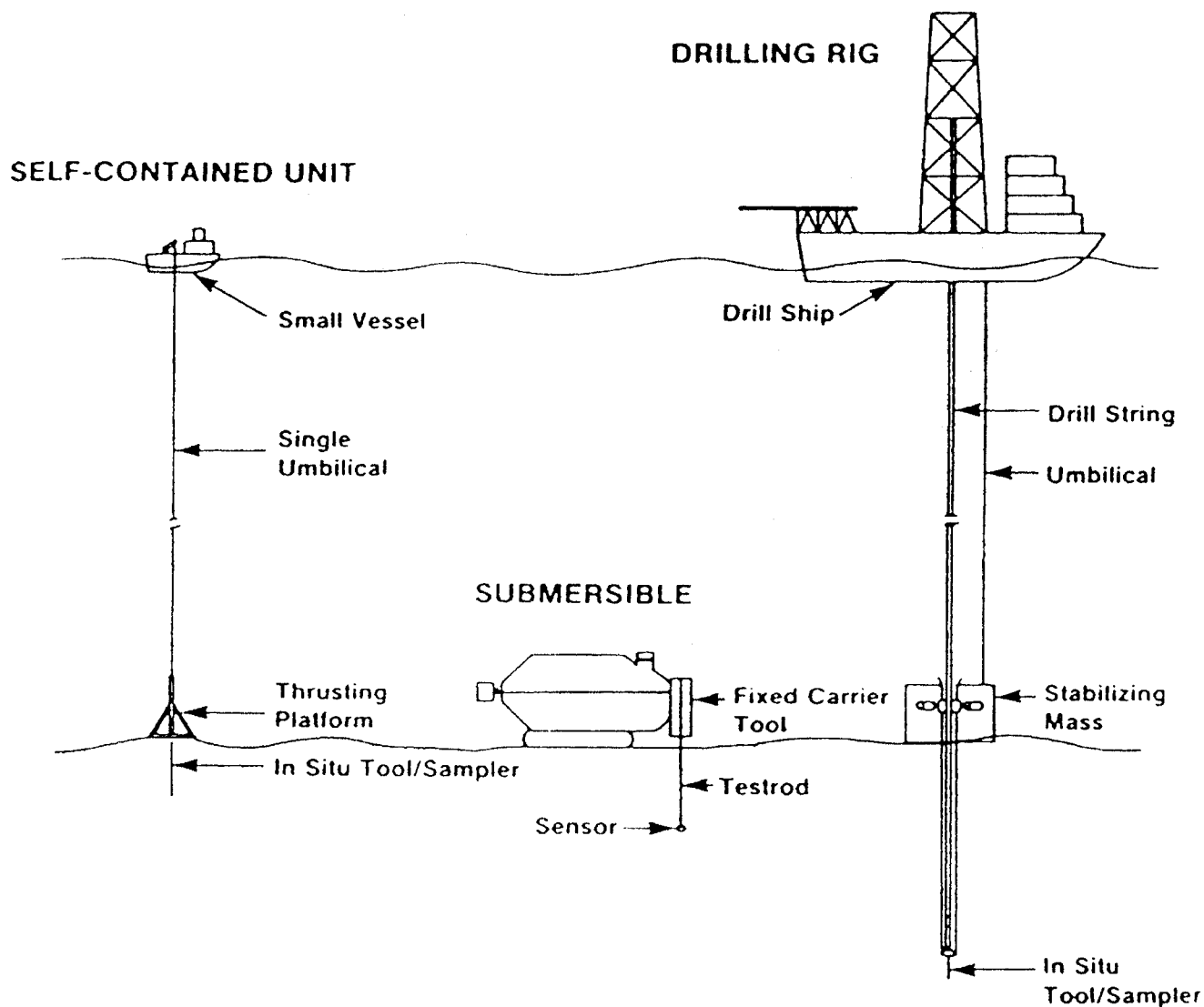
There are many in situ testing devices; these include the vane shear test, cone penetrometer test, pressure meter, shear vane velocity tools, temperature probes, natural gamma logger, and so forth [Young, 1991].

## 84.8 Oil Spills

The best method of controlling oil pollution is to prevent oil spills in the first place. This may include such techniques as rapid removal of oil from stricken tankers, continuous monitoring of oil wells, killing wild wells at sea, and containing oil spills under the water surface. Spilled oil, being lighter than water, floats on the water surface and spreads laterally. As oil is spilled, several regimes are generally assumed: gravity-inertial, gravity-viscous, and surface tension. In the early stage, generally less than one hour, the gravity-inertial regime, or inertial spread, dominates and is described by



**Figure 84.12** Deployment systems used for sampling, in situ, and experimental testings. (Source: Marine Board, National Research Council. 1989. *Our Seabed Frontier*<sup>3/4</sup>*Challenges and Choices*. National Academy Press, Washington, DC.)



$$R = k_4(\Delta g L t^2)^{1/4} \quad (84.29)$$

$R$  = radius of the oil slick

$k_4$  = nondimensional coefficient experimentally determined to be 1.14

$\Delta$  = the ratio of the absolute difference between the densities of sea water and the oil to that of seawater

$g$  = force of gravity

$L$  = original volume of oil spilled

$t$  = time

When the oil film thickness becomes equal to the viscous layer in the water, a transition occurs from the gravity-inertial regime to the gravity-viscous regime. This viscous spreading is described by

$$\text{Radius of oil slick} = R = k_5 \left( \frac{\Delta g L^2 t^{3/2}}{\nu^{1/2}} \right)^{1/6} \quad (84.30)$$

where  $k_5$  is the nondimensional coefficient determined to be about 1.45,  $\nu$  is the kinematic viscosity of water,  $\Delta$  is the ratio of the difference between density of seawater and oil,  $L$  is the original volume of spilled oil, and  $t$  is the time.

The last phase, the surface tension regime, occurs when the oil film thickness drops below a critical level, which is a function of the net surface tension, the mass densities of the oil and the water, and the force of gravity. The surface tension spread is described by

$$R = k_6 \left( \frac{\sigma^2 t^3}{\rho^2 \nu} \right)^{1/4} \quad (84.31)$$

$k_6$  = 2.30, experimentally determined

$\sigma$  = surface tension

$\rho$  = density of water

For large spills, on the order of 10 000 tons, inertial and viscous spreading will dominate for about the first week, with the surface tension spread controlling thereafter.

Although the exact mechanisms that cause the termination of spreading are unknown, the terminal areas of several oil slicks have been observed and used to determine an analytical relationship for the maximum area of a given oil spill based on the properties of the oil. This is described by

$$A_T = K_a \left( \frac{\sigma^2 V^6}{\rho^2 \nu D^3 s^6} \right)^{1/8} \quad (84.32)$$

$K_a$  = undetermined constant or order unit

$V$  = volume of oil that can be dissolved in this layer

$D$  = diffusivity

$s$  = solubility of the significant oil fractions in the water

In addition, the area covered by the oil slick is not allowed to exceed  $A_T$ ; therefore, spreading is terminated at the time

$$t = \left( \frac{V\rho}{s\sigma} \right)^{1/2} \left( \frac{\nu}{D} \right)^{1/4} \left( \frac{K_a}{\pi k_6^2} \right)^{2/3} \quad (84.33)$$

Oil may be set up by wind and current against a barrier; any containment device must take the setup estimates into account. There are a number of containment devices (barriers) that prevent oil from spreading. Most mechanical-type oil containment barriers fail in wave heights greater than 2 ft, when the wave steepness ratio is greater than 0.08, and in currents normal to the barrier greater than about 0.7 knots.

Oil may also be removed from the water surface by skimming devices. Most mechanical skimming devices have only been able to work in waves less than 2 to 3 ft in height, in moderate currents.

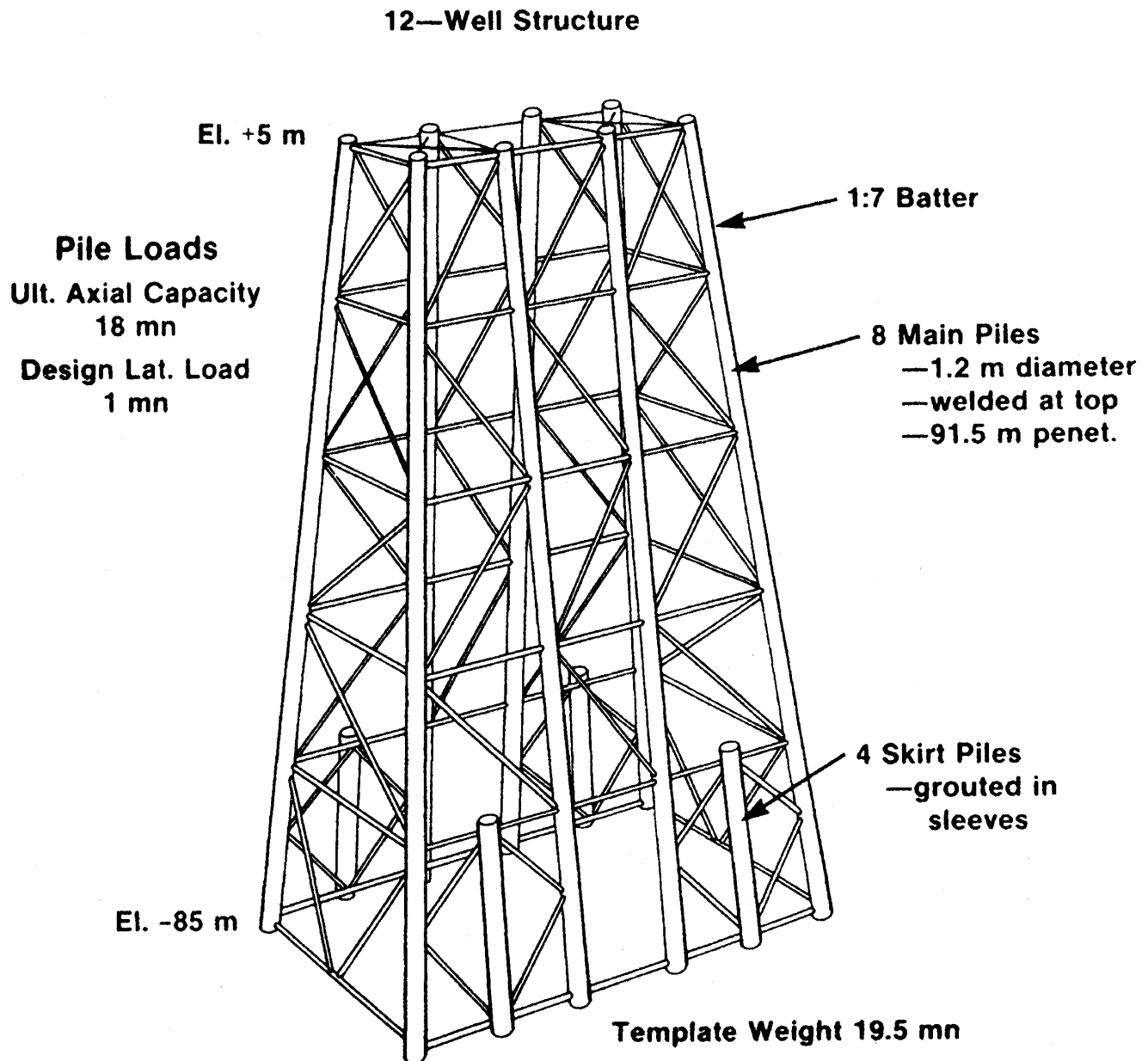
## 84.9 Offshore Structures

---

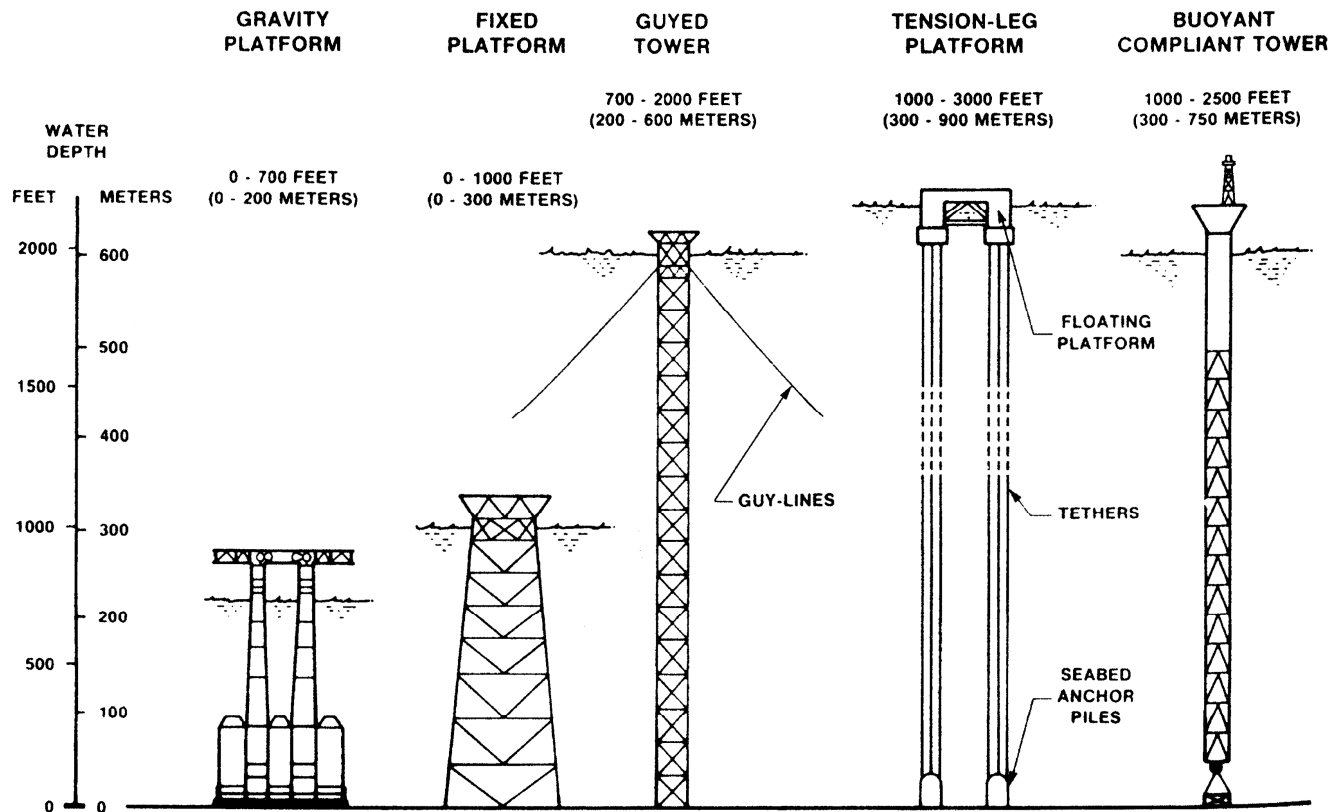
Many types of offshore structures have been developed since 1947, when the first steel structure was installed in 18 feet of water. Since that time over 4100 template-platforms have been constructed on the U.S. continental shelf in water depths less than 600 feet (Fig. 84.13).

Deep-water marine structures include gravity platforms, fixed platforms, guyed tower, tension-leg platform, and a buoyant compliant tower (Fig. 84.14).

**Figure 84.13** Template-type pile foundation structure. (Source: Young, A. G. 1991. Marine foundation studies. In *Handbook of Coastal and Ocean Engineering, Volume 2*, ed. J. B. Herbich. Gulf, Houston, TX. Copyright 1990 by Gulf Publishing Company, Houston, TX. Used with permission. All rights reserved.)



**Figure 84.14** Range of water depths for various types of deep-water marine structures. (Source: Marine Board, National Research Council. 1989. *Our Seabed Frontier*<sup>34</sup>*Challenges and Choices*. National Academy Press, Washington, DC.)



Wave forces on certain types of offshore platforms are computed by the Morrison equation, which is written as the sum of two individual forces, inertia and drag. The equation may be written as

$$f(t) = C_M \rho \frac{\pi}{4} D^2 \dot{u}(t) + \frac{1}{2} C_D \rho D |u(t)| u(t) \quad (84.34)$$

The force,  $f$ , as a function of time,  $t$ , is written as a function of the horizontal water particle velocity,  $u(t)$ , and the horizontal water particle acceleration,  $\dot{u}(t)$ , at the axis of the cylinder, and is dependent on the water density,  $\rho$ . The quantities  $C_M$  and  $C_D$  are defined as the inertia (or mass) coefficient and the drag coefficient, respectively.

The design and dynamic analysis of offshore platforms, which include jacket structures, topside structures, pile foundations, and dynamic analysis, may be found in Hsu [1991]; discussion of wave forces is given in Chakrabarti [1991].

## Defining Terms

**Armor unit:** A relatively large quarry stone or concrete shape that is selected to fit specified geometric characteristics and density. It is usually of nearly uniform size and usually large enough to require individual placement. In normal cases it is used as primary wave protection and is placed in thicknesses of at least two units.

**Artificial nourishment:** The process of replenishing a beach with material (usually sand)

obtained from another location.

**Attenuation:** (1) A lessening of the amplitude of a wave with distance from the origin. (2) The decrease of water-particle motion with increasing depth. Particle motion resulting from surface oscillatory waves attenuates rapidly with depth and practically disappears at a depth equal to a surface wavelength.

**Bar:** A submerged or emerged embankment of sand, gravel, or other unconsolidated material built on the sea floor in shallow water by waves and currents.

**Diffraction:** The phenomenon by which energy is transmitted laterally along a wave crest. When a part of a train of waves is interrupted by a barrier, such as a breakwater, the effect of diffraction is manifested by propagation of waves into the sheltered region within the barrier's geometric shadow.

**Dunes:** (1) Ridges or mounds of loose, wind-blown material, usually sand. (2) Bed forms smaller than bars but larger than ripples that are out of phase with any water-surface gravity waves associated with them.

**Ebb current:** The tidal current away from shore or down a tidal stream, usually associated with the decrease in height of the tide.

**Fetch:** The area in which seas are generated by a wind having a fairly constant direction and speed. Sometimes used synonymously with *fetch length* or *generating area*.

**Flood current:** The tidal current toward shore or up a tidal stream, usually associated with an increase in the height of the tide.

**Groin:** A shore protection structure built (usually perpendicular to the shoreline) to trap littoral drift or retard erosion of the shore.

**Harbor oscillation (harbor surging):** The nontidal vertical water movement in a harbor or bay. The vertical motions are usually low, but when oscillations are excited by a tsunami or storm surge, they may be quite large. Variable winds, air oscillations, or surf beat also may cause oscillations. See **seiche**.

**Hurricane:** An intense tropical cyclone in which winds tend to spiral inward toward a core of low pressure, with maximum surface wind velocities that equal or exceed 33.5 meters per second (75 mph or 65 knots) for several minutes or longer at some points. *Tropical storm* is the term applied if maximum winds are less than 33.5 meters per second.

**Mean high water (MHW):** The average height of the high waters over a 19-year period. For shorter periods of observations, corrections are applied to eliminate known variations and reduce the results to the equivalent of a mean 19-year value.

**Probable maximum water level:** A hypothetical water level (exclusive of wave run-up from normal wind-generated waves) that might result from the most severe combination of hydrometeorological, geoseismic, and other geophysical factors and that is considered reasonably possible in the region involved, with each of these factors considered as affecting the locality in a maximum manner. This level represents the physical response of a body of water to maximum applied phenomena such as hurricanes, moving squall lines, other cyclonic meteorological events, tsunamis, and astronomical tide, combined with maximum probable ambient hydrological conditions such as wave setup, rainfall, runoff, and river flow. It is a water level with virtually no risk of being exceeded.

**Refraction:** (1) The process by which the direction of a wave moving in shallow water at an angle

to the contours is changed: The part of the wave advancing in shallower water moves more slowly than that part still advancing in deeper water, causing the wave crest to bend toward alignment with the underwater contours. (2) The bending of wave crests by currents.

**Scour:** Removal of underwater material by waves and currents, especially at the base or toe of a shore structure.

**Seawall:** A structure separating land and water areas, primarily designed to prevent erosion and other damage due to wave action.

**Seiche:** (1) A standing wave oscillation of an enclosed water body that continues, pendulum fashion, after the cessation of the originating force, which may have been either seismic or atmospheric. (2) An oscillation of a fluid body in response to a disturbing force having the same frequency as the natural frequency of the fluid system. Tides are now considered to be seiches induced primarily by the periodic forces caused by the sun and moon.

**Significant wave:** A statistical term relating to the one-third highest waves of a given wave group and defined by the average of their heights and periods. The composition of the higher waves depends upon the extent to which the lower waves are considered.

**Wave spectrum:** In ocean wave studies, a graph, table, or mathematical equation showing the distribution of wave energy as a function of wave frequency. The spectrum may be based on observations or theoretical considerations. Several forms of graphical display are widely used.

## References

- Boussinesq, J. 1877. *Essai sur la theorie des eaux courantes*. Mem. divers Savants a L'Academie des Science, No. 32:56.
- Bruun, P. 1954. *Coast Erosion and the Development of Beach Profiles*. Tech. Memo. No. 44, 1954. Beach Erosion Board, U.S. Army Corps of Engineers.
- Chakrabarti, S. K. 1991. Wave forces on offshore structures. In *Handbook of Coastal and Ocean Engineering, Volume 2*, ed. J. B. Herbich. Gulf Publishing Co., Houston, TX.
- Dean, R. G. 1977. *Equilibrium Beach Profiles: U.S. Atlantic and Gulf Coasts*. Ocean Engineering T.R. No. 12. Department of Civil Engineering, University of Delaware, Newark, Delaware.
- Dean, R. G. 1990. Stream function wave theory and applications. In *Handbook of Coastal and Ocean Engineering, Volume 1*, ed. J. B. Herbich. Gulf Publishing Co., Houston, TX.
- Dean, R. G. 1991. Beach profiles. In *Handbook of Coastal and Ocean Engineering, Volume 2*, ed. J. B. Herbich. Gulf Publishing Co., Houston, TX.
- Department of the Army. 1984. *Shore Protection Manual*, vols. I and II. Department of the Army, Corps of Engineers, Coastal Engineering Research Center, Waterways Experiment Station, Vicksburg, MS.
- Department of the Army. 1992. *Automated Coastal Engineering System*. Department of the Army, Corps of Engineers, Coastal Engineering Research Center, Waterways Experiment Station, Vicksburg, MS.
- Goda, Y. 1985. *Random Seas and Design of Maritime Structures*. Tokyo University Press, Tokyo, Japan.
- Goda, Y. 1990. Random wave interaction with structures. In *Handbook of Coastal and Ocean*

- Engineering, Volume 1*, ed. J. B. Herbich. Gulf Publishing Co., Houston, TX.
- Herbich, J. B. (Ed.) 1990 (vol. 1), 1991 (vol. 2), 1992 (vol. 3). *Handbook of Coastal and Ocean Engineering*, Gulf Publishing Co., Houston, TX.
- Hsu, T. H. 1991. Design and dynamic analysis of offshore platforms. In *Handbook of Coastal and Ocean Engineering, Volume 2*, ed. J. B. Herbich. Gulf Publishing Co., Houston, TX.
- Kriebel, D. L., Dally, W. R., Dean, R. G. 1986. *Undistorted Froude Number for Surf Zone Sediment Transport*. pp. 1296–1310. Proc. 20th Coastal Engineering Conference, ASCE.
- Le Méhauté, B. 1969. *An Introduction to Hydrodynamics and Water Waves*. Report No. ERL 118-POL3-1&2. U.S. Department of Commerce, Environmental Science Services Administration, Washington, DC.
- Tsuchiya, Y. 1991. Threshold of sand movement. In *Handbook of Coastal and Ocean Engineering, Volume 2*, ed. J. B. Herbich. Gulf Publishing Co., Houston, TX.
- U.S. Army Corps of Engineers. 1983. *Engineering Manual: Hydraulic Design of Deep Draft Navigation Projects*. EM 1110-2-1613. U.S. Army Corps of Engineers, Washington, DC.
- Van der Meer, J. W. 1987. Stability of breakwater armor layers—Design formula. *J. Coastal Engin.* 11(3):219–239.
- Van der Meer, J. W. 1990. Rubble mounds—Recent modifications. In *Handbook of Coastal and Ocean Engineering, Volume 1*, ed. J. B. Herbich. Gulf Publishing Co., Houston, TX.
- Young, A. G. 1991. Marine foundation studies. In *Handbook of Coastal and Ocean Engineering, Volume 2*, ed. J. B. Herbich. Gulf Publishing Co., Houston, TX.

## Further Information

- ASCE Journal of Waterway, Port, Coastal and Ocean Engineering*: Published bimonthly by the American Society of Civil Engineers. Reports advances in coastal and ocean engineering.
- ASCE specialty conference proceedings: Published by the American Society of Civil Engineers. Report advances in coastal and ocean engineering.
- PIANC Bulletin*: Published quarterly by the Permanent International Association of Navigation Congresses, Brussels, Belgium. Reports case studies.
- Coastal Engineering Research Center (Technical reports, contract reports, miscellaneous papers): Published by the Army Corps of Engineers, Waterways Experiment Station, Vicksburg, MS.
- Sea Technology*: Published monthly by Compass Publications, Inc., Arlington, VA.
- IEEE proceedings of ocean conferences: Published by the Institute of Electrical and Electronics Engineers. Report advances in ocean engineering.
- Offshore Technology Conference Preprints: Published by the Offshore Technology Conference, Dallas, TX. Report annually on topics in ocean engineering.
- Marine Board, National Research Council reports: Published by the National Academy Press, Washington, DC.
- American Gas Association project reports: Published by the American Gas Association, Arlington, VA.
- American Petroleum Institute standards: Published by the American Petroleum Institute, Dallas, TX.
- Marine Technology Society conference proceedings: Published by the Marine Technology



Society, Houston, TX.

*World Dredging, Mining & Construction*: Published monthly by Wodcon Association, Irvine, CA.

*Terra et Aqua*: Published by the International Association of Dredging Companies, The Hague, the Netherlands.

*Center for Dredging Studies* abstracts: Published by the Center for Dredging Studies, Texas A&M University, College Station, TX.

Komar, P. D. 1983. *Handbook of Coastal Processes and Erosion*. CRC Press, Boca Raton, FL. A series of papers on coastal processes, beach erosion, and replenishment.

Bruun, P. 1989–90. *Port Engineering*, vols. 1 and 2, 4th ed. Gulf, Houston, TX. A comprehensive treatment on port and harbor design.

*International Dredging Review*: Bimonthly, Fort Collins, CO.

*Technical Standards for Port and Harbour Facilities in Japan*, 1980: Published by the Overseas Coastal Area Development Institute of Japan, 3-2-4 Kasumigaseki, Chiyoda-ku, Tokyo, Japan.

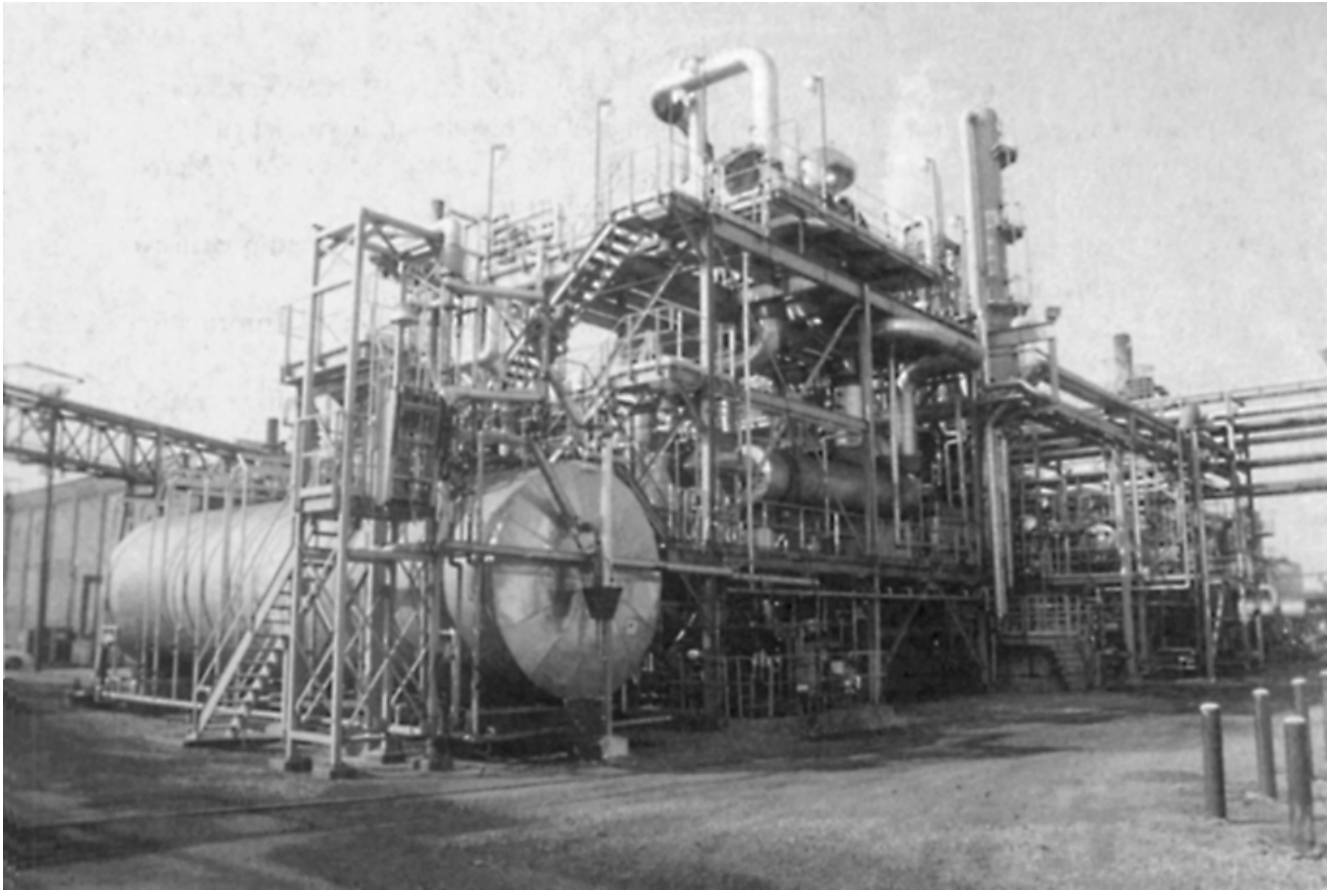
Herbich, J. B., Schiller, R. E., Jr., Watanabe, R. K., and Dunlap, W. A. 1984. *Seafloor Scour*. Marcel Dekker. New York. Design guidelines for ocean-founded structures.

Grace, R. A. 1978. *Marine Outfalls Systems*, Prentice Hall, Englewood Cliffs, NJ. A comprehensive treatment of marine outfalls.

Herbich, J. B. 1981. *Offshore Pipelines Design Elements*. Marcel Dekker. New York. Information relating to design of offshore pipelines.

Herbich, J. B. 1992. *Handbook of Dredging Engineering*. McGraw-Hill, NY. A comprehensive treatise on the subject of dredging engineering.

Jacko, R. B. “ Environmental Systems and Management”  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



Citizens Gas and Coke Utility in Indianapolis, Indiana, which operates as the public Department of Utilities for the City, completed the construction of a \$35 million ammonia destruct/desulfurization facility. The ammonia destruct facility, the first phase of the project, was completed in 1990 and has eliminated more than 5000 pounds of ammonia per day that previously was discharged into the city's wastewater treatment plants. The ammonia destruct system is over 90% efficient. The ammonia molecule is destroyed at high temperature using nickel catalyst, and even the resulting hydrogen gas is used as an energy source for the Citizens Gas and Coke Utility coke oven batteries.

The second-phase of the project, the desulfurization plant, was completed in 1994. The plant eliminates over 4000 tons of hazardous waste each year. 99.99% pure elemental sulfur is recovered for sale. Shown above is the sulfur recovery section of the plant.

The ammonia destruct facility was the first of its kind in the Western Hemisphere, and coupled with its sister desulfurization facility, represents the only combination ammonia destruct/desulfurization facility like it in the world. (Photo Courtesy of Citizens Gas and Coke Utility.)

# XIV

## Environmental Systems and Management

---

**Robert B. Jacko**

*Purdue University*

**85    Drinking Water Treatment** *A. Amirtharajah and S. C. Jones*

Water Quality • Drinking Water Regulations • Water Treatment Processes

**86    Air Pollution** *F. C. Alley and C. D. Cooper*

Control of Particulate Matter • Control of Gaseous Pollutants

**87    Wastewater Treatment and Disposal** *H. S. Peavy*

Wastewater Characteristics • Terminology in Wastewater Treatment • Sludge Advanced Wastewater Treatment • Wastewater Disposal and Reuse • Future of Wastewater Treatment and Reuse

**88    Solid Wastes** *R. E. McKinney*

Regulations • Characteristics • Generation • Collection • Transfer and Transport • Processing and Resource Recovery (Recycling) • Final Disposal

**89    Hazardous Waste Management** *H. M. Cota and D. Wallenstein*

Regulatory Overview • Definition of Hazardous • Waste Management of Hazardous Wastes • Hazardous Waste Treatment • Infectious Waste Management • Radioactive Waste Management • Mixed Wastes • Corrective Action Waste Minimization • Right-to-Know Laws • Computer Usage in Hazardous Waste Management

**90    Soil Remediation** *B. M. Kim and A. P. Shapiro*

Regulations • Treatment • Technologies • Landfilling and Containment • Soil Vapor Extraction • Thermal Treatments • Stabilization • Bioremediation • Soil Washing • Emerging Technologies

THE U.S. HAS LONG SINCE PASSED from being a "frontier society." In those times, the land, air, and water resources at our disposal seemed almost infinite. As this country moved through the Industrial Revolution, we utilized those resources as if they were, indeed, infinite in scope and automatically self-healing. Today we realize our natural resources are neither infinite in size nor totally self-healing from the insults of the pollutants we emit into the atmosphere, discharge into our water courses, or bury in our land. We are quickly realizing that the things we do on this earth in manufacturing, process operations, and in our daily lives must now have a component specifically related to minimizing the impact of these activities on the environment.

We must use the latest technologies to protect the atmosphere, our water resources, and the land we live on. Moreover, we must incorporate into our thinking an "environmental management system" (EMS) approach to everything we do. In process and manufacturing operations, we can no longer design from "cradle to grave"; rather, our design ethic must include a "cradle to cradle" concept. In other words, the product or process design pathway must include an environmental decision block where consideration is given to such things as use of nonhazardous materials; recycling or reuse of waste streams; and reuse of piece parts, fasteners, and so on. Ultimately,

everything we do and use will be given an end-use consideration, and the disposal concept will no longer exist. Cost savings will drive this new thinking. Companies now involved in EMS and this new way of designing are reporting significant cost savings as opposed to the expected negative economic impact of traditional environmental considerations.

On the horizon is a new international standard referred to as ISO 14 000, which is now in the development stage by a number of countries, including the U.S. This new standard, when it is published, will mandate a total EMS approach to the development of exported products and processes. Ideally, in the years to come, an environmental ethic will permeate everything we do, not just in consumer recycling, but in the way commercial and industrial products and processes are designed, manufactured, used, and disposed of—or, I should say, reused.

Amirtharajah, A., Jones, S. C. "Drinking Water Treatment"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Drinking Water Treatment

---

## 85.1 Water Quality

Microbial Contamination • Chemical Contamination • Aesthetic Aspects of Water Quality

## 85.2 Drinking Water Regulations

Total Coliform Rule • Surface Water Treatment Rule • Lead and Copper Rule • Future Regulations

## 85.3 Water Treatment Processes

Coagulation • Sedimentation • Filtration • Disinfection

**Appiah Amirtharajah**

*Georgia Institute of Technology*

**S. Casey Jones**

*Georgia Institute of Technology*

The goal of drinking water treatment is to provide a water supply that is both safe and pleasing to consume. To meet this goal, conventional water treatment plants utilize the physicochemical processes of coagulation, sedimentation, filtration, and disinfection. The following sections will review the important aspects of water quality, drinking water regulations, and the main processes in conventional water treatment.

## 85.1 Water Quality

---

### Microbial Contamination

Microorganisms present in water supplies can cause immediate and serious health problems. Infections by bacteria, viruses, and protozoa usually cause gastrointestinal distress; however, some, such as the bacteria *Vibrio cholerae*, can result in death. The protozoa *Giardia lamblia* and *Cryptosporidium* form chlorine resistant cysts, and just a few cysts can cause disease. Beyond these, a vast number of pathogenic organisms exist, and water suppliers cannot feasibly monitor for all of them. Therefore, they monitor for **indicator organisms** instead. The total coliform group of bacteria is the most common indicator. Unfortunately, some pathogens (e.g., viruses and protozoa) are more resistant to conventional water treatment processes than are total coliforms.

## Chemical Contamination

### Inorganic Contaminants

Toxic metals and other inorganic compounds contaminate water supplies from both human-made and natural sources. Nitrates, common in groundwaters, cause methemoglobinemia or "blue-baby syndrome" in infants. Fluoride, added by many water suppliers in small doses to prevent tooth decay, causes a weakening of the bones called *skeletal fluorosis* at concentrations above 4 mg/L. Radon, a naturally occurring radionuclide, may cause lung cancer from long-term exposures in the air after being released from water.

### Organic Contaminants

Most organic contaminants are either volatile organic chemicals (VOCs) or synthetic organic chemicals (SOCs). Dissolved VOCs transfer to the gas phase when exposed to the atmosphere. They are typically found in groundwaters that have been contaminated by leaks from industrial storage facilities. Examples include trichloroethylene (TCE) and tetrachloroethylene (PCE), both probable carcinogens. SOCs are more soluble in water and include pesticides and pollutants from leaking underground gasoline storage tanks, such as benzene and toluene. The health effects of SOCs range from central nervous system damage to cancer.

## Aesthetic Aspects of Water Quality

### Color and Turbidity

Inorganic metals such as iron and organic compounds such as natural organic matter (NOM) cause color. In addition to being aesthetically undesirable, color in the form of NOM is a precursor to the formation of **disinfection by-products (DBPs)**, which may cause cancer. Turbidity is the cloudiness of a water and is determined by measuring the amount of light scattered by suspended particles in water. The unit of turbidity is the nephelometric turbidity unit (NTU). Although not a direct threat to health, turbidity decreases the efficiency of disinfection, and particles that cause turbidity can transport harmful chemicals through a treatment plant.

### Taste and Odor

Zinc, copper, iron, and manganese can be detected by taste at concentrations of 1 mg/L. Hydrogen sulfide, a common contaminant in groundwaters, is detectable at concentrations of 100 ng/L. Many tastes and odors in surface waters result from biological activity of filamentous bacteria and blue-green algae. They produce geosmin and methylisoborneol (MIB), which cause an earthy or musty smell. Both are detected at concentrations of 10 ng/L [[Tate and Arnold, 1990](#)].

### Alkalinity

**Alkalinity** is a measure of the buffering capacity of a water. Alkalinity determines the magnitude of pH changes during coagulation and affects the solubility of calcium carbonate in the distribution system. In natural waters the carbonate system dominates alkalinity. In such systems, bicarbonate ( $\text{HCO}_3^-$ ), carbonate ( $\text{CO}_3^{2-}$ ), and hydroxide ( $\text{OH}^-$ ) ions are the major species of alkalinity.



## Temperature and pH

Temperature and pH affect coagulation, disinfection, and corrosion control. Equilibrium constants and reaction rates vary with temperature. The hydrogen ion concentration, measured as pH, is an important chemical species in these processes. Furthermore, the density and viscosity of water vary with temperature; thus, it is an important variable in the design of mixing, flocculation, sedimentation, and filtration process units.

## 85.2 Drinking Water Regulations

In the U.S., the Environmental Protection Agency (EPA) sets standards to regulate drinking water quality. Typically, the EPA establishes a maximum contaminant level goal (MCLG) and a maximum contaminant level (MCL) for each contaminant. An MCLG is the level at which no adverse health effect occurs. An MCL is set as close to the MCLG as is economically and technically feasible. A primary MCL is a legally enforceable standard based on a potential health risk. A secondary MCL is a nonenforceable standard based on a potential adverse aesthetic effect. [Table 85.1](#) lists some MCLs of important drinking water contaminants.

**Table 85.1** Some Drinking Water Standards

Contaminant	MCL <sup>1</sup>	Sources
Arsenic	0.05	Geological, pesticide residues; industrial waste and smelter operations
Asbestos	7 mfl <sup>2</sup>	Natural mineral deposits; also in asbestos/cement pipe
Benzene	0.005	Fuel (leaking tanks); solvent commonly used in manufacture of industrial chemicals, pharmaceuticals, pesticides, paints, and plastics
Cadmium	0.005	Natural mineral deposits; metal finishing; corrosion product plumbing
Chromium	0.1	Natural mineral deposits; metal finishing; textile, tanning, and leather industries
Mercury	0.002	Industrial/chemical manufacturing; fungicide; natural mineral deposits
Nitrate (as N)	10	Fertilizers, feedlots, sewage; naturally in soil, mineral deposits
Polychlorinated biphenyls (PCBs)	0.0005	Electric transformers, plasticizers; banned in 1979
Radium 226/228	5 pCi/L <sup>3</sup>	Radioactive waste; geological/natural
Tetrachloroethylene (PCE)	0.005	Dry cleaning; industrial solvent
Toluene	1	Chemical manufacturing; gasoline additive; industrial solvent
Trichloroethyle (TCE)	0.005	Waste from disposals of dry cleaning materials and manufacturing of pesticides, paints, waxes, and varnishes; paint stripper; metal degreaser

<sup>1</sup>Maximum contaminant level, in milligrams per liter unless otherwise noted.

<sup>2</sup>Million fibers per liter, with fiber length greater than 10  $\mu\text{m}$ .

<sup>3</sup>Picocurie (pCi) is the quantity of radioactive material producing 2.22 nuclear transformations per minute.

Source: U.S. Environmental Protection Agency, 1991. *Fact Sheet: National Primary Drinking Water Standards*.

In 1974 the U.S. Congress enacted the Safe Drinking Water Act (SDWA), which was the first law to cover all public drinking water utilities in the U.S. In 1986 Congress amended the SDWA, and, as a result, in 1989 the EPA promulgated the Surface Water Treatment Rule (SWTR). The SWTR established standards for all water treatment plants that use surface water as a source. Subsequently, the EPA also modified the standards for total coliforms, lead, and copper.

## Total Coliform Rule

The total coliform rule set the MCLG for total coliforms at zero. The MCL is currently based on a presence-absence test. Utilities must monitor the distribution system for total coliforms taking a specified number of samples based on the population served by the utility. No more than 5.0% of the samples taken per month should be positive for total coliform. Small systems taking fewer than forty samples per month can have no more than one positive sample.

## Surface Water Treatment Rule

### Turbidity

Filtered water turbidity must never exceed 5 NTU and should meet the MCL in 95% of the samples taken either continuously or every four hours. The MCL is either 0.5 or 1.0 NTU depending on the type of filter used. For conventional and direct filtration the MCL is 0.5 NTU.

### Disinfectant Residual

The disinfectant concentration or residual entering the distribution system cannot be less than 0.2 mg/L for more than four hours. A disinfectant residual must be detectable in 95% of the samples taken monthly at consumers' taps for two consecutive months.

### Treatment Techniques

Instead of setting an MCL for the pathogens *Giardia lamblia* and viruses, the U.S. EPA specified a treatment technique, that is, disinfection and possibly filtration. To ensure the required disinfection, utilities must determine *CT* values, where *C* is the concentration of disinfectant and *T* is the contact time between disinfectant and water. The *CT* value required for *Giardia lamblia* corresponds to 99.9% (3 log) removal or inactivation, and for viruses it corresponds to 99.99% (4 log) removal or inactivation. Filtration must be practiced unless the utility demonstrates that adequate treatment occurs without it and that the water source is free from significant contamination. Conventional filtration plants receive a *CT* credit of 2.5 log removal for *Giardia lamblia* and 2.0 log removal of viruses.

## Lead and Copper Rule

The EPA also specified a treatment technique for lead and copper. However, in this case, the treatment technique need not be implemented unless an action level is exceeded. Action levels of 0.015 mg/L for lead and 1.3 mg/L for copper are set at the consumer's tap. If these action levels are exceeded during an extensive sampling program, the utility must perform a corrosion control study to determine the best technique to minimize the corrosivity of the water entering the distribution system.

## Future Regulations

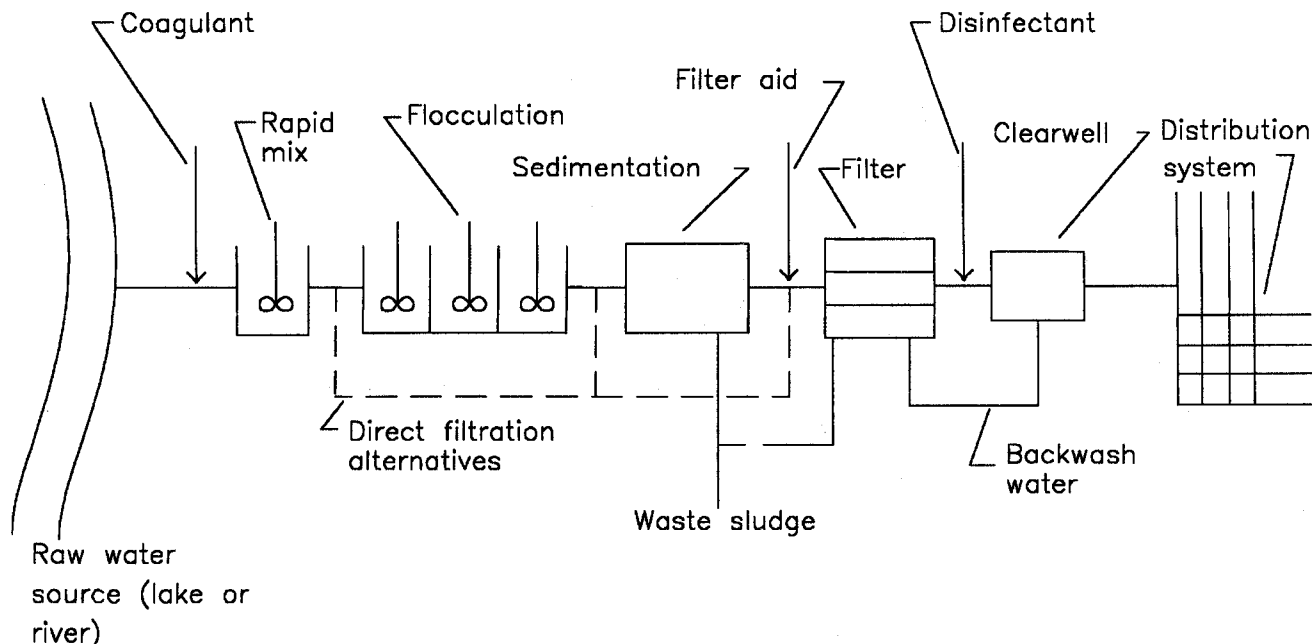
At present, no regulation exists regarding the protozoa *Cryptosporidium*, in part because of a lack of technical information about the organism. The EPA is in the process of collecting this information. Regulations are also expected in the near future for DBPs. The EPA is expected to lower the MCLs for total trihalomethanes (THMs) from 0.10 mg/L to 0.08 mg/L and to set MCLs for total haloacetic acids at 0.04 mg/L. In addition, removals of total organic carbon, a precursor to DBP formation, are to be specified under an Enhanced Surface Water Treatment Rule.

## 85.3 Water Treatment Processes

---

Figure 85.1 shows the treatment processes in a conventional surface water treatment plant. After being withdrawn from a source (lake or river), raw water is a suspension of small, stable colloidal particles whose motions are governed by molecular diffusion. In coagulation these particles are destabilized by the addition of a coagulant during rapid mixing. Flocculation promotes the collisions of these unstable particles to produce larger particles called *flocs*. In sedimentation, these flocs settle under the force of gravity. The particles that do not settle are removed during filtration. A disinfectant such as chlorine is then added, and, after a certain amount of contact time, the treated water is distributed to consumers. Direct filtration plants omit the sedimentation and occasionally the flocculation processes. These plants are suitable for raw waters with low to moderate turbidities and low color. The following sections describe the underlying theory and design of each of the major processes: coagulation, sedimentation, filtration, and disinfection.

**Figure 85.1** Schematic of a conventional water treatment plant.



## Coagulation

In coagulation, small particles combine into larger particles. Coagulation consists of three separate and sequential processes: coagulant formation, particle destabilization, and interparticle collisions. The first two steps occur during rapid mixing, whereas the third occurs during flocculation. In natural waters, particles (from 10 nm to 100  $\mu\text{m}$  in size) are stable, because they have a negative surface charge.

### Mechanisms of Destabilization

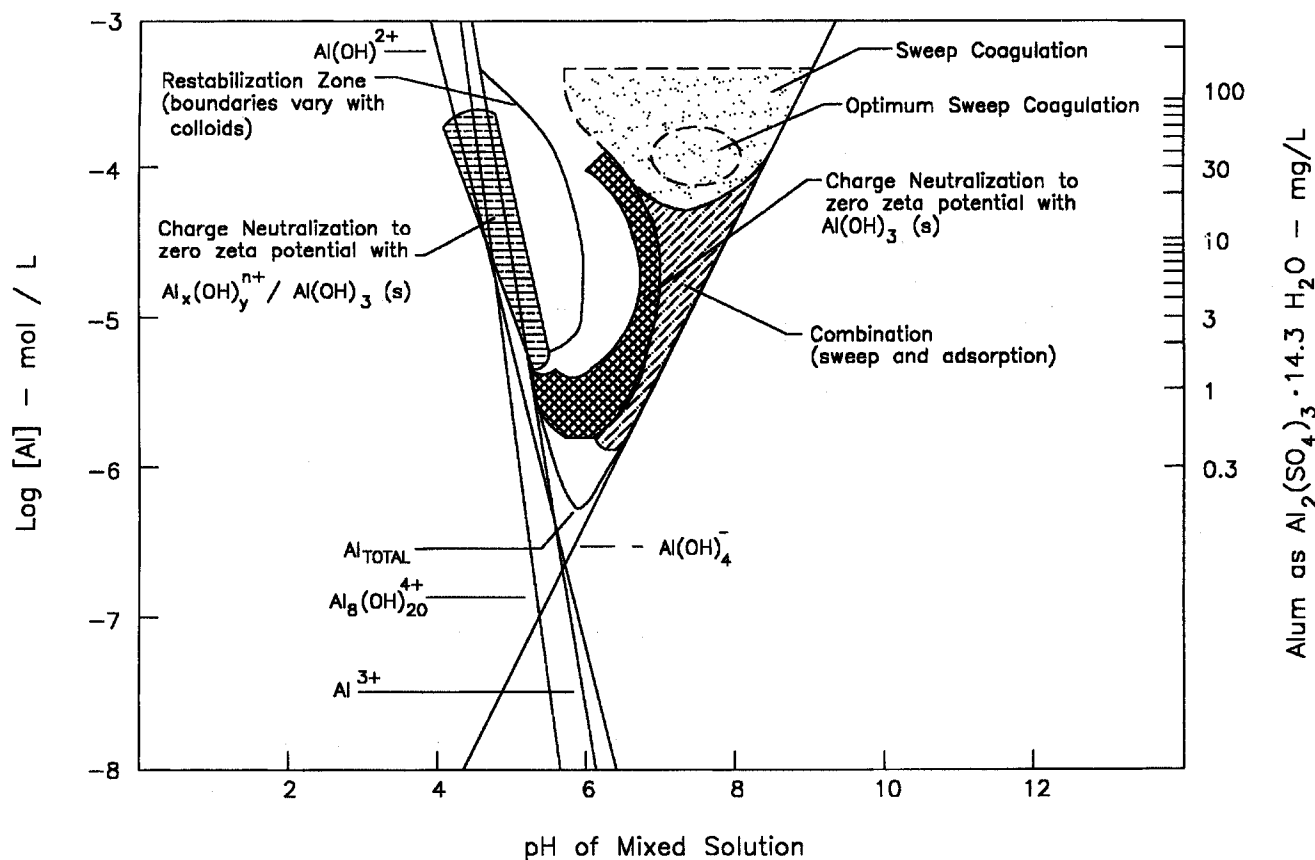
The possible mechanisms of particle destabilization are double layer compression, polymer bridging, **charge neutralization**, and **sweep coagulation**. In water treatment the last two mechanisms predominate; however, when organic polymers are used as coagulants, polymer bridging can occur. In charge neutralization the positively charged coagulant, either the hydrolysis species of a metal salt (alum or ferric chloride) or polyelectrolytes, adsorbs onto the surface of the negatively charged particles. As a result, the particles have no net surface charge and are effectively destabilized. In sweep coagulation a metal salt is added in concentrations sufficiently high to cause the precipitation of a metal hydroxide (e.g., aluminum hydroxide). The particles are enmeshed in the precipitate, and it "sweeps" the particles out of the water as it forms and settles.

### Solution Chemistry

With metal salt coagulants, the mechanism of coagulation is determined by the coagulant dose and the pH of the equilibrated solution. The most common coagulant is alum  $[\text{Al}_2(\text{SO}_4)_3 \cdot 14.3 \text{ H}_2\text{O}]$ . The alum coagulation diagram, shown in Fig. 85.2, indicates the regions where each mechanism dominates. A similar diagram exists for ferric chloride [Amirtharajah and O' Melia, 1990].

Important considerations in using the alum coagulation diagram are that the boundaries of the restabilization zone vary with the surface area of the raw water particles and that significant concentrations of NOM rather than turbidity can control the alum dose required for effective treatment.

**Figure 85.2** The alum coagulation diagram that defines the mechanism of coagulation based on pH and alum dose. (Source: Amirtharajah, A. and Mills, K. M. 1982. Rapid mix design for mechanisms of alum coagulation. *J. AWWA*. 74(4):210–216.)



### Rapid Mix Design

At a fundamental level the rapid-mixing unit provides encounters between molecules and particles in the source water and the coagulant species. These encounters are controlled by the hydrodynamic parameters and geometry of the mixer, molecular properties of the source water, and the kinetics of the coagulation reactions. Research indicates that coagulation by sweep coagulation is insensitive to mixing intensity. Although its applicability is questionable on theoretical grounds, the **G-value** is widely used to represent mixing intensity in both rapid mix and flocculation units. The G-value is computed as

$$G = \sqrt{\frac{P}{\mu V}} = \sqrt{\frac{\varepsilon}{\nu}} \quad (85.1)$$

where  $G$  is the velocity gradient ( $\text{s}^{-1}$ ),  $P$  is the *net* power input to the water (W),  $\mu$  is the dynamic viscosity of water ( $\text{Ns/m}^2$ ),  $V$  is the mixing volume ( $\text{m}^3$ ),  $\varepsilon$  is the rate of energy dissipation per mass of fluid ( $\text{W/kg}$ ), and  $\nu$  is the kinematic viscosity ( $\text{m}^2/\text{s}$ ). Mixing time,  $t$ , is an important design parameter, and it can vary from less than one second in some in-line mixers to over a minute in back-mix reactors. In general, short times ( $< 1$  s) are desired for the charge neutralization mechanism and longer times (10 to 30 s) for sweep coagulation.

### Flocculator Design

In flocculation, physical processes transform smaller particles into larger aggregates or flocs. Interparticle collisions cause the formation of flocs, and increased mixing with increased velocity gradients accelerates this process. However, if the mixing intensity is too vigorous, turbulent shear forces will cause flocs to break up. Studies of the kinetics of flocculation [Argaman and Kaufman, 1970] indicate that a minimum time exists below which no flocculation occurs regardless of mixing intensity and that using tanks in series significantly reduces the overall time required for the same degree of flocculation. Figure 85.3 illustrates these two conclusions. In current designs,  $G$ -values are tapered from one tank to the next with the highest  $G$ -value in the first tank and decreasing in each successive compartment.  $G$ -values are between 60 and  $10 \text{ s}^{-1}$ , and total detention times are close to 20 minutes.

### Sedimentation

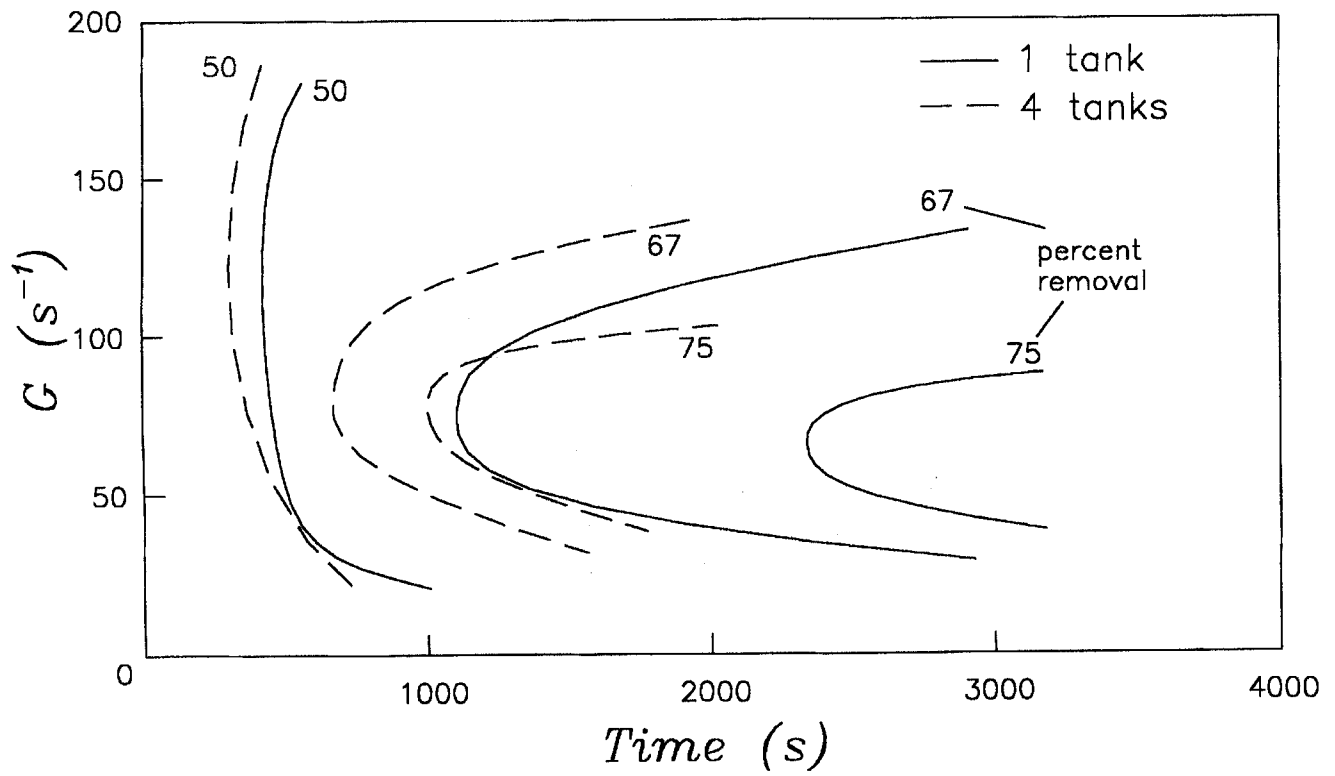
During sedimentation, gravity removes the flocs produced during the preceding flocculation process. These flocs continue to aggregate as they settle, and, as a result, experimental techniques are required to describe their settling behavior. Rectangular sedimentation basins are the most common in water treatment practice. Designs are based on the overflow rate, which is the flow rate divided by the surface area. The overflow rate indicates the settling velocity of the discrete (nonflocculant) particles that are removed with 100% efficiency. Typical overflow rates are 1.25 to 2.5 m/h. Plate and tube settlers are often added to the last two thirds of a basin to increase the overflow rate by a factor of 3.

### Filtration

In the U.S., the most common filters are dual-media filters, in which water flows by gravity through a porous bed of two layers of granular media. The top layer is anthracite coal, and the bottom layer is sand. Filters are operated until one of two criteria is exceeded—the effluent turbidity standard or the allowable head loss through the filter. The filters are cleaned by backwashing to remove the particles that have been collected on the filter media.

The removal of particles in a dual-media filter occurs within the pores of the filter and is mediated by transport mechanisms that carry small particles across fluid streamlines to distances

**Figure 85.3** A graph illustrating the benefit of tanks in series for flocculation. (Source: Argaman, Y. and Kaufman, W. J. 1970. Turbulence and flocculation. *ASCE J. Sanitary Engineering Div.* 96(SA2):223–241.)



close to the filter grains (also called *collectors*). When particles are very close to the collectors, short-range surface forces cause the collector to capture the particle. The dominant transport mechanisms in water filtration are diffusion and sedimentation. Diffusion is transport resulting from random Brownian motion by bombardment of the particle by molecules of water. Diffusion is increasingly important for particles less than 1  $\mu\text{m}$  in size. Sedimentation is due to the force of gravity and the associated settling velocity of the particle, which causes it to cross streamlines and reach the collector. This mechanism becomes increasingly important for particles greater than 1  $\mu\text{m}$  in size (for a size range of 5 to 25  $\mu\text{m}$ ). The combination of these two mechanisms results in a minimum net transport efficiency for a size of approximately 1  $\mu\text{m}$ . It is interesting to extrapolate this result to two important microbial contaminants. Cysts of *Giardia lamblia*, with dimensions of 10 to 15  $\mu\text{m}$ , are probably removed by the sedimentation mechanism, whereas *Cryptosporidium*, with a dimension close to 3 to 5  $\mu\text{m}$ , is probably close to the minimum net transport efficiency. Unfortunately, a theory of filtration that is sufficiently general and predictive does not yet exist. Therefore, designers must rely on empirical evidence from pilot-scale tests for guidance.

### **Chemical Pretreatment**

Evidence clearly shows that chemical pretreatment for particle destabilization is the most important variable affecting filtration efficiency. Plant-scale studies have shown that filtration rates between 5 and 15 m/h can treat water equally well given adequate chemical pretreatment.

### **Initial Degradation and Filter Ripening**

Just after backwashing, filters typically have poor effluent quality, and the quality improves during the course of the run. This improvement with time is called *filter ripening*. Studies have shown that greater than 90% of the particles passing through a well-operated filter do so during the initial stages of filtration. If the duration of filter ripening is short, then the initial filter effluent can be wasted until the effluent turbidity reaches a desired level. Particles removed during a filter run function as collectors themselves, resulting in filter ripening.

### **Rate Changes**

Any rate changes during filtration cause a significant deterioration of the effluent quality. The degradation in quality can be quantitatively correlated directly with the magnitude of the rate change and inversely with the duration of the rate change. This fact forms the basis for the superiority of variable declining-rate filters over the traditional constant-rate filters. In variable declining-rate filters the rate through the filter is controlled by the hydraulics of the filter; therefore, clean filters have a higher rate than dirty ones. However, in constant-rate filters, the rate is controlled by a mechanically actuated valve in the effluent piping. Each time the valve opens, the rate through the filter changes slightly, possibly causing a poorer effluent quality. With well-destabilized suspensions, the difference in effluent quality from these systems is minimal [Amirtharajah, 1988].

### **Polymers**

Polyelectrolytes can assist in improving effluent quality when added in small amounts as a filter



aid (0.1–1.0 mg/L) just prior to filtration. Polymers cause the attachment forces to be stronger, and, therefore, backwashing is more difficult. Polymers are important in direct-filtration plants where sedimentation and possibly flocculation are not included. However, polymers are not a substitute for adequate pretreatment.

### **Backwashing**

In the U.S., filters have traditionally been backwashed by fluidizing the filter media for a specific period of time. However, very few particle contacts occur between fluidized particles. Hence, particulate fluidization with water alone is an intrinsically weak process. Air scour with subfluidization water wash causes abrasions between particles throughout the depth of the bed. Surface wash causes collisions at the top of the bed. Both processes are effective auxiliaries for cleaning. When polymers are used, air scour or surface wash is necessary.

### **Disinfection**

A variety of disinfectants are available in water treatment, including chlorine, chloramines, chlorine dioxide, and ozone. In the U.S., however, chlorine is the most common disinfectant. Chlorine gas is added to water to form hypochlorous acid (HOCl). At pHs between 6 and 9, HOCl dissociates to form the hypochlorite ion ( $\text{OCl}^-$ ) and hydrogen ion ( $\text{H}^+$ ). HOCl has the greatest disinfection power. The extent of disinfection in a water treatment plant is determined by computing *CT* values as mentioned in section 85.2. The *CT* value required varies with chlorine concentration, pH, and temperature.

Although increasing the *CT* value may provide a large factor of safety against microbial contamination, disinfection causes the formation of disinfection by-products (DBPs), which are suspected carcinogens. DBPs result from reactions between disinfectants and NOM, which is ubiquitous in natural waters. The most common DBPs from chlorine are the THMs: chloroform, bromodichloromethane, dibromochloromethane, and bromoform. Other technologies such as membranes and adsorption will be used in the future, together with the traditional water treatment processes, to reduce the threat from both microorganisms and DBPs.

### **Defining Terms**

**Alkalinity:** The ability of a water to resist changes in pH. Includes the following species: carbonate, bicarbonate, and hydroxide.

**Charge neutralization:** A mechanism of coagulation in which positively charged coagulants adsorb to the surface of negatively charged colloids. The resulting colloids, with no net surface charge, are effectively destabilized.

**Disinfection by-products (DBPs):** By-products that result from the reactions of disinfectants such as chlorine with the natural organic matter that is present in all natural waters. DBPs such as trihalomethanes are suspected carcinogens.

**G-value:** A measure of mixing intensity in turbulent mixing vessels. Used in the design of both rapid mix and flocculation units.

**Indicator organisms:** Easily detectable organisms that act as a surrogate for water-borne

pathogens. Although not pathogenic themselves, indicators are present in the same environs as pathogens in larger concentrations.

**Sweep coagulation:** A mechanism of coagulation in which particles are enmeshed in a precipitate of metal hydroxide, such as aluminum hydroxide.

## References

- Amirtharajah, A. and Mills, K. M. 1982. Rapid mix design for mechanisms of alum coagulation. *J. AWWA*. 74(4):210–216.
- Amirtharajah, A. 1988. Some theoretical and conceptual views of filtration. *J. AWWA*. 80(12):36–46.
- Amirtharajah, A. and O'Melia, C. R. 1990. Coagulation processes: Destabilization, mixing, and flocculation. In *Water Quality and Treatment*, 4th ed., ed. F. W. Pontius, pp. 269–365. McGraw-Hill, New York.
- Argaman, Y. and Kaufman, W. J. 1970. Turbulence and flocculation. *ASCE J. Sanit. Eng. Div.* 96(SA2):223–241.
- Tate, C. H. and Arnold, K. F. 1990. Health and aesthetic aspects of water quality. In *Water Quality and Treatment*, 4th ed., ed. F. W. Pontius, pp. 63–156. McGraw-Hill, New York.
- U.S. Environmental Protection Agency. 1991. *Fact Sheet: National Primary Drinking Water Standards*. U.S. E.P.A., Washington, DC.

## Further Information

### Design Textbooks

- Kawamura, S. 1991. *Integrated Design of Water Treatment Facilities*. John Wiley & Sons, New York.
- James Montgomery Consulting Engineers. 1985. *Water Treatment Principles and Design*. John Wiley & Sons, New York.

### Introductory Textbooks

- Davis, M. L. and Cornwell, D. A. 1991. *Introduction to Environmental Engineering*, 2nd ed. McGraw-Hill, New York.
- Peavy, H. S., Rowe, D. R., and Tchobanoglous, G. 1985. *Environmental Engineering*. McGraw-Hill, New York.

### Specialized Textbooks

- Amirtharajah, A., Clark, M. M., and Trussell, R. R., eds. 1991. *Mixing in Coagulation and Flocculation*. American Water Works Association Research Foundation, Denver, CO.
- Stumm, W. and Morgan, J. J. 1981. *Aquatic Chemistry*, 2nd ed. John Wiley & Sons, New York.
- Snoeyink, V. L. and Jenkins, D. 1980. *Water Chemistry*. John Wiley & Sons, New York.

### Journals

*Journal of the American Water Works Association*

*Water Research*  
*ASCE Journal of Environmental Engineering*  
*Environmental Science and Technology*

Alley, F. C., Cooper, C. D. "Air Pollution"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

### 86.1 Control of Particulate Matter

Electrostatic Precipitators • Fabric Filtration • Wet Scrubbing

### 86.2 Control of Gaseous Pollutants

Absorption Processes • Adsorption • Incineration Processes

#### **F. Chris Alley**

*Clemson University (Emeritus)*

#### **C. David Cooper**

*University of Central Florida*

Air pollution is defined as contamination of the atmosphere with solid particles, liquid mists, or gaseous compounds in concentrations that can harm people, plants, or animals, or reduce environmental quality. Particulate matter (PM) constitutes a major class of air pollution. Ambient (outdoor) air quality standards exist for particulate matter less than  $10\ \mu\text{m}$  (PM-10), and many industries have emission limits on total PM. Particles have various shapes, different chemical and physical properties, and a wide range of sizes—from less than  $0.01\ \mu\text{m}$  to over  $100\ \mu\text{m}$  in **aerodynamic diameter**. The major gaseous pollutants that are emitted into the air include sulfur oxides, nitrogen oxides, carbon monoxide, and volatile organic compounds (such as petroleum fuels and organic chemicals). Major sources of air pollution include combustion processes (especially fossil fuel power plants), motor vehicles, and the materials processing and petrochemical industries.

## **86.1 Control of Particulate Matter**

---

Control of PM involves separation and removal of the PM from a flowing stream of air. A control device is chosen based on the size and properties (density, corrosivity, reactivity) of the particles, the chemical and physical characteristics of the gas (flow rate, temperature, pressure, humidity, chemical composition), and the collection efficiency,  $E$ , desired. The calculation of  $E$  is based on the fraction of mass removed:

$$E = (M_i - M_e)/M_i \quad (86.1)$$

where  $E$  is the collection efficiency (fraction),  $M$  is the mass flow rate of the pollutant (g/s), and  $i$

and  $e$  are subscripts indicating the input or exit stream. The efficiency of most devices is a strong function of particle size, decreasing rapidly as particle size decreases.

A key operating parameter of most control devices is the pressure drop ( $\Delta P$ ). In general, to increase efficiency of a device, more energy must be expended, which often shows up as  $\Delta P$  through the system. But moving the air through that  $\Delta P$  can account for a major portion of the operating cost. The fan power required is given by

$$W = Q\Delta P/\eta \quad (86.2)$$

where  $W$  is the power (kW),  $Q$  is the volumetric flow rate ( $\text{m}^3/\text{s}$ ),  $\Delta P$  is the pressure drop (kPa), and  $\eta$  = fan/motor efficiency.

Cyclones are moderate-efficiency mechanical separators that depend on centrifugal force to separate PM from the air stream. These precleaners are typically much less expensive than the more efficient devices discussed herein: electrostatic precipitators, fabric filters, and wet scrubbers [Cooper and Alley, 1994].

## Electrostatic Precipitators

An electrostatic precipitator (ESP) applies electrical force to separate PM from the gas stream. A high voltage drop is established between the electrodes (many sets of large plates in parallel, with vertical wires in between). The particles being carried by the gas flowing between the electrodes acquire a charge, and then are attracted to an oppositely charged plate, while the cleaned gas flows through the device. The plates are cleaned by rapping periodically, and the dust is collected in hoppers in the bottom of the device.

The classic Deutsch equation (first published in 1922) is used for preliminary design and performance evaluation:

$$E = 1 - \exp(-wA/Q) \quad (86.3)$$

where  $w$  is the drift velocity,  $A$  is the collection area, and  $Q$  is the gas volumetric flow rate, all in a consistent set of units.

In design, this equation is used to estimate the total plate area needed, and thus the size and cost of the ESP. The drift velocity,  $w$ , is a key parameter, and is a function of many variables (including the particle's diameter and resistivity, the electrical field strength, and the gas viscosity). Ranges of values of some operating and design parameters are given in Table 86.1.

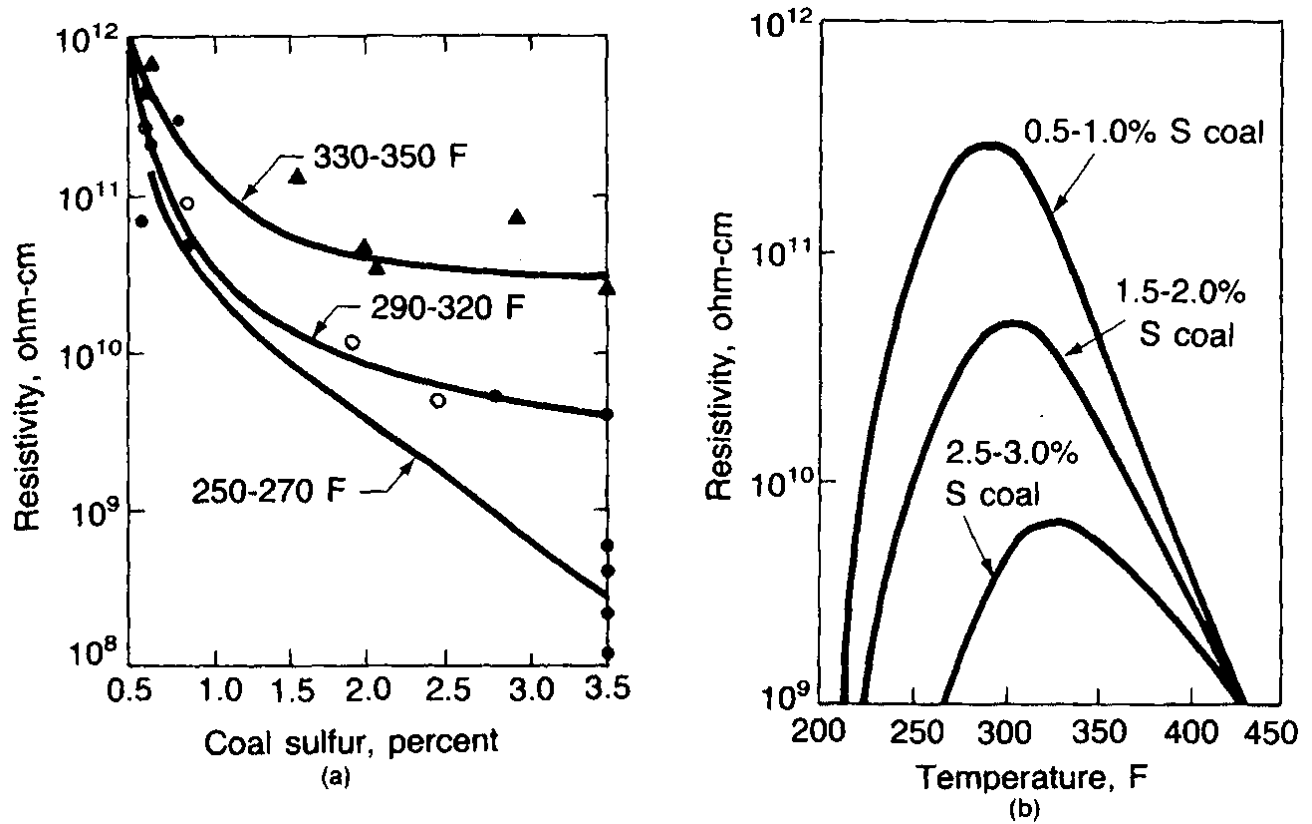
The resistivity of the dust is a measure of its resistance to electrical conduction. Resistivity is very important to the design and operation of an ESP because (1) for different dusts, it can range over many orders of magnitude, and (2) for a given ESP, an increase in dust resistivity can decrease the drift velocity markedly, reducing the collection efficiency. Resistivity is also quite dependent on operating conditions. For example, for coal fly ash, the resistivity is a strong function of temperature and sulfur content of the coal (see Fig 86.1).

**Table 86.1** Ranges of Values for Key ESP Parameters

Parameter	Range of Values
Drift velocity $w_e$	1.0–10 m/min
Channel width $D$	15–40 cm
Specific collection area (plate area/gas flow)	0.25–2.1 m <sup>2</sup> /(m <sup>3</sup> /min)
Gas velocity $u$	1.2–2.5 m/s (70–150 m/min)
Aspect ratio $R$ (duct length/height)	0.5–1.5 (not less than 1.0 for $\eta > 99\%$ )
Corona power ratio $P_c/Q$ (corona power/gas flow)	1.75–17.5 W/(m <sup>3</sup> /min)
Corona current ratio $I_c/A$ (corona current/plate area)	50–750 $\mu\text{A}/\text{m}^2$
Power density versus resistivity	
Ash Resistivity, ohm-cm	Power Density, W/m <sup>2</sup>
$10^4$ – $10^7$	43
$10^7$ – $10^8$	32
$10^9$ – $10^{10}$	27
$10^{11}$	22
$10^{12}$	16
$10^{13}$	10.8
Plate area per electrical set $A_s$	460–7400 m <sup>2</sup>
Number of electrical sections $N_s$	
a. In the direction of gas flow	2–8
b. Total	1–10 bus sections/(1000 m <sup>3</sup> min)

Source: Cooper, C. D. and Alley, F. C. 1994. *Air Pollution Control<sup>3/4</sup>A Design Approach*, 2nd ed. Waveland Press, Prospect Heights, IL. With permission.

**Figure 86.1** Variations of resistivity of coal fly ash. (Source: Cooper, C. D. and Alley, F. C. 1994. *Air Pollution Control<sup>3</sup>/4A Design Approach*, 2nd ed. Waveland Press, Prospect Heights, IL. With permission.)





ESPs can be designed for efficiencies above 99% on many types of dry dusts or wet fumes, and can handle large volumes of air with very low  $\Delta P$  (less than 1 in. H<sub>2</sub>O). However, they tend to have high capital costs and take up a lot of space. The capital cost is a quantitative function of the plate area [Vatavuk, 1990].

## Fabric Filtration

A fabric filter (or baghouse) operates on the age-old principle of filtering air through a cloth to remove dust. The air passes through the cloth, leaving the dust behind and providing a clean air stream. The dust builds up as a loosely packed mat on the cloth until removed by shaking or blowing; the dust is then collected in hoppers. There are three main types of baghouses (classified by cleaning method): shaker, reverse-air, and pulse-jet. The first two types have parallel compartments; each compartment is taken off-line, cleaned, and returned to service sequentially while the filtering continues in the other compartments. The pulse-jet baghouse has only one compartment and the bags are cleaned in sequence by blasts of high-pressure air while filtration occurs.

The efficiency of a fabric filter system is extremely high and is almost independent of particle size because of the mat of dust that builds up on the cloth. However, as the thickness of the dust mat increases, so does the pressure drop. The baghouse  $\Delta P$  can be related to key operating variables through the filter drag model:

$$\Delta P/V = K_1 + K_2 W \quad (86.4)$$

where  $V$  is the filtering velocity (m/min),  $W$  is the fabric area dust density (g/m<sup>2</sup>), and  $K_1$  and  $K_2$  are empirical constants.  $V$  is defined simply as the volumetric gas flow rate divided by the on-line fabric area. The filtering velocity is a key design parameter; some values are shown in Table 86.2.

**Table 86.2** Recommended Filtering Velocities in Baghouses

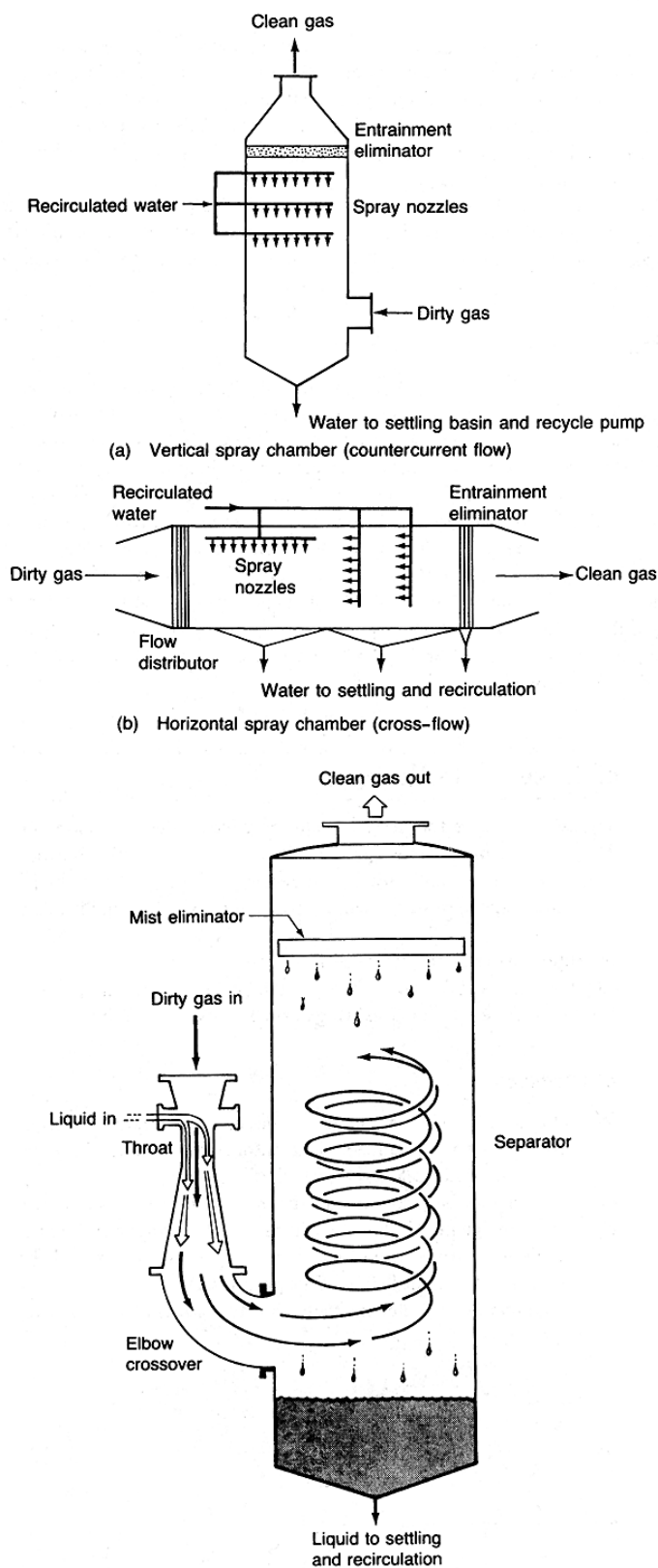
Type of Dust	Filtering Velocity, m <sup>3</sup> /m <sup>2</sup> or m/min	
	Shaker or Reverse-Air	Pulse-Jet
Carbon, graphite, sugar, paint pigments	0.6–0.7	1.6–2.4
Glass, gypsum, limestone, quartz	0.8–0.9	3.0–3.5
Leather, tobacco, grains	1.0–1.1	3.8–4.2

Baghouses often achieve efficiencies above 99.9% on many types of dry dusts, are fairly flexible to operating changes, and can handle large volumes of air with reasonable pressure drops (6–12 in. H<sub>2</sub>O). However, they cannot handle highly humid gases, they have high capital costs, and they take up a lot of space. The capital cost is related quantitatively to the fabric area [Vatavuk, 1990].

## Wet Scrubbing

A wet scrubber employs the principle of impaction and interception of the particles by droplets of water. The larger, heavier droplets are then easily separated from the gas. Later, the particles are separated from the water stream, and the water treated prior to reuse or discharge. Spray chambers and venturi scrubbers are shown in Fig. 86.2.

**Figure 86.2** Common types of wet scrubbers. (a),(b) Spray chambers. (c) venturi scrubber with cyclone separator. (Source: Cooper, C. D. and Alley, F. C. 1994. *Air Pollution Control—A Design Approach*, 2nd ed. Waveland Press, Prospect Heights, IL. With permission.)



During design, the collection efficiency of scrubbers can be related to a number of gas and liquid parameters through specific equations [Cooper and Alley, 1994]. Some of the most important variables include the liquid-to-gas ratio, the particle diameter, and, in a venturi, the gas velocity (which is directly related to pressure drop). In high-efficiency venturis,  $\Delta P$  can exceed 60 in. H<sub>2</sub>O. Extrapolating the performance of existing scrubbers can be done successfully using the contacting power approach. Contacting power is the energy expended per unit volume of gas treated and is related to collection efficiency through the following equation:

$$E = 1 - \exp [a(P_T)^b] \quad (86.5)$$

where  $P_T$  is the contacting power, kW/1000 m<sup>3</sup>/hr, and  $a$  and  $b$  are empirical constants.

Scrubbers have been used on a wide variety of dry or wet PM with efficiencies as high as 99%, while removing some soluble gases as well. Successful use of wet scrubbers also requires (1) good humidification of the gases prior to entering the scrubber, and (2) good separation of the water mist before exhausting the scrubbed gases. The disadvantages include potential for corrosion and the production of a liquid effluent that must be further treated. Capital costs of scrubbers have been related to the gas scrubbing capacity [Peters and Timmerhaus, 1991].

## 86.2 Control of Gaseous Pollutants

---

The most widely used processes for gaseous pollution control are absorption, adsorption, and incineration. The selection of the most practical process for a specific control application is based primarily on the chemical and physical properties of the pollutant to be removed. Some of the properties include vapor pressure, chemical reactivity, toxicity, solubility, flammability, and corrosiveness.

The following sections describe the design basics for each process and some typical applications. This section will refer to gases and vapors where gases are defined as substances far removed from the liquid state and vapors are substances existing near condensation conditions.

### Absorption Processes

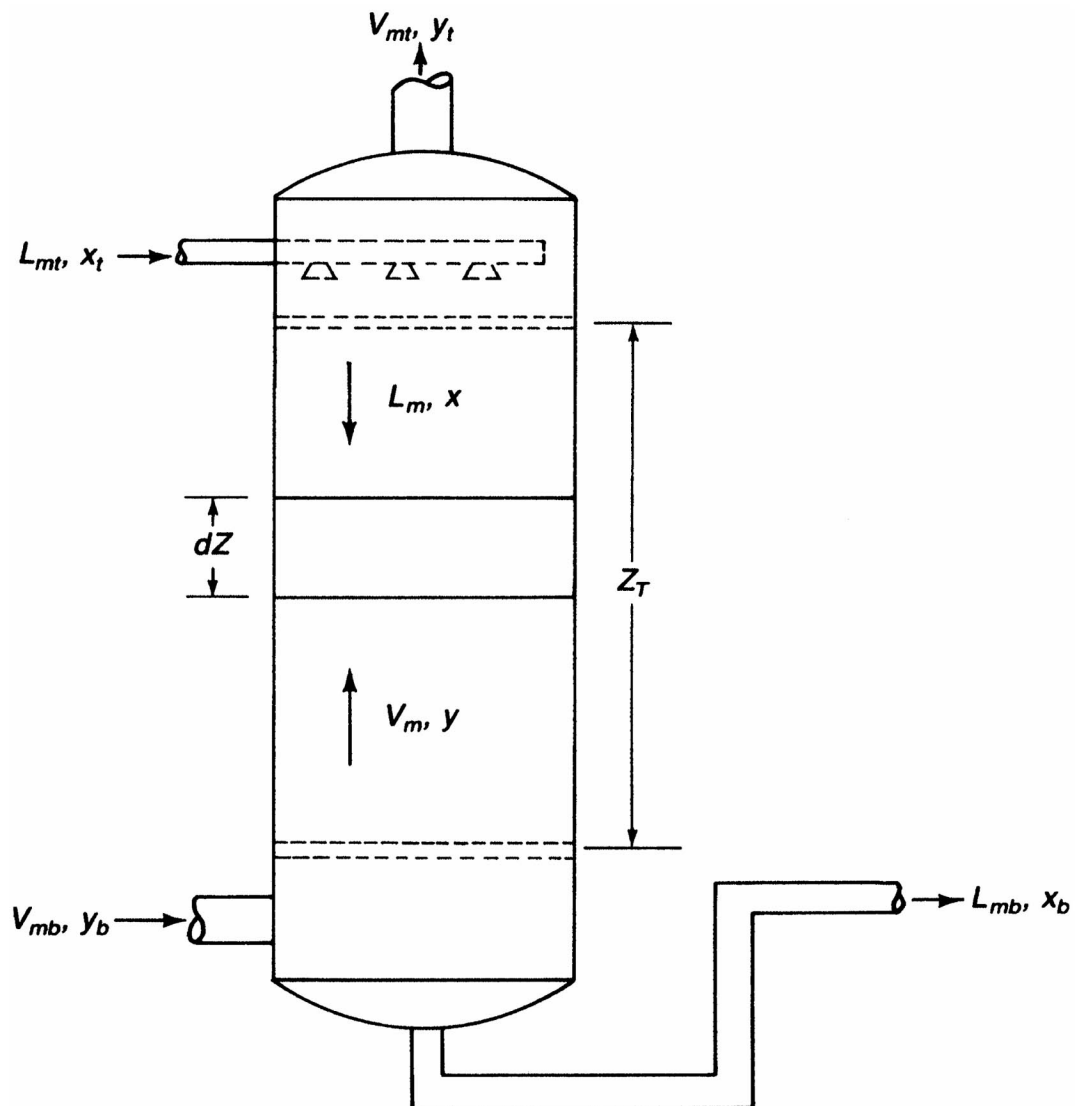
Absorption refers to the transfer of a material from a gas or vapor mixture to a contacting liquid, which in most cases is either water or an aqueous alkaline or acidic solution. The contacting process is typically carried out in a cylindrical tower partially filled with an inert packing to provide a large wetted surface to contact the incoming gas stream. The process is shown schematically in Fig. 86.3, where  $V$  and  $L$  designate the flow rates in mol/h of the vapor and liquid streams. The concentrations of the liquid streams in mol fraction are designated by  $x$  and the vapor phase concentrations are designated by  $y$ . The height of the packed section is shown as  $Z_t$  and subscripts indicate the location of the variable,  $t$  referring to the top of the tower and  $b$  the bottom. The height of the packed bed is a function of the rate of transfer of material (which we will refer to as the solute) from the gas phase to the liquid phase. The transfer rate is dependent on the concentration driving force across the gas-liquid interface, which is approximated by  $(y - y^*)$ , where  $y^*$  is the

mol fraction of the solute in the gas phase that would be in equilibrium with the solute concentration in the bulk liquid phase. The following relationship referred to as the operating line for the tower is developed from an overall material balance around the tower by substituting the solute-free liquid and gas flow rates,  $L'$  and  $V'$ .

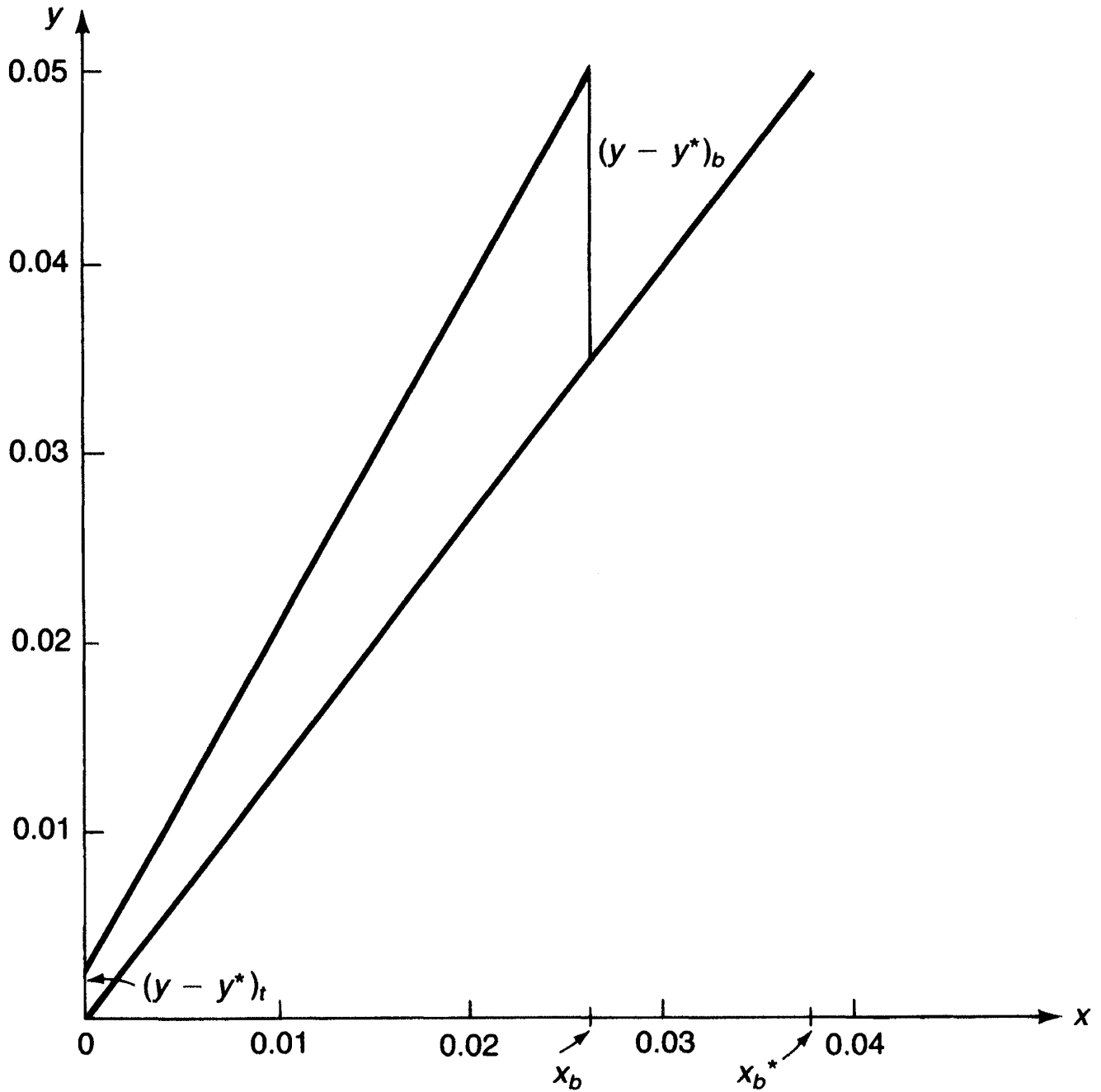
$$\frac{y}{1-y} = \frac{L'_m}{V'_m} \left( \frac{x}{1-x} \right) + \left[ \frac{y_b}{1-y_b} - \frac{L'_m}{V'_m} \left( \frac{x_b}{1-x_b} \right) \right] \quad (86.6)$$

where  $L'_m = L(1-x)$  and  $V'_m = V(1-y)$ . The operating line provides a relationship between the solute concentrations in the bulk liquid and gas phases. The overall driving force in the packed bed is shown graphically in Fig. 86.4, where the operating line is the upper curve and the lower line is the equilibrium line for the solute at the operating temperature and pressure of the tower ( $y^*$  vs.  $x$ ). These two lines will be linear only at low solute concentrations (lean gas mixtures), which fortunately is the case in many air pollution control applications.

**Figure 86.3** Schematic diagram of packed tower. (Source: Cooper, C. D. and Alley, F. C. 1994. *Air Pollution Control<sup>3</sup>A Design Approach*, 2nd ed. Waveland Press, Prospect Heights, IL. With permission.)



**Figure 86.4** Operating and equilibrium lines. (Source: Cooper, C. D. and Alley, F. C. 1994. *Air Pollution Control<sup>3</sup>4A Design Approach*, 2nd ed. Waveland Press, Prospect Heights, IL. With permission.)



The basic design equation for the height of the packed bed is

$$\left( \frac{K_y a}{G_{my}} \right) Z_t = \int_{y_t}^{y_b} \frac{dy}{(1-y)(y-y^*)} \quad (86.7)$$

The reciprocal of the term in parentheses, the overall mass transfer rate divided by the molar gas phase flux, is referred to as  $H_y$ , the height of a transfer unit based on the overall gas phase driving force. The value of the integral may then be thought of as the number of transfer units,  $N_{ty}$ . Hence,

Eq. (86.7) may also be written as

$$Z_t = (N_{ty})(H_y) \quad (86.8)$$

For the case of linear operating and equilibrium lines, the integral can be replaced with

$$\int_{y_t}^{y_b} \frac{dy}{(1-y)(y-y^*)} = \frac{y_b - y_t}{\Delta y_{LM}} \quad (86.9)$$

where

$$\Delta y_{LM} = (y - y^*)_{LM} = \frac{(y_b - y_b^*) - (y_t - y_t^*)}{\ln \left[ \frac{(y_b - y_b^*)}{(y_t - y_t^*)} \right]}$$

The value of  $H_y$  may be calculated from empirical relations that are presented in many texts or from experimental plots provided by manufacturers of tower packings. Typically,  $H_y$  values range from 1.0 to 3.0 feet.

The diameter of an absorption tower must be such that the gas and liquid velocities will provide good phase contact, but at the same time will not result in a gas pressure drop sufficient to cause liquid hold-up or flooding. Flooding–gas velocity correlations are also provided by suppliers of tower packing and are discussed in detail by McCabe *et al.* [1985] and Cooper and Alley [1994].

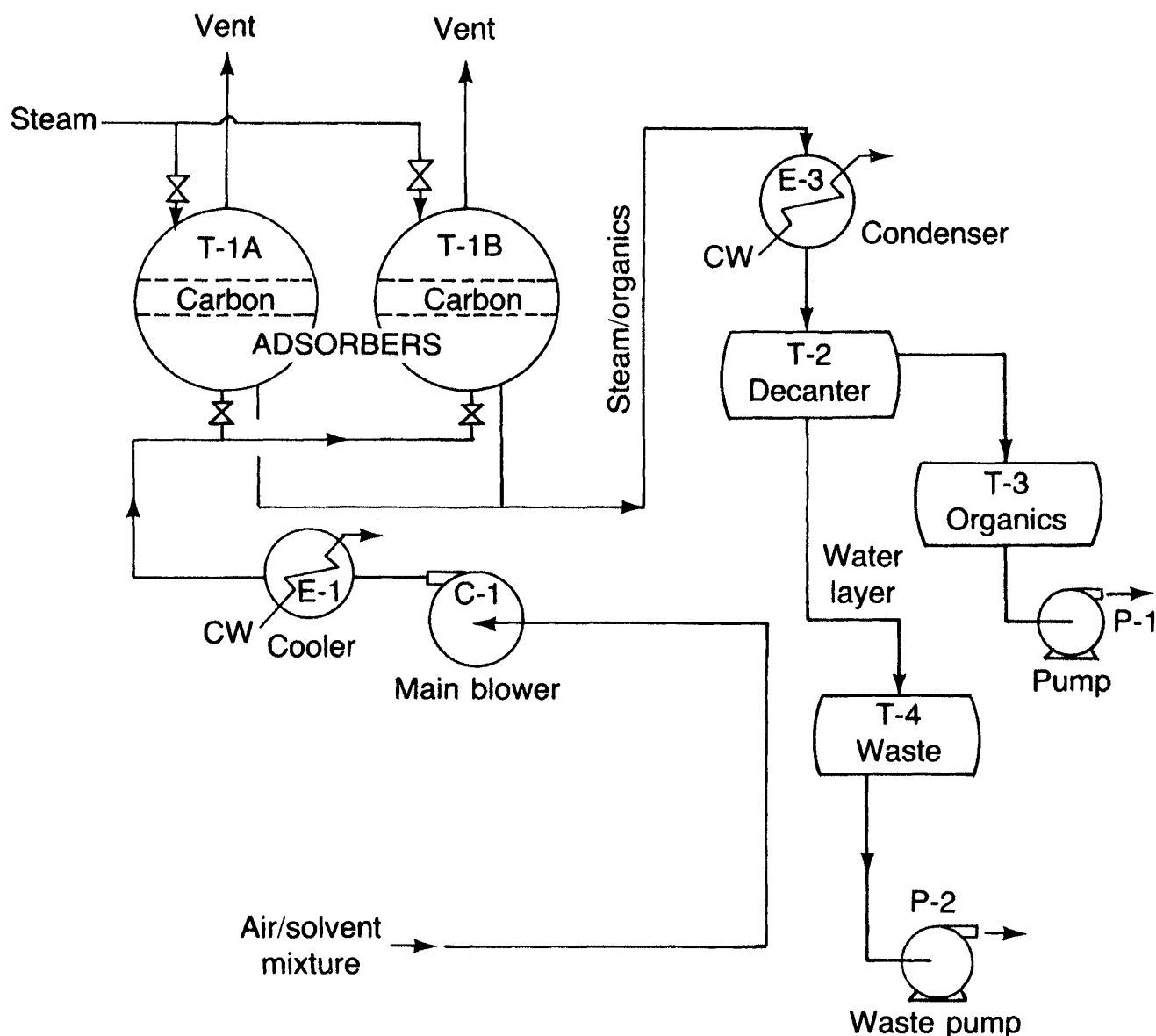
Tower costs are estimated based on the total weight of the tower shell plus 10 to 20% additional for manholes and fittings. To this is added the packing cost. Costing procedures for towers and auxiliaries are presented in detail by Peters and Timmerhaus [1991]. Major operating costs include blower power, spent absorbent liquid disposal, and (in the case of acid gas treatment) chemical costs.

In many applications gas absorbers are utilized to recover valuable chemicals from exhaust streams, which reduces the load on subsequent pollution control equipment. Examples of this application include recovery of light hydrocarbon ends from wellhead gas and recovery of HCl gas in hydrochloric acid production. Strictly air pollution control applications include removal of oxides of sulfur and nitrogen, chlorine, and ammonia.

## Adsorption

The removal of low-concentration gases and vapors from an exhaust stream by the adherence of these materials to the surface of a porous solid adsorbent is referred to as adsorption. Adsorbents used in air pollution applications include activated carbon, alumina, bauxite, and silica gel. Activated carbon (most frequently used) is prepared by partially oxidizing lignite, bituminous coal, and petroleum coke. A typical fixed-bed system employing carbon as the adsorbent is shown schematically in Fig. 86.5.

**Figure 86.5** Flow sheet for a fixed-bed system. (Source: Cooper, C. D. and Alley, F. C. 1994. *Air Pollution Control<sup>3</sup>A Design Approach*, 2nd ed. Waveland Press, Prospect Heights, IL. With permission.)

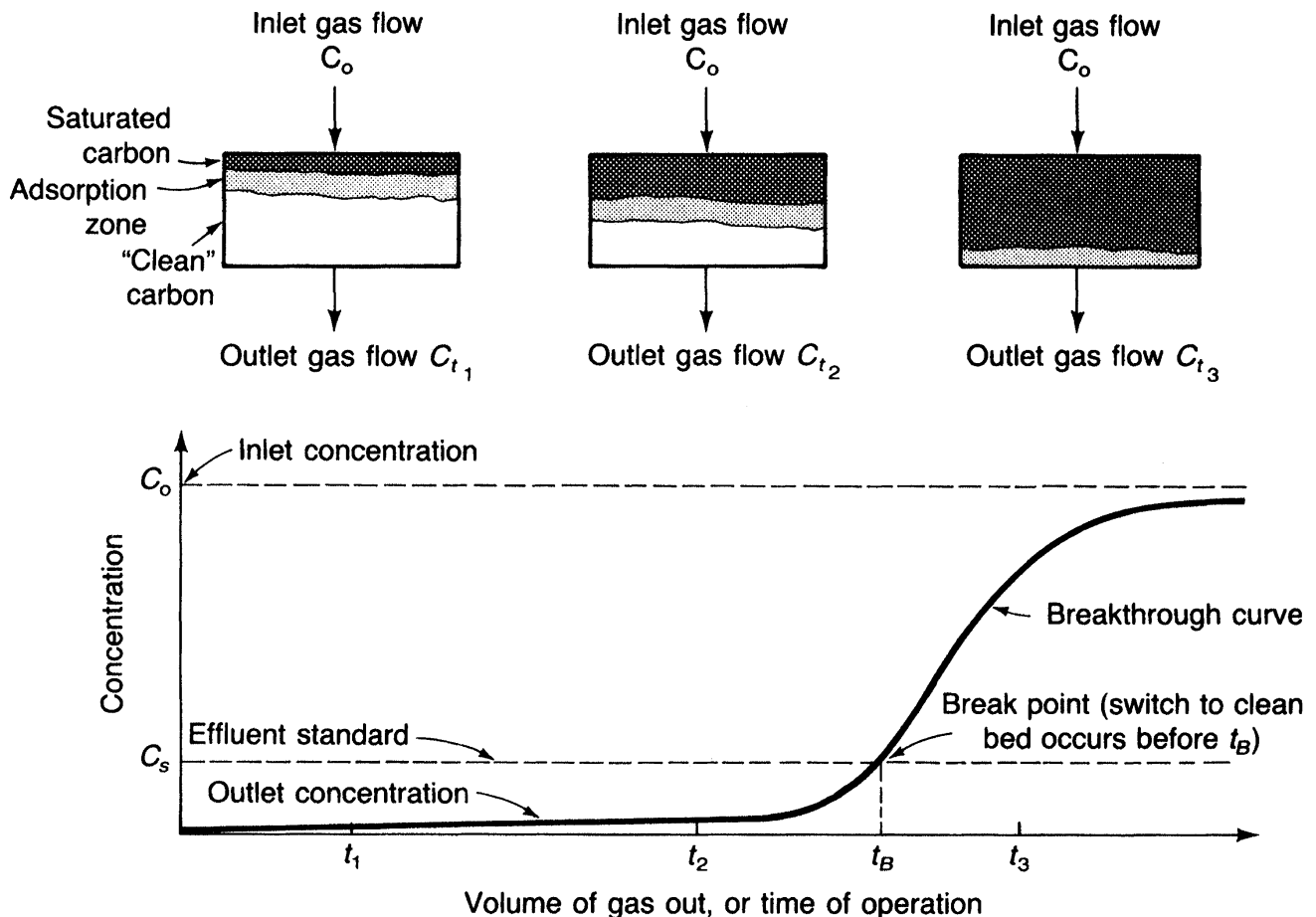


This system incorporates two horizontal cylindrical vessels, each containing a bed of granular activated carbon supported on a heavy screen. An air-solvent or air-pollutant mixture from a plant source flows through a cooler and then into one carbon bed while the other bed is being regenerated. A monitor-controller located in the vent stack detects a preset maximum concentration (breakthrough) and switches the incoming exhaust stream to the other bed, which has completed a regeneration cycle. The bed undergoing regeneration is contacted with saturated steam, which drives off the adsorbed material and carries it to a condenser, where it is condensed along with the steam. If the adsorbed material is insoluble in water, the condensed mixture goes to a decanter and is separated for final disposal.

When a material is adsorbed it may be weakly bonded to the solid (physical adsorption) or it may react chemically with the solid (chemisorption). Physical adsorption permits economical regeneration of the adsorbent and is practical in air pollution control applications such as solvent recovery and the removal of low-concentration odorous and toxic materials. The design of fixed-bed physical adsorption systems is described below.

The actual adsorption onto a carbon bed occurs in a concentration zone called the adsorption zone, as shown in Fig. 86.6. The zone is bounded by saturated carbon upstream and clean carbon downstream. Unlike the steady state conditions in an absorption tower, adsorption is an unsteady state process in which the mass transfer driving force goes to zero as the carbon bed becomes saturated with respect to the partial pressure of the adsorbate (material adsorbed) in the gas stream.

**Figure 86.6** The adsorption wave and breakthrough curve. (Source: Cooper, C. D. and Alley, F. C. 1994. *Air Pollution Control<sup>3</sup>4A Design Approach*, 2nd ed. Waveland Press, Prospect Heights, IL. With permission.)

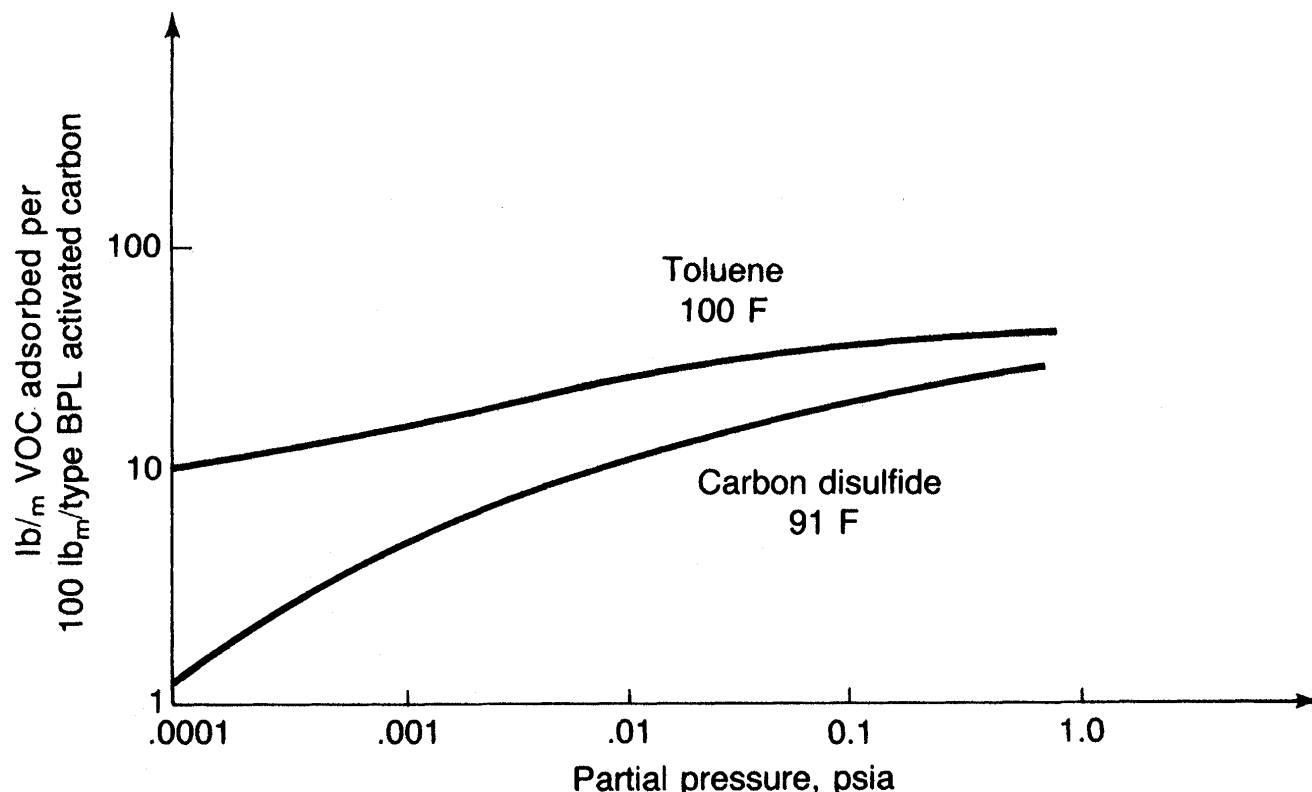


The quantity of adsorbent in each bed is based on the saturation capacity for the adsorbate at the operating conditions of the adsorber. The operating adsorbent capacity is normally supplied by the manufacturer of the adsorbent or estimated by using 30% of the saturation capacity, as shown on



the capacity curve (isotherm) illustrated in Fig. 86.7.

**Figure 86.7** Adsorption isotherms for activated carbon. (Source: Cooper, C. D. and Alley, F. C. 1994. *Air Pollution Control<sup>3</sup>4A Design Approach*, 2nd ed. Waveland Press, Prospect Heights, IL. With permission.)



For a two-bed system each bed must contain sufficient adsorbent to adsorb all solvent or pollutant in the incoming gas stream while the other bed is being regenerated, usually a period of 0.5 to 1.0 hour. The cross-sectional area of each bed is determined by assuming that the superficial velocity of the gas in the bed is in the range of 70 to 100 ft/s. The system blower power is based on the pressure drop through the bed, fittings, and condenser. In most applications, the expended bed is regenerated by passing low pressure-saturated steam through it at a rate of 0.3 to 0.5 pounds of steam per pound of carbon. The size of the condenser is based on 70 to 80% of this heat load. Cooper and Alley [1994] present examples of typical fixed-bed adsorber design.

Vatavuk [1990] reported that the purchased equipment cost (PEC) of packaged adsorber systems containing up to 14 000 pounds of carbon can be estimated by

$$PEC = 129(M_C)^{0.848} \quad (86.10)$$

where

PEC = purchase cost, 1988 dollars

$M_C$  = mass of carbon, pounds

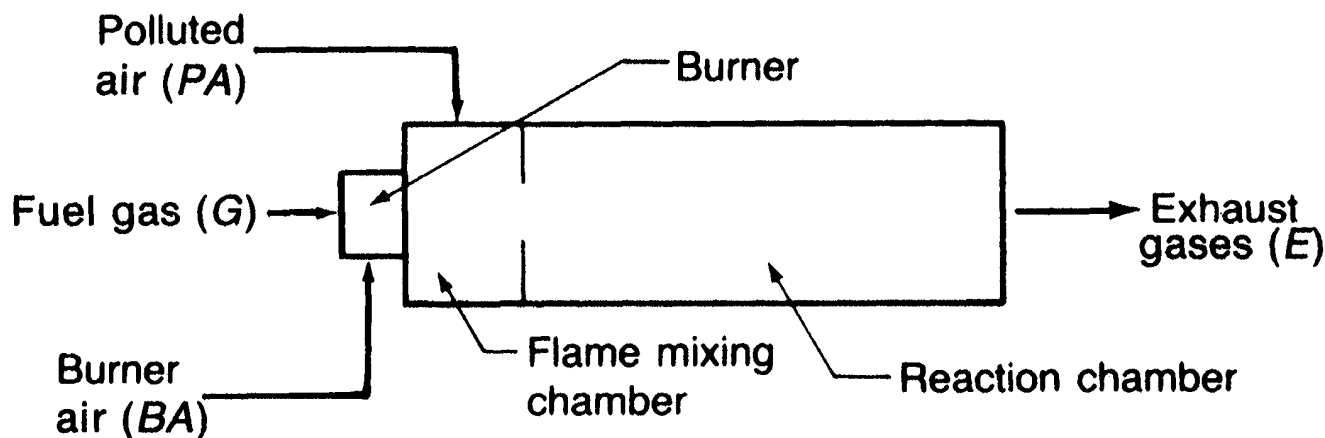
Typical grades of carbon utilized in adsorbers cost from two to three dollars per pound and have a normal service life of five to seven years. The major utility costs are for steam, blower power, and cooling water. Average blower horsepower requirement is 6–8 hp/1000 SCFM of gas throughput.

Carbon adsorption has been shown to be applicable for the control of many types of gaseous pollutants, including volatile organic compounds (VOC), chlorinated solvents, and odorous materials.

## Incineration Processes

Incineration of polluted exhaust streams offers an effective but costly means of controlling a wide variety of contaminants, including VOCs, which constitute a major fraction of the total annual emissions in the U.S. and play an important role in the **photochemical smog** process. A pollution control incinerator, also referred to as an afterburner or thermal oxidizer, is shown schematically in Fig. 86.8. The incoming polluted air is injected into the flame mixing chamber, heated, and held in the reaction chamber until the desired pollutant destruction efficiency is obtained.

**Figure 86.8** Schematic diagram of a vapor incinerator. (Source: Cooper, C. D. and Alley, F. C. 1994. *Air Pollution Control<sup>34</sup>A Design Approach*, 2nd ed. Waveland Press, Prospect Heights, IL. With permission.)



The process design of an incinerator requires specification of an operating temperature and a residence time in the reaction chamber. Typical operating condition ranges to provide the required "three T's" of combustion are temperature (1000–1500°F), residence time (0.3–0.5 s), and turbulence (mixing velocity 20–40 ft/s). Burner fuel requirements are estimated from a steady state enthalpy balance, as shown in Eq. (86.11):

$$0 = \dot{M}_{PA} h_{PA} + \dot{M}_G h_G + \dot{M}_{BA} h_{BA} - \dot{M}_E h_E + \dot{M}_G (-\Delta H_c)_G + \sum \dot{M}_{VOC_i} (-\Delta H_c)_{VOC_i} X_i - q_L \quad (86.11)$$

where

- $\dot{M}$  = mass flow rates, kg/min or lbm/min
- $h$  = specific enthalpy, kJ/kg or Btu/lbm
- $-\Delta H_c$  = net heat of combustion (lower heating value), kJ/kg or Btu/lbm
- $X_i$  = fractional conversion of  $\text{VOC}_i$
- $q_L$  = rate of heat loss from the incinerator, kJ/min or Btu/min

In most designs the heat generated by combustion of the pollutant is ignored, but at concentrations in the 1000 ppm range, this heat may be roughly 10% of the burner heat requirement. Insurance regulations limit the combustible pollutant concentration entering the incinerator to 25% of the **lower explosion limit (LEL)**.

To reduce the required operating temperature in the reaction chamber, the combustion mixture may be passed over a catalyst bed, which greatly increases the pollutant oxidation rate. Typically, the catalyst is either palladium or platinum, although Cr, Mn, Co, and Ni are used. The catalyst is deposited on an alumina support or wire screen which is placed just downstream of the burner mix chamber. The total amount of catalyst surface area required is in the range of 0.2 to 0.5 ft<sup>2</sup> per SCFM of waste gas. Typical operating temperature for catalytic incinerators is in the range of 600 to 900°F.

Combustion kinetics are extremely complicated and often incinerator design is based on empirical data or the results of pilot scale tests. The application of simplified kinetic reaction models to incinerator design is described by Cooper and Alley [1994].

The choice between a thermal or catalytic incinerator is in most cases a trade-off between capital and operating cost. The installed cost of a catalytic unit will normally be 40 to 70% higher than a thermal unit, but will require 30 to 50% less fuel for the same waste gas throughput. Offsetting this saving, however, is the service life of the catalyst, which is usually four to six years. Fuel costs for both units may be lowered by incorporating heat recovery equipment in the overall design. The cost of a packaged thermal unit handling up to 30 000 SCFM with facilities for 50% heat recovery may be estimated with the following equation [Vatavuk, 1990]:

$$P = \$4920Q^{0.389} \quad (86.12)$$

where

- $P$  = manufacturer's f.o.b. price, 1988 dollars
- $Q$  = waste gas flow rate, SCFM

The cost of catalytic units (not including the catalyst) equipped for 50% heat recovery may be found from [Vatavuk, 1990]

$$P = \exp(11.7 + 0.0354Q) \quad (86.13)$$

where

- $P$  = manufacturer's f.o.b. price, 1988 dollars

$Q$  = waste gas flow rate, thousands of SCFM

Catalyst costs in 1988 were \$3000 per cubic foot for precious metals and \$600 per cubic foot for common metals.

Both catalytic and thermal incinerators find widespread use in the chemical process industries, in coating and film printing operations, and in food processing. Recent VOC emission regulations should promote increased sales of these units.



This photograph, taken in 1994, clearly illustrates the consequences of acid rain and pollution in the Shenandoah forests of Virginia. It is not possible to tell exactly where the pollution originated because air streams can carry weather fronts hundreds of miles before dumping the precipitation. It is, however, more common to find this type of damage in forests near industrial areas. Although the effect of acid rain is easily visible in forests, acid rain is less visible but extremely detrimental to agricultural crops and human health.

The entire eastern seaboard in the U.S. is experiencing acid rain to some degree. Effects of acid rain are not isolated, but can be seen all over the world. Europe, particularly the vast area of southern Scandinavia, eastern England, Northern France, Germany, and sweeping east through the former Eastern Block countries, have also been greatly affected.

Due to regulation in the U.S., this damaging trend has been slowly reversing over the past 20 years. Discussion in Congress of deregulation threatens the improvements already made. (Photo courtesy of Noam Lior, University of Pennsylvania.)

## Defining Terms

**Aerodynamic diameter:** The diameter of a unit density sphere ( $\rho_p = \rho_w = 1000 \text{ kg/m}^3$ ) that has the same settling velocity as the particle in question.

**Contacting power:** The quantity of energy dissipated per unit volume of gas treated.

**Lower explosion limit:** The lowest concentration of a flammable gas, vapor, or solid in air that will provide flame propagation throughout the mixture.

**Photochemical smog:** An air pollution condition resulting from a series of photochemical reactions involving various volatile organic compounds and oxides of nitrogen in the lower atmosphere.

## References

Cooper, C. D. and Alley, F. C. 1994. *Air Pollution Control<sup>3</sup>4A Design Approach*, 2nd ed. Waveland Press, Prospect Heights, IL.

McCabe, W. L., Smith, J. C., and Harriott, P. 1985. *Unit Operations of Chemical Engineering*, 4th ed. McGraw-Hill, New York.

Peters, M. S. and Timmerhaus, K. D. 1991. *Plant Design and Economics for Chemical Engineers*, 4th ed. McGraw-Hill, New York.

Vatavuk, W. M. 1990. *Estimating Costs of Air Pollution Control*. Lewis, Chelsea, MD.

## Further Information

*Journal of the International Air and Waste Management Association*. Published monthly by A&WMA, Pittsburgh, PA.

*Air Pollution Engineering Manual*, Anthony J. Buonicore and Wayne T. Davis, eds., Van Nostrand Reinhold, New York, 1993.

Peavy, H. S. "Wastewater Treatment and Disposal"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Wastewater Treatment and Disposal

---

- 87.1 Wastewater Characteristics
- 87.2 Terminology in Wastewater Treatment
- 87.3 Sludge
- 87.4 Advanced Wastewater Treatment
  - Nutrient Removal • Solids Removal
- 87.5 Wastewater Disposal and Reuse
- 87.6 Future of Wastewater Treatment and Reuse

## Howard S. Peavy

*University of Idaho*

The practice of collecting and treating wastewater prior to disposal is a relatively recent undertaking. Although remains of sewers have been found in ancient cities, the extent of their use for wastewater carriage is not known. The elaborate drainage system of ancient Rome was not used for waste disposal, and wastes were specifically excluded from the sewerage systems of London, Paris, and Boston until well after the turn of the nineteenth century.

The invention of the flush toilet in the nineteenth century drastically changed waste disposal practices. To cope with the drastically increased volume of liquid waste, cities began to use natural drainage systems and storm sewers for wastewater carriage. The first "modern" sewerage system for wastewater carriage was built in Hamburg, Germany, in 1842 and included many of the principles that are still in use today. Construction of combined sewers was commonplace in large cities during the latter half of the nineteenth century. Since storm drain systems naturally ended at watercourses, waterborne wastes were discharged directly to streams, lakes, and estuaries without treatment. Gross pollution often resulted, and health problems were transferred from the sewered community to downstream users of the water.

The treatment of wastewater lagged considerably behind its collection. Treatment was considered necessary only after the self-purification capacity of the receiving waters was exceeded and nuisance conditions became intolerable. Various treatment processes were tried in the late 1800s and early 1900s, and by the 1920s wastewater treatment had evolved to those processes in common use today. Design of wastewater treatment facilities remained empirical, however, until mid-century. In the last 30 to 40 years, great advances have been made in understanding wastewater treatment, and the original processes have been formulated and quantified. The science of wastewater treatment is far from static, however. Advanced wastewater treatment processes are currently being developed that will produce potable water from domestic wastewater. Problems

associated with wastewater reuse will no doubt challenge the imagination of engineers for many years to come.

Philosophies concerning the treatment and ultimate disposal of wastewater have evolved over the years. Operating under the original assumption that the "solution to pollution is dilution," the assimilative capacity of streams was utilized before treatment was deemed necessary. For many years, little, if any, treatment was required of small communities located on large streams, whereas a high level of treatment was required by large cities discharging to small streams. In more recent times the policy has shifted to require a minimum level of treatment of all waste discharges, regardless of the capacity of the receiving stream. Under current practice in the U.S., all dischargers are given a permit stating the maximum amount of each pollutant that they are allowed to discharge. Discharge permits are no longer intended to just protect the self-purification capacity of the streams but are focused on maintaining a high level of quality in all receiving streams.

Where extensive treatment of wastewater is necessary to meet stringent discharge permits, the quality of the treated effluent may approach or exceed that of the receiving stream. This effluent should be considered a valuable water resource, particularly where water is scarce. Regulatory agencies encourage utilization of these wastewaters for irrigation, non-body contact recreational activities, groundwater recharge, some industrial processes, and other nonpotable uses.

## 87.1 Wastewater Characteristics

Wastewaters are usually classified as industrial wastewater or municipal wastewater. Industrial wastewater with characteristics compatible with municipal wastewater may be discharged to the municipal sewers. Many industrial wastewaters require pretreatment to remove noncompatible substances prior to discharge into the municipal system.

Water collected in municipal wastewater systems, having been put to a wide variety of uses, contains a wide variety of contaminants. Quantitatively, constituents of wastewater may vary significantly, depending upon the percentage and type of industrial waste present and the amount of dilution from infiltration/inflow into the collection system. The most significant components of wastewater are usually suspended solids, biodegradable organics, and pathogens. The source and environmental significance of contaminants commonly found in wastewater are shown in [Table 87.1](#).

**Table 87.1** Wastewater Contaminants

Contaminant	Source	Environmental Significance
Suspended solids	Domestic use, industrial wastes, erosion by infiltration/inflow	Cause sludge deposits and anaerobic conditions in aquatic environment
Biodegradable organics	Domestic and industrial waste	Cause biological degradation, which may use up oxygen in receiving water and result in undesirable conditions



Pathogens	Domestic waste	Transmit communicable diseases
Nutrients	Domestic and industrial waste	May cause eutrophication
Refractory organics	Industrial waste	May cause taste and odor problems; may be toxic or carcinogenic
Heavy metals	Industrial waste, mining, etc.	Are toxic; may interfere with effluent reuse
Dissolved inorganic solids	Increases above level in water supply by domestic and/or industrial use	May interfere with effluent reuse

---

Source: Peavy, H. S., Rowe, D. R., and Tchobanoglous, G. 1985. *Environmental Engineering*. McGraw-Hill, New York.

**Suspended solids** are primarily organic in nature and are composed of some of the more objectionable material in sewage. Body wastes, food waste, paper, rags, and biological cells form the bulk of suspended solids in wastewater. Even inert materials such as soil particles become fouled by adsorbing organics to their surface. Removal of suspended solids is essential prior to discharge or reuse of wastewater.

Although suspended organic solids are biodegradable through hydrolysis, biodegradable material in wastewater is usually considered to be **dissolved solids** of organic origin. Soluble organics in domestic wastewater are composed chiefly of proteins (40 to 60%), carbohydrates (25 to 50%), and lipids (approximately 10%). Proteins are chiefly amino acids, whereas carbohydrates are compounds such as sugars, starches, and cellulose. Lipids include fats, oil, and grease. All of these materials contain carbon that can be converted to carbon dioxide biologically, thus exerting a **biochemical oxygen demand**. Proteins also contain nitrogen, and thus a nitrogenous oxygen demand is also exerted. The biochemical oxygen demand (BOD) test is therefore used to quantify biodegradable organics.

All forms of waterborne **pathogens** may be found in domestic wastewater. These include bacteria, viruses, protozoa, and helminths. These organisms are discharged by persons who are infected with the disease. Although pathogens causing some of the more exotic diseases may not be present in viable form, it is a safe assumption that a sufficient number of pathogens are present in all untreated wastewater to represent a substantial health hazard. Fortunately, few of the pathogens survive wastewater treatment in a viable state.

Traditional wastewater treatment processes are designed to reduce suspended solids, biodegradable organics, and pathogens to acceptable levels prior to disposal. Additional wastewater treatment processes may be required to reduce levels of nutrients if the wastewater is to be discharged to a delicate ecosystem. Processes to remove refractory organics and heavy metals and to reduce the level of inorganic dissolved solids may be required where wastewater reuse is anticipated.

## 87.2 Terminology in Wastewater Treatment

---

The terminology used in wastewater treatment is often confusing to the uninitiated person. Terms such as *unit operations*; *unit processes*; *reactors*; *systems*; and *primary*, *secondary*, and *tertiary treatment* frequently appear in the literature, and their usage is not always consistent. The

meanings of these terms as used in this chapter are discussed in the following paragraphs.

Methods used for treating municipal wastewaters are often referred to as either **unit operations** or **unit processes**. Generally, unit operations involve contaminant removal by physical forces (screening, sedimentation, and filtration), whereas unit processes are conversion processes involving biological and/or chemical reactions.

The term *reactor* refers to the vessel, or containment structure, along with all of its appurtenances, in which the unit operation or unit process takes place. Although unit operations and processes are natural phenomena, they may be initiated, enhanced, or otherwise controlled by altering the environment in the reactor. Reactor design is a very important aspect of wastewater treatment and requires a thorough understanding of the unit processes and unit operations involved.

A wastewater treatment system is composed of a combination of unit operations and unit processes designed to reduce certain constituents of wastewater to an acceptable level. Many different combinations are possible. Although practically all wastewater treatment systems are unique in some respects, a general grouping of unit operations and unit processes according to target contaminants has evolved over the years. Unit operations and processes commonly used in wastewater treatment are listed in [Table 87.2](#) and are arranged according to conventional grouping.

**Table 87.2** Unit Operations, Unit Processes, and Systems for Wastewater Treatment

Contaminant	Unit Operation, Unit Process, or Treatment System
Suspended solids	Sedimentation
	Screening and comminution
	Filtration variations
	Flotation
	Chemical-polymer addition
	Coagulation/sedimentation
	Land treatment systems
Biodegradable organics	Activated-sludge variations
	Fixed-film; trickling filters
	Fixed-film; rotating biological contactors
	Lagoon and oxidation pond variations
	Intermittent sand filtration
	Land treatment systems
	Physical-chemical systems
Pathogens	Chlorination
	Hypochlorination
	Ozonation
	Land treatment systems
Nutrients	
Nitrogen	Suspended-growth nitrification and denitrification variations
	Fixed-film nitrification and denitrification variations

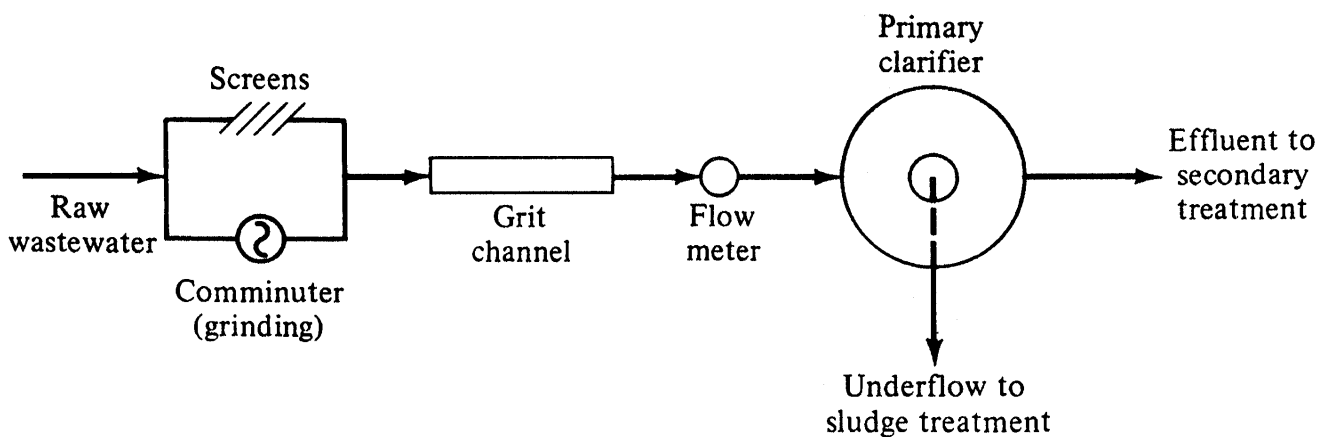
	Ammonia stripping
	Ion exchange
	Breakpoint chlorination
	Land treatment systems
Phosphorus	Metal-salt addition
	Lime coagulation/sedimentation
	Biological-chemical phosphorus removal
	Land treatment systems
Refractory organics	Carbon adsorption
	Tertiary ozonation
	Land treatment systems
Heavy metals	Chemical precipitation
	Ion exchange
	Land treatment systems
Dissolved inorganic solids	Ion exchange
	Reverse osmosis
	Electrodialysis

---

Source: Peavy, H. S., Rowe, D. R., and Tchobanoglous, G. 1985. *Environmental Engineering*. McGraw-Hill, New York.

Municipal wastewater treatment systems are often divided into primary, secondary, and tertiary subsystems. The purpose of **primary treatment** is to remove suspended solid materials from the incoming wastewater. Large debris may be removed by screens or may be reduced in size by grinding devices. Inorganic solids are removed in grit channels, and much of the organic suspended solids is removed by sedimentation. A typical primary treatment system (Fig. 87.1) should remove approximately one half of the suspended solids in the incoming wastewater. The BOD associated with these solids accounts for about 30% of the influent BOD.

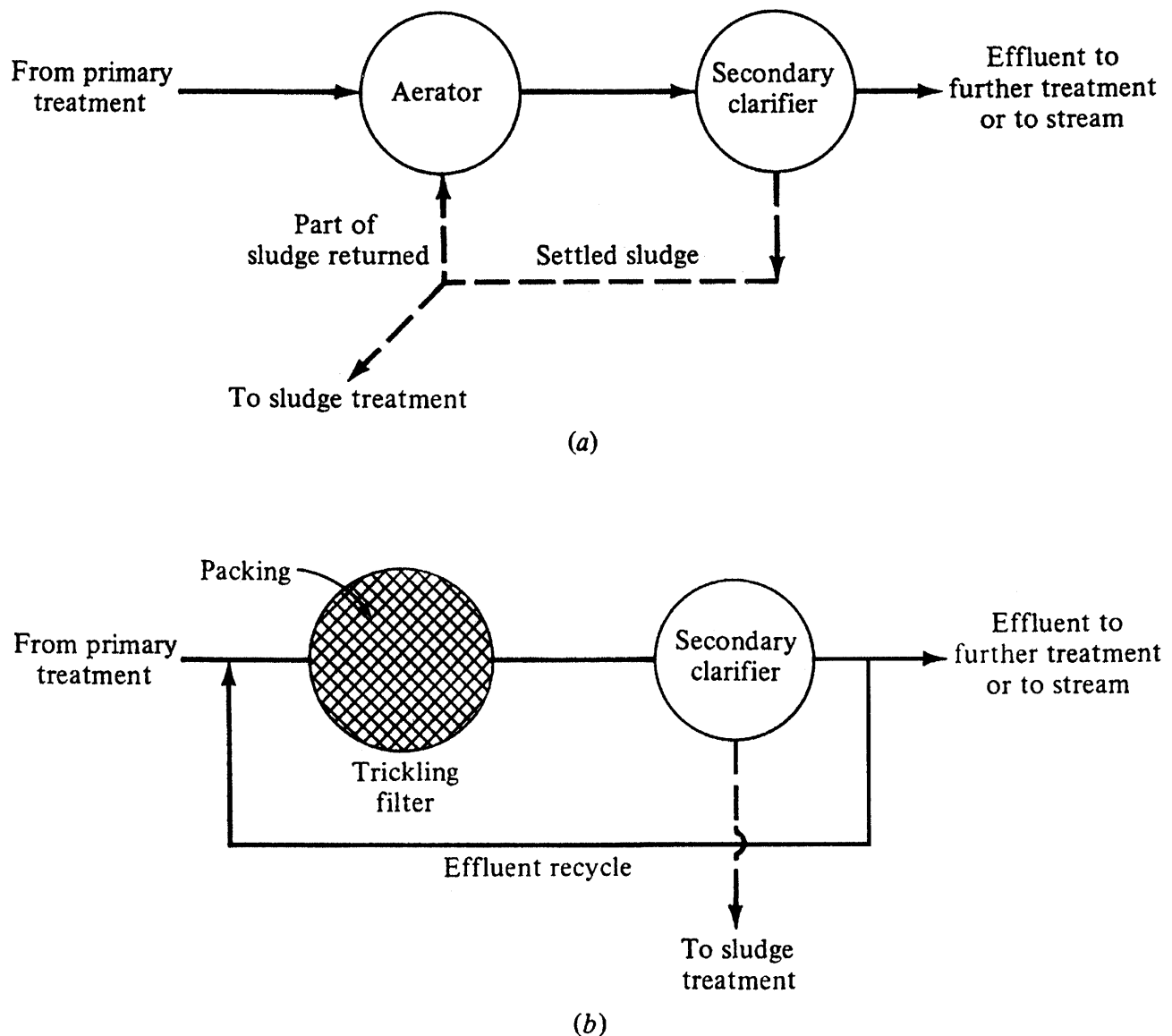
**Figure 87.1** Typical primary treatment system. (Source: Peavy, H. S., Rowe, D. R., and Tchobanoglous, G. 1985. *Environmental Engineering*. McGraw-Hill, New York.)



**Secondary treatment** usually consists of biological conversion of dissolved and colloidal

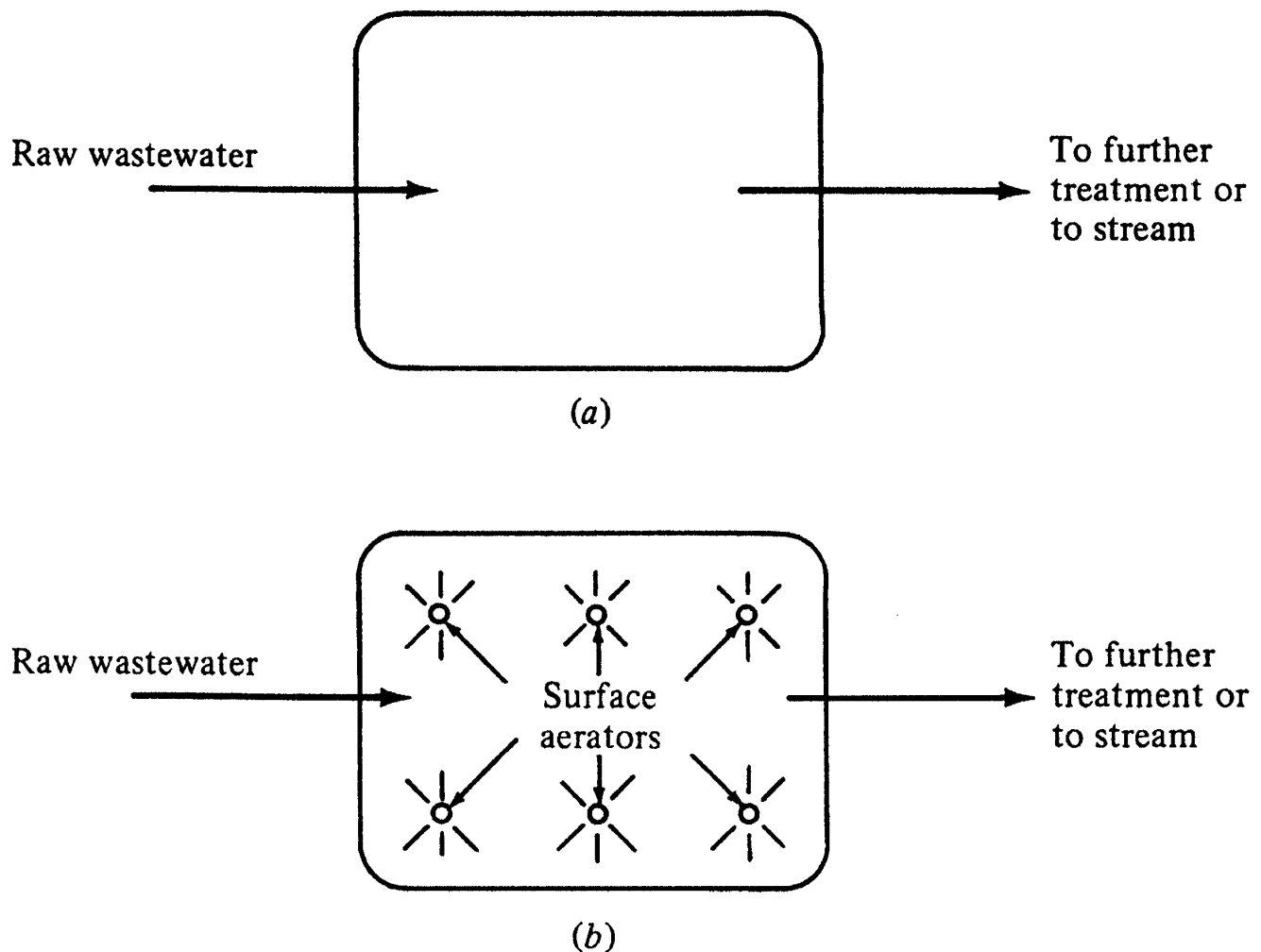
organics into biomass that can subsequently be removed by sedimentation. Contact between microorganisms and the organics is optimized by suspending the biomass in the wastewater or by passing the wastewater over a film of biomass attached to solid surfaces. The most common suspended biomass system is the activated-sludge process shown in Fig. 87.2(a). Recirculating a portion of the biomass maintains a large number of organisms in contact with the wastewater and speeds up the conversion process. The classical attached-biomass system is the trickling filter shown in Fig. 87.2(b). Stones or other solid media are used to increase the surface area for **biofilm** growth. Mature biofilms peel off the surface and are washed out to the settling basin with the liquid underflow. Part of the liquid effluent may be recycled through the system for additional treatment and to maintain optimal hydraulic flow rates.

**Figure 87.2** Typical secondary treatment systems: (a) activated sludge system and (b) trickling filter system. (Source: Peavy, H. S., Rowe, D. R., and Tchobanoglous, G. 1985. *Environmental Engineering*. McGraw-Hill, New York.)



Sometimes primary and secondary treatment can be accomplished together, as shown in Fig. 87.3. The oxidation pond [Fig. 87.3(a)] is a large impoundment that most nearly approximates treatment by natural systems, with oxygen being supplied by algal photosynthesis and surface reaeration. This oxygen seldom penetrates to the bottom of the pond, and the solids that settle are decomposed by anaerobic bacteria (bacteria that require no oxygen). In an aerated lagoon system [Fig. 87.3(b)] oxygen is supplied by mechanical aeration, and the entire depth of the pond is aerobic. The small quantity of excess sludge that is produced is retained in the bottom sediments. Due to large space requirements, oxidation ponds are only feasible for small communities.

**Figure 87.3** Primary and secondary treatment by ponds: (a) oxidation pond and (b) aerated pond.  
(Source: Peavy, H. S., Rowe, D. R., and Tchobanoglous, G. 1985. *Environmental Engineering*. McGraw-Hill, New York.)



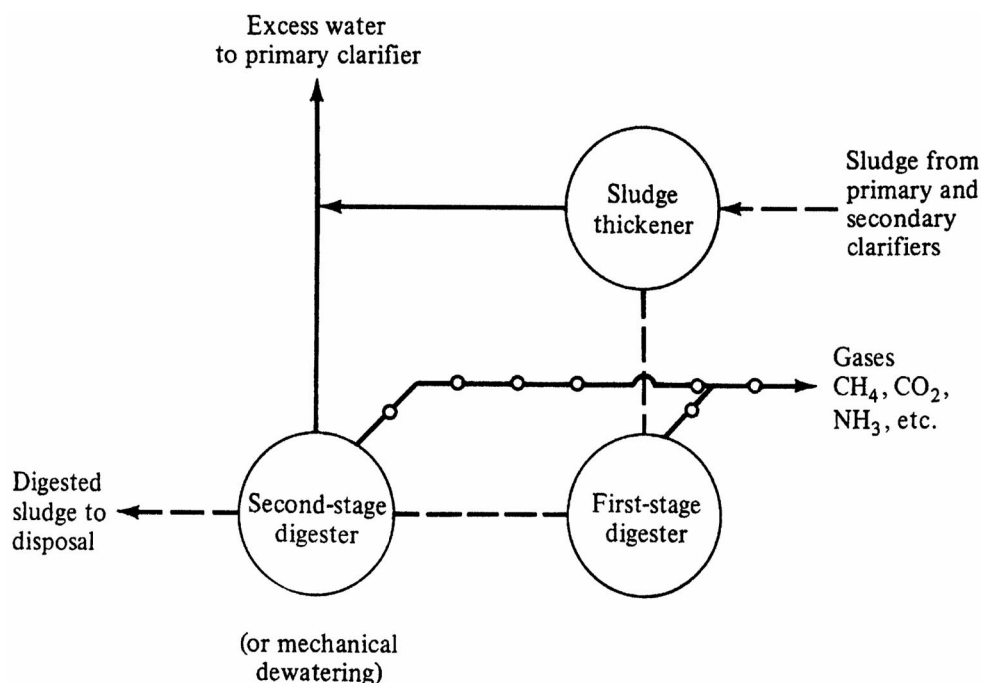
In most cases, secondary treatment of municipal wastewater is sufficient to meet effluent standards. In some instances, however, additional treatment may be required. **Tertiary treatment** most often involves further removal of suspended solids and/or the removal of nutrients. Solids removal may be accomplished by filtration, and phosphorus and nitrogen compounds may be removed by combinations of physical, chemical, and biological processes.

## 87.3 Sludge

A careful inspection of Figs. 87.1 through 87.3 leads to an interesting observation. The "removal" processes in wastewater treatment are essentially concentrating, or thickening, processes. Suspended solids are removed as sludges, and dissolved solids are converted to suspended solids and subsequently become removable sludges. Primary and secondary treatment, followed by sludge thickening, may concentrate the per capita contribution of organic material (represented by 250 mg/L of suspended solids and 200 mg/L BOD in 375 L of municipal wastewater) to 2.0 L of sludge containing 50 000 mg/L of solids. Most of the objectionable material initially in the wastewater is concentrated in the sludges and must be disposed of in a safe and environmentally acceptable manner. The majority of the expenses, effort, and problems of wastewater treatment and disposal are associated with the sludges.

Wastewater sludges are usually combined, thickened, and treated by two-stage anaerobic biological processes, as shown in Fig. 87.4. The results are gaseous end products, principally methane ( $\text{CH}_4$ ) and carbon dioxide ( $\text{CO}_2$ ), and liquids and inert solids. The methane has significant heating value and may be used to meet part of the power requirements of the treatment plant. The liquids contain large concentrations of organic compounds and are recycled through the treatment plant. The solid residue has a high mineral content and may be used as a soil conditioner and fertilizer on agricultural lands. Other means of solids disposal may be by incineration or by landfilling.

**Figure 87.4** Sludge thickening and treatment system. (Source: Peavy, H. S., Rowe, D. R., and Tchobanoglous, G. 1985. *Environmental Engineering*. McGraw-Hill, New York.)



## 87.4 Advanced Wastewater Treatment

---

The quality of effluent provided by secondary treatment may not always be sufficient to meet discharge requirements. This is often the case when large quantities of effluent are discharged into small streams or when delicate ecosystems are encountered. In these instances, additional treatment to polish the effluent from secondary systems will be required, or an alternative method of wastewater disposal must be found.

Additional treatment often involves the removal of nitrogen and phosphorus compounds, plant nutrients associated with eutrophication. Further treatment may be required to remove additional suspended solids, dissolved inorganic salts, and **refractory organics**. Combinations of the above processes can be used to restore wastewater to potable quality, although at considerable expense. Referred to as *reclamation*, this complete treatment of wastewater can seldom be justified except in water-scarce areas where some form of reuse is mandated.

The term **advanced wastewater treatment** is frequently used to encompass any or all of the above treatment techniques, and this term would seem to imply that advanced treatment follows conventional secondary treatment. This is not always the case, as some unit operations or unit processes in secondary or even primary treatment may be replaced by advanced treatment systems. The targets of most advanced wastewater treatment processes and operations are dissolved nutrients and suspended solids.

### Nutrient Removal

Excess nutrients can be very troublesome in natural water systems because they stimulate growth of algae and other aquatic plants. Although the quantities of nutrients contributed by wastewater discharges may be less than those contributed by agricultural runoff and other sources, the point-source nature of wastewater discharges makes them more amenable to control techniques. Thus, wastewater treatment plants that discharge to water bodies that are delicately balanced with respect to nutrient loads may have nutrient limitations imposed on their effluent. The nutrients most often of interest are nitrogen and phosphorous compounds.

### Solids Removal

Removal of suspended solids, and sometimes dissolved solids, may be necessary in advanced wastewater treatment systems. The solids removal processes employed in advanced wastewater treatment are essentially the same as those used in the treatment of potable water, although application is made more difficult by the overall poorer quality of the wastewater.

As an advanced treatment process, suspended solids removal entails the removal of particles and flocs too small or too lightweight to be removed in gravity settling operations. These solids may be carried over from the secondary clarifier or from tertiary systems in which solids were precipitated.

Several methods are available for removing residual suspended solids from wastewater. Removal by centrifugation, air flotation, mechanical microscreening, and granular-media filtration have all been used successfully. In current practice, granular-media filtration is the most commonly used process. Basically, the same principles that apply to filtration of particles from potable water

apply to the removal of residual solids in wastewater. However, differences in operational modes for wastewater filtration versus potable water filtration may range from slight to drastic.

Although applied less frequently than suspended solids removal processes, dissolved solids removal is sometimes necessary in the treatment of wastewaters. Both secondary treatment and nutrient removal decrease the content of dissolved organic solids in wastewater. Neither process, however, completely removes all dissolved organic constituents, and neither process removes significant amounts of inorganic dissolved solids. Further treatment will be required where substantial reductions in the total dissolved solids of wastewater must be made.

Ion exchange, microporous membrane filtration, adsorption, and chemical oxidation can be used to decrease the dissolved solids content of water. These processes were developed to prepare potable water from a poor-quality raw water. Their use can be adapted to advanced wastewater treatment if a high level of pretreatment is provided. The removal of suspended solids is necessary prior to any of these processes. Removal of the dissolved organic material (by activated carbon adsorption) is necessary prior to microporous membrane filtration to prevent the larger organic molecules from plugging the micropores.

Advanced wastewater treatment for dissolved solids removal is complicated and expensive. Treatment of municipal wastewater by these processes can be justified only when reuse of the wastewater is anticipated.

## **87.5 Wastewater Disposal and Reuse**

---

An insignificant volume of the influent wastewater accompanies sludges and other materials disposed of during wastewater treatment processes. The bulk of the wastewater remains to be disposed of after the treatment processes have been completed. Ultimate receptors of treated wastewaters include surface water and groundwater bodies, land surfaces, and, in some instances, the atmosphere. Recognition of the value of wastewater as a water resource has resulted in an increase in the reuse of treated effluent, particularly in water-scarce regions.

The most common method of wastewater disposal is by dilution in surface waters. As previously discussed, treatment processes are applied to make the wastewater compatible with the natural water system to which it is discharged. Other practices include land disposal and, in arid climates, evaporation to the atmosphere. Land application may involve irrigation, in which the water is incorporated in plants and/or transpired to the atmosphere, or infiltration, in which the water is percolated through the soil and is ultimately discharged to groundwater. Land disposal practices are generally seasonal in nature and generally require large storage facilities for off-season flows.

Under some circumstances, wastewater reuse may be an appropriate method of disposing of treated effluent. In water-scarce areas, wastewater may constitute a major portion of the available resource. Where delicate ecosystems necessitate stringent effluent requirements, reuse of the wastewater may help to offset the cost of advanced wastewater treatment, or a reuse that will accept a lower level of treatment may obviate the expense of advanced treatment prior to



discharge. Wastewater is currently being reused for several purposes, including creation or enhancement of recreational facilities, industrial water supplies, groundwater recharge, and even direct reuse in potable supplies.

## 87.6 Future of Wastewater Treatment and Reuse

---

Many areas of the world are presently experiencing water shortages, or expect to experience them in the foreseeable future. In these areas, wastewaters must be considered a valuable resource and be integrated into the available water supply.

The principal concerns involving the reuse of wastewater are public health and public acceptance. From a public health standpoint, there must be reasonable assurance that pathogenic microorganisms and toxic chemicals are reduced to concentrations that pose acceptable health risks. Thus, wastewater treatment processes and operations must be upgraded and/or developed to consistently treat wastewater to a higher level of purity, and at an affordable cost. Professionals in the wastewater industry, in engineering practice, and in academia are currently active in this pursuit. The development of new technology, new materials, and better analytical techniques to measure results creates challenges and opportunities to which professionals in the wastewater field will be responding well into the foreseeable future.

In addition to technological advances, public acceptance is also a necessary element of wastewater reuse. Experience at the Santee project in California—in which treated wastewater was used to enhance a recreational resource—indicates that public acceptance is greatly enhanced by informing and involving the public at all stages of planning and implementation of wastewater reuse. Following this lead, the city of Denver has launched a massive drive for public acceptance of wastewater recycling. Nonpotable reuse of effluent from a demonstration plant is planned, with extensive research on health and toxicological studies being performed. Concurrently, a public education program has been designed to gain public acceptance of eventual reuse in the potable system, should the health studies show this to be practical. These programs are to continue for 10 to 15 years and, if successful, will result in the construction of a full-scale plant from which reuse will include direct recycle to the potable system. The city of San Diego, California, is presently embarking on a similar project.

Other projects have chosen to limit the utilization of reclaimed wastewater to nonpotable uses for the present time. In California, Los Angeles and Orange counties conducted an extensive market analysis and identified a long-term nonpotable reuse potential of  $1.15 \cdot 10^6 \text{ m}^3/\text{d}$  of treated wastewater. These uses include irrigation of public property, various industrial uses, and groundwater recharge. The uses require varying levels of treatment and, thus, varying costs depending on the water quality acceptable to each user. At the project's initiation, a combination of users that optimized the cost of treatment and delivery of the wastewater was selected. This type of approach has considerable merit when the demand for treated wastewater exceeds the supply.

As demand for water increases, more consideration will necessarily be given to fitting the quality of the water to the intended use. Currently, water distributed through public systems is of potable quality, although less than one half the water distributed through these systems is used in a manner necessitating potable water quality. Thus, the opportunity for the use of water of less-than-potable

quality is abundant, and reclaimed wastewater could conceivably be used in many instances where potable water is now being used. This would require separate delivery and plumbing systems, an expensive but not technologically complicated procedure, along with an extensive public education program.

## Defining Terms

**Advanced wastewater treatment:** Processes or operations used to "polish" wastewater after conventional treatment. Generally aimed at specific contaminants.

**Biochemical oxygen demand:** The amount of oxygen that is required for microorganisms to convert the organic material in wastewater to carbon dioxide and other stable end products.

**Biofilm:** A community of microorganisms that attach themselves to surfaces in thick, slimy layers.

**Dissolved solids:** Ions of elements and or compounds that are held by or within the molecular structure of water. The residue that is left when a filtered sample of water is evaporated to dryness.

**Pathogen:** Microorganisms that are capable of causing disease in humans.

**Primary treatment:** The initial set of wastewater treatment operations that are intended to remove or alter suspended solids from wastewater.

**Refractory organics:** Organic material that is not readily biodegradable due to its molecular structure or toxicity to microorganisms.

**Secondary treatment:** A set of processes and operations whose purpose is to convert dissolved organics to biological solids and to remove those solids from the wastewater stream. Usually follows primary treatment.

**Suspended solids:** Material that is held in suspension by the buoyant or turbulent forces of water. Can be removed by filtering through a standard filter paper.

**Tertiary treatment:** Removal of nutrients—usually compounds of phosphorous and nitrogen—from wastewater. Usually follows secondary treatment.

**Unit operations:** The removal of substances from wastewater by physical forces, for example, settling, filtration, and so on.

**Unit processes:** The conversion of dissolved substances to suspended material so that it can be more readily removed from wastewater.

## References

- Hadeed, S. J. 1972. Potable water from wastewater—Denver's program. *J. WPCF*. 49(8):1757.
- Metcalf & Eddie, Inc. 1979. *Wastewater Engineering: Treatment, Disposal, Reuse*, 2nd ed. McGraw-Hill, New York.
- Peavy, H. S., Rowe, D. R., and Tchobanoglous, G. 1985. *Environmental Engineering*. McGraw-Hill, New York.

## Further Information

Information in this chapter was excerpted from chapter 5, Engineered systems for wastewater treatment and disposal, in Peavy, H. S., Tchobanoglous, G., and Rowe, D. R. 1985.

*Environmental Engineering*. McGraw-Hill, New York. The reader seeking more information on the subject of wastewater treatment is referred to that book and to the following texts and journals.

A good introductory-level coverage of wastewater treatment is presented by Vesilind, Peirce, and Weiner in *Environmental Engineering*, 3rd ed. (Butterworth-Heinemann, Newton, MA, 1994).

A thorough, rigorous, and complete coverage of wastewater treatment is presented in *Wastewater Engineering*, 3rd ed. (Metcalf & Eddie, McGraw-Hill, New York, 1991).

*Water Environment & Technology*, a monthly journal published by the Water Environment Federation, contains articles on the practice and politics of wastewater treatment. A more scholarly journal, *Water Environment Research*, is published every other month by the same organization.

The *Journal of Environmental Engineering*, published every other month by the American Society of Civil Engineers, contains articles of a more technical nature on wastewater treatment.

McKinney, R. E. "Solid Wastes"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

- 88.1 Regulations
- 88.2 Characteristics
- 88.3 Generation
- 88.4 Collection
- 88.5 Transfer and Transport
- 88.6 Processing and Resource Recovery (Recycling)
- 88.7 Final Disposal

**Ross E. McKinney***Consulting Engineer*

**Solid wastes** are primarily a problem for urban and industrial areas. Although solid wastes have been a problem for urban centers since the beginning of time, it was the rapid growth of the industrial society that generated the current problems. As the standard of living increased, the consumption of goods produced solid wastes that created environmental pollution when discarded without proper treatment. Recently, the industrialized countries began to examine solid wastes and to develop methods for reducing their production as well as processing them for return to the environment without creating pollution. Although progress has been made, more is required.

---

**88.1 Regulations**

---

The primary solid waste regulations are at the local government level. One of the primary functions of local government is to protect the health of its citizens. Solid waste collection and processing is primarily a public health problem that can only be managed at the local level. Most communities have developed ordinances to handle solid waste storage, collection, and processing.

Local government derives its authority from state government. Since solid wastes are public health problems, many states have placed their solid waste regulations under state departments of health. In recent years a number of states have moved control of solid wastes to departments of environmental pollution control or departments of natural resources. The authority for all solid waste regulations resides in each state legislature, with the development of detailed regulations assigned to the appropriate state agency.

In 1965 the federal government became involved in solid waste legislation when Congress passed the first federal solid waste legislation as an addendum to the 1965 Federal Air Pollution

legislation. The primary purpose of this federal legislation was to reduce the number of open burning dumps that were creating air pollution and to determine the status of state solid waste programs. The Resource Recovery Act was passed by Congress in 1970 to assist in the development of new systems for processing solid wastes and recovering materials for reuse. This was followed by the Resource Conservation and Recovery Act of 1976 (RCRA). The most important components of this legislation were Subtitle C, which dealt with hazardous solid wastes, and Subtitle D, which dealt with sanitary landfills. Problems with buried hazardous wastes resulted in Congress's passing the Comprehensive Environmental Response, Compensation and Liability Act of 1980 (CERCLA), also known as the Superfund Act. RCRA was amended by Congress in 1980 and in 1984. The 1984 amendments were designated the Hazardous and Solid Waste Amendments. In 1986 the Superfund Amendments and Reauthorization Act (SARA) passed Congress. Congress has had recent difficulty passing new legislation, but it has allowed existing legislation to continue, permitting the federal EPA to develop regulations covering the areas of concern. As a net result, it is essential to keep informed on the continuous stream of regulations published in the *Federal Register*.

Overall, federal regulations provide national policies in solid wastes, whereas the state regulations are designed to implement the federal regulations. However, it must be recognized that ultimate control of solid waste operations is at the local level. It is at the local level that federal and state policies and regulations are converted to action. For the most part, regulations are being developed in response to public perception of environmental problems and the two are not always in a logical sequence.

## 88.2 Characteristics

---

One of the more interesting aspects of solid waste management is the simple fact that very little real data exists on current solid waste characteristics. It is too difficult to collect detailed data on solid waste characteristics for even the largest cities to gather data. A few detailed studies have been made when the economic considerations for specific solid waste-processing systems were high enough to justify determination of solid waste characteristics. Most engineers use the data generated by the U.S. EPA [EPA, 1990] as the basis for solid waste characteristics. The EPA data represent approximate characteristics based on a material flow methodology developed at Midwest Research Institute in 1969. The EPA composition data are a reasonable estimate on a national basis. It should be recognized that solid waste characteristics vary with the section of the country, the seasons of the year, and economic conditions.

The EPA Office of Solid Waste recognized that solid wastes were composed of many different materials, making composition analyses difficult if a common set of criteria were not used. As a result, a broad classification of ten categories was used. Although several subcategories have been added for specific projects, the ten categories have gained widespread acceptance. Table 88.1 gives the percentage of total solid wastes for the ten categories as estimated by the EPA for the year 1988 [EPA, 1990]. It was estimated that 83% of the solid wastes was combustible and 17% noncombustible. The combustible characteristics of solid wastes are important when solid wastes are to be processed by incineration, whereas the general solid waste characteristics are used for evaluating solid waste recycling and solid waste reduction projects. It should be noted that the EPA

solid waste characteristics are based on an "as collected" basis, which includes about 20% moisture and 5% ash from the combustible organics. Only about 58% of the solid wastes are combustible organics, with 22% noncombustible ash. Variations in moisture content are very important when evaluating solid waste characteristics for incineration.

**Table 88.1** Characteristics of Solid Wastes Generated in 1988

Type of Waste	Percentage of Total Solid Waste
Paper and paperboard	40.0
Glass	7.0
Metals	8.5
Plastics	8.0
Rubber and leather	2.6
Textiles	2.2
Wood	3.6
Food wastes	7.3
Yard wastes	17.6
Other wastes	3.2

*Source:* Environmental Protection Agency, 1990. *Characterization of Municipal Solid Waste in the United States: 1990 Update*, p. ES-7. EPA/530-SW-90-042.

## 88.3 Generation

Data on solid waste generation are a little better than data on solid waste characteristics, since more communities have gathered data on the weight of solid wastes handled than on solid waste analyses. However, it should be recognized that weight data may not be much better than characterization data. Accurate weight data require checking the scales at regular intervals to ensure accuracy. It is also important to weigh trucks both before and after unloading to determine the real weight of the solid wastes. Use of a truck tare weight to minimize the number of weighings will often result in high values. The key to proper evaluation of generation data is regular collection and weighing of solid wastes from the same areas of the community.

Generation data are generally reported in pounds per person per day. With a uniform generation rate it is easy to determine the total solid waste production for various sizes of communities. The EPA data on solid waste generation [EPA, 1990] indicated a change from 2.66 lb/person/day in 1960 to 4.00 lb/person/day in 1988. It has been found that solid waste generation is primarily a function of economic conditions. During recessions solid waste generation decreases. With rapid economic growth the production of solid wastes increases. Concern over the generation of solid wastes has resulted in changes in the packaging of materials. Smaller packages and lighter packaging have contributed to the decrease in weight of packaging materials. Increased advertising has generated more pages of newspapers and magazines that quickly become outdated. Careful evaluation of social conditions indicates that there is a limit to how much solid waste a person can actually generate on a daily basis. A study for the Institute for Local Self-Reliance [Platt *et al.*, 1990] found that residential solid waste generation for 15 cities ranged from 1.9 to 5.5

lb/person/day with a median of 3.3 lb/person/day. Unfortunately, the economics of some solid wastes projects depend upon a reasonable estimation of the solid waste generation rate. The ease of collecting solid waste generation data should allow engineers to obtain sufficient data prior to the design of any solid waste project.

## 88.4 Collection

---

Solid wastes must be collected where they are generated. This means that a system must be developed to go to every building in the community at some time frequency. Currently, residential solid wastes can be conveniently collected once weekly. A few communities employ twice weekly collection. Commercial buildings may require more frequent collection, depending on the rate of generation and the size of the storage container. Restaurants often require daily collection.

Every solid waste generator must have a suitable storage container to hold the solid wastes between collection. Residences tend to use individual containers, each having a maximum capacity of 30 gallons to minimize the weight in any single container. Recently, large plastic containers with wheels have been used in some communities. The householder is able to wheel the solid wastes from the garage to the curb for pickup without difficulty. Special mechanisms on the collection trucks allow the solid waste collector to transfer the solid wastes from the wheeled container to the collection truck with a minimum of effort. Apartment buildings, office buildings, and commercial establishments produce larger quantities of solid wastes than residences and require larger solid waste containers. Storage containers range in size from 1 to 10 yd<sup>3</sup>. Small storage containers can be dumped into rear-loading collection trucks or side-loading trucks, whereas large storage containers require front-loading collection trucks.

Solid wastes weigh between 200 and 250 lb/yd<sup>3</sup> in loose containers. Collection trucks use hydraulic compaction to maximize the quantity of solid wastes that can be collected before dumping. Although it is possible to compact solid wastes to 600–700 lb/yd<sup>3</sup>, most collection trucks average 500 lb/yd<sup>3</sup> after compaction. Rear-loading compaction trucks range in size from 9 to 32 yd<sup>3</sup>, with 20 yd<sup>3</sup> trucks being the most commonly used for residential collection. Side-loading compaction trucks range from small, 6 yd<sup>3</sup>, to large, 35 yd<sup>3</sup>. Front-loading compaction trucks range from 22 to 42 yd<sup>3</sup>, with 30 and 35 yd<sup>3</sup> trucks being most widely used for commercial routes. Roll-off units are used by industrial and special commercial accounts having large quantities of nonputrescible solid wastes. Roll-off units range in size from 10 to 30 yd<sup>3</sup> and may be attached to stationary compaction units to increase the capacity of the roll-off unit. When the roll-off unit is full, the collection truck pulls the roll-off unit onto the truck frame for transport to the processing site. The collection trucks all have special unloading mechanisms to remove the compacted solid wastes at the desired processing site.

Collection crew size depends upon the specific collection pattern of each community. One-person crews are used for some side-loading residential collections, roll-off collections, and some front-loading commercial collections. Two-person crews are used in which one person is the driver and the other person is the collector. The two-person crews are quite common in areas where collection can only be made on one side of the street at a time. Three-person crews are most common in residential collection from both sides of the street. Four-person crews have been used



in large communities. The size of collection crews and collection trucks depends on the size of the collection routes and the time allotted for collection activities. Time-motion studies are very important for evaluating the crew collection efficiency, whereas good operation and maintenance records are important for evaluating the drivers and determining equipment replacement.

## 88.5 Transfer and Transport

---

Transport from the collection route to the solid waste processing site is an important parameter affecting the cost of operations. It is part of the rest time for collectors. If the transport distance is too long, valuable collection time can be lost. As a result, transfer stations have been constructed in large cities to permit smaller collection trucks to unload and return to collection routes. The solid wastes in the transfer station are placed into larger transfer trucks, 50 to 70 yd<sup>3</sup> capacity, for transport to the final processing site. The large transfer trucks normally have only a single driver, reducing the labor costs. Transfer stations are economical only in very large cities and require special design for efficient traffic flow. Small transfer stations where individuals bring their solid wastes in their own automobiles or trucks have been used in rural areas.

Recently, large cities have examined transporting solid wastes by rail for final processing. Rail transport is very expensive and has been used only in crowded areas where several communities are contiguous and no processing site can be found in any of the communities. Barge transport has also been used to move large quantities of solid wastes more economically. Transport of solid wastes from one area to other distant areas has created social and political problems that have limited long-distance transport of unprocessed solid wastes.

## 88.6 Processing and Resource Recovery (Recycling)

---

Solid waste processing has moved from incineration and open dumps to burial in sanitary landfills to partial recycling. **Incineration** has been the primary processing system for solid wastes in large cities over the past 100 years. The British pioneered incineration with energy recovery at the end of the last century, but it has been in only the last two decades that many energy recovery incineration systems have been constructed in the U.S. Poor designs and poor operations limited the use of energy recovery incineration even though they were quite successful in Europe. Incineration has application only in the very crowded areas of the country where heat energy can be easily used by adjacent industrial operations. It is important to understand that incineration simply converts combustible solid wastes to gaseous wastes that must be discharged to the atmosphere for final disposal and the noncombustible solid wastes to **ash** that must be returned to the land. Because of the potential for air pollution, the EPA has required extensive air pollution control equipment for solid waste incinerators, increasing the capital and operating costs. The U.S. Supreme Court has indicated that municipal solid waste ash must pass the hazardous waste criteria or be treated as hazardous waste. The high cost of incineration has limited its use to highly populated areas that produce large quantities of solid wastes on a continuous basis. Even in these areas, social and political pressures have limited the use of incineration for processing solid wastes.

Sanitary landfilling is the engineered burial of solid wastes. Fundamentally, sanitary landfills are sound processing systems for municipal solid wastes. The two most common forms of sanitary landfills are **trench landfills** and **area landfills**. Both methods employ heavy equipment to compact the solid wastes to at least 1000 lb/yd<sup>3</sup>. Unfortunately, inadequate designs and improper operations have combined to create a poor public image for sanitary landfills. The major problems with sanitary landfills have been (1) the failure to apply adequate soil cover on the compacted solid waste at the end of each day, allowing solid wastes to blow over the landfill surface and odors to escape; and (2) allowing water to enter the landfill, creating leachate and methane gas. Leachate passing through the sanitary landfill has contaminated groundwater as well as surface waters in some locations. Methane gas formation has produced fires in a few landfills, creating concern in the media over the safety of the sanitary landfills. The simplicity of sanitary landfills has made them the most popular form of solid waste processing with local government. The failure of local government to operate sanitary landfills properly has led the EPA to develop more complex regulations under Subtitle D of RCRA. Leachate and gas collection systems—as well as water barriers and sampling systems to prove that leachate and gas are not leaving the landfill and creating environmental pollution—will be required in the future. The cost of sanitary landfill processing of solid wastes will increase significantly as the EPA develops more stringent operational requirements.

**Reuse and recycling** have long been important as a solid waste processing system. Most countries with limited economic bases do not have serious solid waste problems. These countries use and reuse all their resources as a necessity. As countries grow economically, their production of solid wastes increases. In 1970 the U.S. demonstrated that current solid wastes could be recycled. Unfortunately, recycling of solid wastes was not as cost-effective as using raw materials and did not gain much acceptance. Environmental groups brought pressure on various industrial groups to begin recycling. The aluminum can industry developed the best recycling system and survived the challenge from environmentalists. The plastic industry recognized the dangers that adverse media exposure posed and began to develop methods for plastic recycling. Glass was easily recycled but it was too heavy to ship very far and was made from an abundant material widely available in nature. Office paper was the easiest paper product to recycle. Corrugated cardboard was also recyclable. Newspapers were abundant but had to be used at a lower product level or de-inked for fiber recovery. The problem with lower product formation was that ultimately the material had to be processed back into the environment. De-inking was expensive and produced water pollution with concentrated solids that had to be processed before being returned to the environment.

The EPA established a Municipal Solid Waste Task Force in 1988 to develop a national agenda for solid wastes [EPA, 1989]. With the help of various environmental organizations they published *The Solid Waste Dilemma: An Agenda for Action*. They called for a national goal of a reduction of 25% in solid wastes by source reduction/recycling by 1992. It was recognized that 17% of the solid wastes was composed of yard wastes that could be removed and composted for reuse. This fact meant that only 8% more solid waste reduction was necessary in the form of metal cans, glass bottles, and newspapers. A number of state legislatures quickly accepted the 25% reduction goal and eliminated the discharge of yard wastes to sanitary landfills. Local communities had to eliminate collection of yard wastes or establish separate collections and composting programs. A

few states legislated mandatory recycling. Unfortunately, mandating recycling created a sudden increase in recyclables that overwhelmed the markets, driving the prices down and increasing the costs. Unfortunately, the excess solid wastes were processed into sanitary landfills rather than being recycled. It was quickly learned that even the commodity of solid waste was governed by supply and demand economics. Distortion of economics by government leads to chaos rather than to a positive solution. The basic problem has been excessive social and political pressure for an instant solution to problems that require time and patience to resolve. However, some progress has been made with partial recycling of some solid waste components. A recent survey by Lisa Rebasca in *Waste Age* [1994] indicated that 40 states had data on recycling rates. The rates of recycling ranged from 3 to 56% with a median of 14%.

## 88.7 Final Disposal

---

There is no doubt that solid waste **reduction**, reuse, and recycling is the best method for processing solid wastes. The major issue is the same as it was in 1970—economics. Society has not learned how to reduce the production of solid wastes while maintaining an economically viable operation. Solving the problem of solid waste involves determining what materials should not be produced and how to efficiently employ people in meaningful endeavors. Reuse and recycling is more attractive since work is required to reprocess the waste materials into useful products. It should be recognized that all materials come from the earth and most materials are still available in the solid wastes. Recycling is the ultimate solid waste process, with wastes being converted back to raw materials for manufacture to new products to benefit all people. The strength of democratic societies lies in their ability to adapt to current conditions with a minimum of wasted effort. Until complete recycling systems are developed that can economically process solid wastes, storage will continue to be the final disposal of solid wastes. Storage may take many forms, including burial below the ground surface or above the ground surface, as in a sanitary landfill. When society needs the buried materials, the sanitary landfills will become the mines for future generations. Few people recognize that the future of current societies depends on their ability to process solid wastes back into the environment without creating serious pollution problems. The processing of polluted air and polluted water results in the production of solid wastes that must be returned to the environment. Survival does not depend on the heads of state or various legislative bodies, but, ultimately, on the action of a few individuals at the local level. Those dedicated individuals who collect and process all the solid wastes that society produces provide the environment that makes life worth living and allows the rest of society to do the things it finds most enjoyable. The future is bright because the solid waste profession has met all the challenges to date.

### Defining Terms

**Area landfill:** A landfill constructed by placing the solid waste against a natural hill, compacting the solid waste to 1000 lb/yd<sup>3</sup>, and covering with a minimum of 2 ft soil for final cover. The next cell is constructed by placing the solid waste against the previous cell and raising the land surface to the top of the landfill.

**Ash:** Residual inorganic material remaining after incineration.

**Incineration:** Controlled burning of mixed solid wastes at a temperature of 1800°F for a sufficient time to oxidize 99% of the organic matter in the solid waste.

**Solid wastes:** All solid materials that have no significant economic value to the owner and are discarded as wastes.

**Trench landfill:** A landfill constructed by digging a trench 10 to 15 ft deep and 20 to 30 ft wide, filling it with solid waste, compacting the solid waste to 1000 lb/yd<sup>3</sup>, and covering with a minimum of 2 ft soil for final cover.

**Waste recycling:** The reuse of solid waste materials after separation and reprocessing to new products.

**Waste reduction:** The reduction in the production of solid wastes, primarily by reducing the production and consumption of goods by society.

**Waste reuse:** The direct reuse (without major processing) of solid wastes by members of society.

## References

- Environmental Protection Agency. 1989. *The Solid Waste Dilemma: An Agenda for Action*. EPA/530-SW-89-019. EPA, Washington, DC.
- Environmental Protection Agency. 1990. *Characterization of Municipal Solid Waste in the United States: 1990 Update*. EPA/530-SW-90-042. EPA, Washington, DC.
- Platt, B., Doherty, C., Broughton, A. C., and Morris, D. 1990. *Beyond 40 Percent: Record-Setting Recycling and Composting Programs*. Institute for Local Reliance, Washington, DC.
- Rabasca, L. 1994. State recycling rates plateau. *Waste Age*. 25(6):48–52.

## Further Information

Current federal regulations can be obtained from the Office of Solid Wastes, U.S. Environmental Protection Agency, Washington, D.C.

EPA reports are available through the National Technical Information Service (NTIS).

The Institute for Local Reliance is located at 2425 18th Street NW, Washington, D.C. 20009. It is a nonprofit research and educational organization, providing technical information to local governments and interested citizens.

*Waste Age* is published monthly by the Environmental Industry Association, 4301 Connecticut Ave. NW, Suite 300, Washington, D.C. 20008.

Cota, H. M., Wallenstein, D. "Hazardous Waste Management"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

This chapter is not available because of  
copyright issues

Bang Mo Kim, B., Shapiro, A. P. "Soil Remediation"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

- 90.1 Regulations
- 90.2 Treatment Technologies
- 90.3 Landfilling and Containment
- 90.4 Soil Vapor Extraction
- 90.5 Thermal Treatments
- 90.6 Stabilization
- 90.7 Bioremediation
- 90.8 Soil Washing
- 90.9 Emerging Technologies

**Bang Mo Kim**

*General Electric Corporate Research and Development*

**Andrew P. Shapiro**

*General Electric Corporate Research and Development*

## 90.1 Regulations

---

Among the federal and state regulations related to soil contamination and cleanup, the Comprehensive Environmental Response Compensation Liability Act (CERCLA, or Superfund) and the Resource Conservation and Recovery Act (RCRA) Corrective Action are most important. CERCLA, which was enacted in 1980, is the first comprehensive set of federal laws addressing releases of hazardous substances into the environment [U.S. EPA, 1990a]. The main objective of CERCLA is to establish an organized and cost-effective mechanism for responding to releases of hazardous materials, or to abandoned or uncontrolled hazardous waste sites, that present a serious threat to human health and environment. To accomplish this, CERCLA mandates two types of responses: an emergency response action for handling major chemical spills or incidents requiring immediate action and a remedial response capability for undertaking the long-term cleanup of hazardous waste disposal sites. The regulatory framework developed to guide these responses became the National Contingency Plan (NCP), which first outlined the level of cleanup necessary at Superfund sites and established basic procedures pertaining to discovery and treatment of hazardous waste sites. The federal Superfund cleanup process begins with the EPA or state agency conducting a preliminary assessment of a suspected site. If the results show that hazardous



substances have possibly been released to the environment, the agency will conduct a site inspection. Based on this information the agency determines whether the site is added to the National Priority List (NPL). Once the site is on the NPL, the following procedures are followed: (1) remedial investigation (RI)/feasibility study (FS), (2) public comment period, (3) record of decision (ROD), and (4) remedial design (RD)/**remedial action** (RA) [Hopper, 1989].

Whereas the Superfund regulation relates to the past hazards and cleanup requirements for hazardous waste sites, the RCRA amendments of 1984 focus on the cleanup of operating manufacturing facilities and the control of hazardous waste releases. RCRA imposes regulatory standards and permit requirements for ongoing hazardous waste management activities. Discretion is left to federal/state agencies on a site-specific basis to determine if, when, and what corrective actions are required. A facility that treats, stores, or disposes hazardous wastes may be required to initiate a corrective action to prevent a release of hazardous substances to the environment. Corrective action must address all releases of hazardous substances in the entire facility. The steps in RCRA corrective action include: (1) RCRA facility assessment, (2) RCRA facility investigation, (3) identification of remedies, (4) implementation of remedies, and (5) operation and maintenance [U.S. EPA, 1990a]. Table 90.1 lists acronyms commonly used in regulation for soil and remediation of soil.

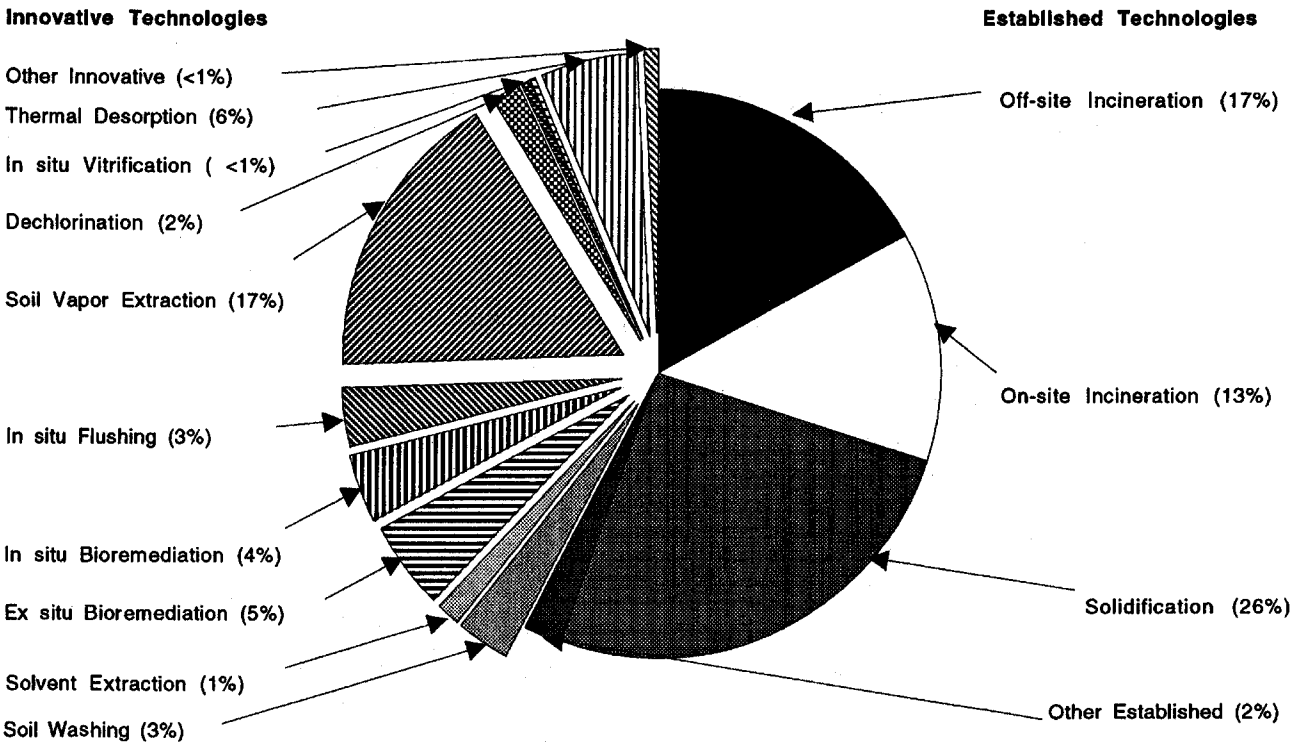
**Table 90.1** Acronyms Used in Soil Remediation

ARAR	Applicable or relevant and appropriate requirements
CERCLA	Comprehensive Environmental Response Compensation and Liability Act
HRS	Hazardous ranking system (EPA's system for ranking the priority of cleanup sites)
MCL	Maximum contaminant level, enforceable standards under the Safe Drinking Water Act
NCP	National Contingency Plan
NPL	National Priority List, a list of sites eligible for federal cleanup
PRPs	Potentially responsible parties, who may be held liable for cleanup costs pursuant to CERCLA
RA	Remedial action
RCRA	Resource Conservation and Recovery Act
RD	Remedial design
RI/FS	Remedial investigation/feasibility study
ROD	Record of decisions
SARA	Superfund Amendment Reauthorization Act
SITE	Superfund Innovative Technology Evaluation
TCLP	Toxicity characteristics leaching procedure
TSCA	Toxic Substance Control Act

# 90.2 Treatment Technologies

As of 1991, landfilling and containment [U.S. EPA, 1992a] accounted for about 37% of all the remedial actions specified in Superfund RODs. The obvious drawbacks of landfilling are that the process does not eliminate the contamination or the liability associated with it and that the number of available sites for landfilling is constantly being reduced. Therefore, there are great incentives for developing alternatives to landfilling. The other 63% of the remedial actions, specified in the 1991 RODs, included both well-demonstrated and innovative technologies. The distribution of the nonlandfilling remedial actions for 1991, grouped by technology type, is shown in Fig. 90.1. The most commonly used technologies are thermal processes such as incineration, followed by solidification/stabilization and soil vapor extraction. The selection of a technology depends most on the physicochemical properties of the contaminant and soil, the local hydrogeology, and the amount of soil requiring treatment. Table 90.2 summarizes treatment technologies investigated for remediation of Superfund sites.

**Figure 90.1** Alternative treatment technologies through 1991. (Source: U.S. EPA, 1992.)



**Table 90.2** Treatment Technologies Investigatedfor Remediation of Superfund Sites

Technology	Contaminants
Soil vapor stripping	VOCs, <sup>1</sup> SVOCs, <sup>2</sup>
Thermal desorption	VOCs, SVOCs, PCBs, pesticides, <sup>3</sup> metals <sup>4</sup>
Soil washing	Metals, dioxins, PAHs, <sup>5</sup> PCBs, pesticides
Solvent extraction	PCBs, VOCs, SVOCs,

In situ flushing	VOCs, SVOCs, PAHs, metals
Solidification	Metals, pesticides, SVOCs, VOCs
Bioremediation	PAHs, SVOCs, VOCs
Dechlorination	PCBs, SVOCs, dioxins, pesticides

Examples of each classification are as follows:

<sup>1</sup>TCE, TCA, DCE, DCA, PCE, BTEX, TX, MEK, acetone, benzene carbon tetrachloride, chloroform, isopropyl alcohol, Freon, methylene chloride, styrene, vinyl chloride

<sup>2</sup>PCP, DNT, Aniline, Bis 2-ethyl hexyl phthalate, chlorobenzene, dimethylphenol, ethyl benzene, naphthalene, phenol, polychlorinated phenols

<sup>3</sup>DDT, DDE, DDD, ethyl and methyl parathion,, silvex, Toxaphene

<sup>4</sup>Arsenic, cadmium, chromium, copper, lead, mercury, nickel, silver

<sup>5</sup>Benzo anthracene, benzo pyrene, cresol, creosote, petroleum hydrocarbon

## 90.3 Landfilling and Containment

Landfilling and containment are the most widely used "treatments" for contaminated soil and account for about a third of all remedial actions. Although neither of these approaches decontaminate the soil, they do attempt to reduce the risk of the pollutants coming in contact with people or the surrounding environment. Ironically, about 25% of today's Superfund sites are former landfills. Modern landfills are usually lined with a clay layer or synthetic impermeable geomembrane that prevents leaching of contamination into the local groundwater. Landfills are classified as either *secured hazardous waste landfills* or *sanitary landfills*. Most contaminated soil that is intended for landfilling will be put in secured hazardous waste landfills.

Containment implies installation of impermeable barriers around the contaminated region. Examples of barriers are trenches backfilled with a bentonite slurry (slurry walls) or lined with a synthetic geomembrane. The top of the contaminated site is usually capped with a clay layer. The obvious difficulty in containment is installing a barrier beneath the contaminated region. In some cases the contaminated soil must be excavated and the empty hole lined with a barrier material before the contaminated soil is backfilled into the hole.

## 90.4 Soil Vapor Extraction

Soil vapor extraction (SVE) or soil vapor stripping is an **in situ** process in which several wells are sunk into and around the contaminated soil. In the simplest applications, air blowers attached to well heads create a partial vacuum within the wells, causing air and vapors in the surrounding soil to flow towards the wells and out of the soil. Volatile contaminants in the soil vapor are removed with the vapors and are usually treated above ground by thermal oxidation or adsorption on activated carbon. As the contaminant vapors are removed, liquid contaminant present in the soil evaporates to maintain equilibrium between the liquid and vapor phases of the contaminant. Thus liquid contaminants are removed from the soil by purging the contaminant's vapor phase until the liquid has completely evaporated.

Many variations of SVE exist that improve performance and extend the range of applicability of

the process. In some cases positive pressure is applied to wells to help control the flow direction. Heated air or steam can be injected into some wells to warm the soil and thereby increase the vapor pressure of volatile compounds. Increasing the temperature can dramatically increase the rate and degree of contaminant removal.

SVE is applicable to sites that have high soil permeability ( $K_h > 10^{-9} \text{ m}^2$ , or 1 cm/s) and are contaminated with volatile organic compounds (VOCs), as listed in Table 90.2. The vapor pressure of the contaminants should be greater than about 0.5 torr for effective performance. SVE is not appropriate for semivolatile compounds such as PCBs and many pesticides. Drier soils are easier to clean than soils with high moisture content, but the process has been applied to partially saturated soils. The performance and cost estimate of soil vapor extraction and other processes are shown in Table 90.3.

**Table 90.3** Performance and Cost of Remediation Technologies

Contaminants	Performance		Operating Condition	Cost Estimate
	Before	After (% removal)		
Soil Vapor Extraction				
Hydrocarbons	70 ppm	0.002 ppm	Steam enhanced	\$35–40/ton
Carbon tetrachloride			Ambient condition	
Benzene, TCE, PCE, DCA, MEK		Tot. VOC < 10 ppm Avg. VOC < 1 ppm	Continuous operation	
TCE		(95%)		\$40/ton(\$10–150/ton)
TCE, TetCE		(85 to 99% CVOC)  (55% SVOC)	Steam: 450°F & 450 psi,  hot air: 300°F & 250 psi	\$250–350/yd <sup>3</sup>
Thermal Desorption				
VOC	32 000 ppm	4.5 ppm	T = 160°C, heated screw	
TCE	1000 ppm	0.1 ppm	T = 300°F, t = 6–8 min	
PCB	1.04–2.3%	< 2 ppm	T = 1000°F	\$185/ton
Coal tars, wood preservatives		(Nearly 100%)	T = 260 to 815°C	\$260–1000/ton
PCB, PAH, pesticides		(99.9999%)	T = 850°C	
PCB		(99.99%)	Heated up to 1850°F	\$800/ton
Solidification/Stabilization				
Chromium		TCLP < 0.5 ppm		\$73/ton
PCB, organics	PCB: very low	TCLP: ND	In situ	
Organics, heavy metals	Pb: 2.2%; oil & grease: 25%;	TCLP = Pb: 100ppb; oil & grease: 4	Cement, fly ash,chloranan	\$40–60/ton for metals;\$75–100/ton

Hydrocarbons, metals	VOC: 100 ppm	ppm TCLP = organics < detection limit Ba: 2 mg/L	additives Emulsified asphalt	for organics \$88–100/ton for organics
Pb, Cd, As, Sb	Pb: total = 1.2% TCLP = 620 mg/L	TCLP: 0.4 mg/L	Dry, wet additives	\$67/ton
Bioremediation				
PAHs	800–2000 ppm	10–80 ppm	Time = 3–6 months	
PAH (creosote)	13 000 ppm	500 ppm	Slurry with 30% soil nutrients & surfactants	\$350/yd <sup>3</sup>
Gasoline, diesel, Jet-A, JP-4	88 000 ppm	(BTEX > 30%)	Microaerophilic bacteria micronutrients	\$25–50/yd <sup>3</sup>
Waste oil Toxaphene	470 ppm	(TPH > 80%) 180 ppm	Anaerobic, pH = 8.3–9.8	
Soil Washing				
Hg, PAH, organic Pesticides		(90%)	Flotation	\$125–200/ton
Hydrocarbons	15 000 ppm	(85–99%)	Surfactants, biosurfactants	\$250/ton \$40–140/ton
As, Cr	As: 2–6200 ppm Cr: 4–6200 ppm Cr: 4–6200 ppm	As < 1 ppm Cr: 627 ppm Cr: 627 ppm	Ambient, pH = 2–9	

Compiled from Anderson, 1994; U. S. EPA, 1990, 1991, 1992.

## 90.5 Thermal Treatments

There are many techniques that use heat to remediate soil, and they vary widely in the manner in which the soil is heated and the temperatures at which they operate. At one end of the spectrum is in situ steam stripping, described in the previous section as an SVE technique. In this process the soil temperature may be increased 20 to 40°C, which greatly assists the evaporation of volatile organics but does not destroy them. At the other extreme is incineration, which typically involves heating excavated soil with burning fuel in a furnace. For incinerators permitted to treat PCB-contaminated soil, for instance, temperatures must reach 1000°C with residence times of at least two seconds. Incineration can treat soils contaminated with organics and volatile metals like mercury and arsenic. The ash resulting from incineration may or may not be hazardous, depending primarily on its heavy metal content.

Another widely used **ex situ** thermal treatment is low-temperature thermal desorption. In this process, excavated soil is heated to 300 to 600°C so that volatile and semivolatile contaminants are vaporized. At these temperatures only partial destruction of some organic compounds is expected, and vapor treatment therefore is necessary. Usually the off-gases are either condensed and concentrated or passed through a thermal oxidizer to destroy the remaining organics. Carbon

adsorption and wet scrubbers are often used as polishing steps.

Thermal treatments account for about 36% of nonlandfilling remedial actions (Fig. 90.1). The advantage of thermal treatments is that they are robust. If the temperature is hot enough and the residence time long enough, any organic contaminant can be either volatilized, pyrolyzed, or oxidized. In ex situ processes like incineration and low-temperature desorption, the required operating conditions can be carefully controlled. However, ex situ thermal processes require excavation and handling of the soil, both steps adding expense and complexity to the process. In situ thermal processes generally require less complex machinery but also have less control of the operating conditions. In addition to steam-assisted SVE, in situ processes are being developed that make use of radio frequency and electric resistance heating for removal of VOCs and SVOCs.

## 90.6 Stabilization

---

Soil stabilization is the process of mixing the soil with an additive, such as a cement or grout, so that contaminants are bound in the resulting mixture and cannot leach out to the surrounding environment. Stabilization has been used to treat soils contaminated with heavy metals and organics such as PCBs. In the case of organics contaminants, special additives with high organic affinity must be added to the stabilizing mixture. Usually the soil is excavated and processed above ground before being put back into its original location. The additives may compose about 20 to 30% of the volume of the resulting mixture so that significantly more space is required for the treated soil.

There are also techniques for mixing the stabilizing agents directly into the soil without excavation. This is accomplished using machinery such as the Mectool, which is a large auger attached to a drill rig. As this auger is driven down through the soil, stabilizing agents are pumped through the auger blades and mixed with the soil.

## 90.7 Bioremediation

---

Many organic compounds or contaminants are degraded by microorganisms present in the soil, producing environmentally benign compounds. Biodegradation of contaminants can occur naturally if the soil has a proper mix of microorganisms, nutrients, and other environmental conditions such as pH, molecular oxygen, concentration, and temperature. Natural restoration (i.e., simply waiting for the site to improve, with minor site manipulation) is the most cost-effective way to decontaminate sites, if it is applicable. Sites should be investigated for the potential of natural restoration by determining the rate of contaminant degradation and formation of by-products or intermediates, and their environmental impacts.

*Bioremediation* refers to enhancement of biodegradation by altering conditions to increase its rate and effectiveness. Most applications of bioremediation make use of naturally occurring microorganisms that can use the organic contaminants as a food or energy source under favorable conditions. Applying bioremediation to a site usually entails manipulating the conditions in order to stimulate microbial activity. Examples of techniques used to stimulate biodegradation include injecting air into the soil to promote aerobic activity, supplying nutrients to promote growth of the microbial population, and increasing the temperature to stimulate microbial

metabolism.

Bioremediation is only appropriate for certain organic contaminants and depends on the hydrogeology of the site. As with most other in situ processes, bioremediation works best in sandy soils. In such soils there is enough transport of water and vapors to provide nutrients and oxidizing and reducing agents to microbes and to remove microbial wastes.

## 90.8 Soil Washing

---

Soil washing processes can take on many forms. In ex situ processes the soil is usually prepared for washing by grinding and size fractionation. Because large soil particles have a relatively small surface area per mass, they adsorb small amounts of contaminants; therefore, size fractionation often greatly reduces the volume of soil that needs to be washed. The washing process can be done with water alone or with water plus **surfactants**, organic solvents, or acids. Surfactants and organic solvents are used to remove organic contaminants, and acids are used to remove metals. In situ soil washing is a less common technique that involves pumping a washing solution into injection wells and extracting it at production wells installed in the contaminated soil.

Soil washing works best with soils that have low surface-area-to-mass ratios, such as sandy soils. Clay soils present problems in terms of desorbing the contaminants and separating the fine particles from the washing solution.

## 90.9 Emerging Technologies

---

Several technologies are being developed that may solve problems that present technologies cannot address. For radioactive and mixed radioactive/hazardous waste soils, in situ **vitrification** may be able to destroy the organic contaminants and encapsulate the radioactive compounds in glassy material. For low permeability soils in which SVE and bioremediation cannot work, in situ **electrokinetic** processes have been shown to remove metals and water-soluble organics [Shapiro, 1993]. In situ chemical destruction techniques are also being developed, such as a method to destroy TCE in groundwater as it flows through an installed permeable wall made of iron particles.

### Defining Terms

**Electrokinetic:** A process in which water and dissolved materials are caused to move soil as a result of an applied electric field.

**Ex situ:** A process conducted on excavated soil.

**In situ:** A process conducted in place, without excavation of soil.

**PCBs:** Polychlorinated biphenyls, dielectric compounds previously used in high-voltage electrical equipment such as transformers and capacitors.

**Remedial action:** Specific steps taken to decontaminate a site.

**Surfactants:** Surface active chemicals, typically found in soaps, that assist solubilization of nonpolar organic materials in water.

**SVOCs:** Semivolatile organic compounds, which slightly volatilize at room temperature.

Examples are shown in [Table 90.2](#).

**Vitrification:** Process of heating soil to sufficient temperature to cause melting of silica minerals and produce a glass-like material.

**VOCs:** Volatile organic compounds, which readily volatilize at room temperature. Examples are shown in [Table 90.2](#).

## References

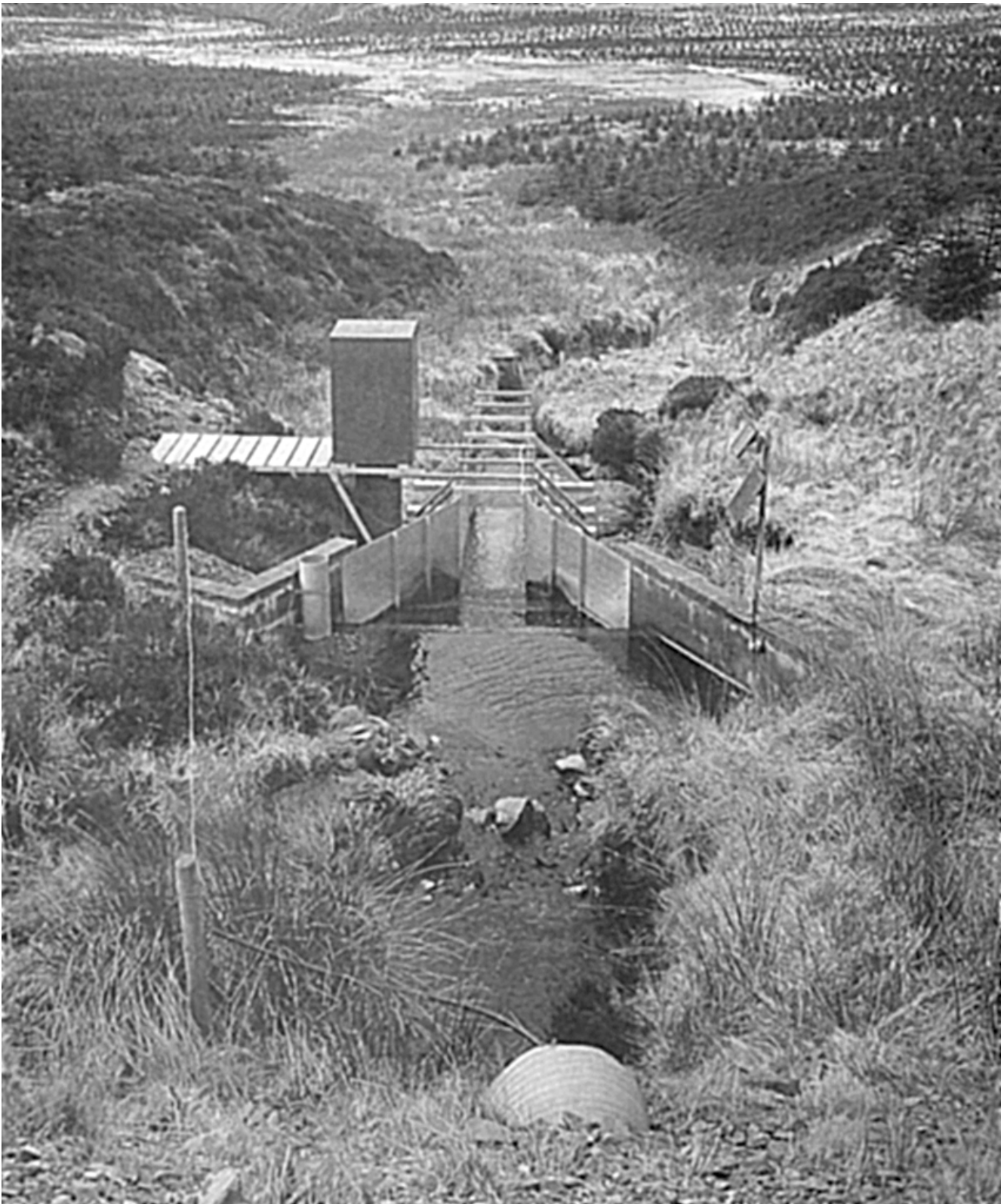
- Anderson, W. C. 1994a. *Innovative Site Remediation Technology: Stabilization/Solidification*. American Academy of Environmental Engineers, Annapolis, MD.
- Anderson, W. C. 1994b. *Innovative Site Remediation Technology: Chemical Treatment*. American Academy of Environmental Engineers, Annapolis, MD.
- U.S. EPA. 1990a. *Handbook on In Situ Treatment of Hazardous Waste-Contaminated Soils*. EPA/540-2-90/002. U.S. EPA, Washington, DC.
- U.S. EPA. 1990b. *Federal Register*. Vol. 55. No 145. July 27, 1990.
- U.S. EPA. 1991. *The Superfund Innovative Technology Evaluation Program*. EPA/540/5-91/008. U.S. EPA, Washington, DC.
- U.S. EPA. 1992a. *Innovative Treatment Technologies: Semi Annual Status Report*. EPA/540/2-91/1001 No. 3. U.S. EPA, Washington, DC.
- U.S. EPA. 1992b. *Abstract Proceedings: Fourth Forum on Innovative Hazardous Waste Treatment Technologies*. EPA/540/R-92/081. U.S. EPA, Washington, DC.
- Hopper, D. R. 1989. Cleaning up contaminated waste sites. *Chem. Eng.* August 1989, p. 95.
- Shapiro, A. P. and Probstein, R. F. 1993. Removal of contaminants from saturated clay by electroosmosis. *Env. Sci. & Tech.* 27:283.

## Further Information

- Wilson, D. J. and Clarke, A. N. 1994. *Hazardous Waste Site Soil Remediation*. Marcel Dekker, New York.
- Kostecki, P. T. and Calabrese, E. J. 1988. *Petroleum Contaminated Soils*, Vol. 1. Lewis, Chelsea, MI.
- Calabrese, E. J. and Kostecki, P. T. 1989. *Petroleum Contaminated Soils*, Vol. 2. Lewis, Chelsea, MI.



Delleur, J. W. "Water Resources Engineering"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



The processes of surface runoff and infiltration into the ground are normally effective in removing excess water from the land. Where these processes are impeded, wetland habitats, with their moist or permanently saturated soils, may develop. Plant life in wetlands is highly specialized in order to cope with excess water, lack of oxygen, and frequently a scarcity of nutrients in the soil. The rural economy of wetland areas is also distinctive, based on drainage and careful water level control or on the cultivation of wetland crops such as rice and freshwater fish and the production of traditional materials such as reed and peat. The conservation of wetland areas requires an understanding of their hydrology.

The photo is of the lower half of the Ceunant Ddu subcatchment in central-Wales. The view shows the wetland area within the young forest and the flow-measuring structure for the upper half of the catchment. (Courtesy of the Natural Environment Research Council.)

# XV

## Water Resources Engineering

---

**Jacques W. Delleur**

*Purdue University*

**91    Hydraulics    *B. Hauser***

Flow Characteristics • Equation of Continuity • Pressure Characteristics • Effects of Pressure—Dynamic Systems • Pressure Loss • Open Channel Flow • Flow Measurement • Centrifugal Pump

**92    Hydrology    *V. P. Singh***

Classification of Hydrology • Hydrologic Cycle • Laws of Science • Approaches to Hydrologic Problems • Tools for Hydrologic Analyses • Components of Hydrology Cycle—Deterministic Hydrology • Statistical Hydrology • Hydrologic Design

**93    Sedimentation    *E. V. Richardson***

Fluvial Geomorphology • Sediment Properties • Beginning of Motion • Sediment Yield • Bed Forms • Sediment Transport • Reservoir Sedimentation

WATER RESOURCES ENGINEERING encompasses a number of fields associated with design, construction, and management activities related to water resources. Basic to all these fields are hydraulics and hydrology, which are concerned with the flow of water in the human-made and natural environments, respectively.

Records of different applications of hydraulics go back to antiquity. There are ancient Egyptian pictographs of water bailers and of siphoning. The principles of hydrostatics and flotation began with the early Greek Archimedes. The elements of hydraulics of open channels were known to the Romans, who used the techniques very efficiently in their aqueducts, such as the famous Pont du Gard in southern France and the Cloaca Maxima or main sewer of ancient Rome. But we have to look to the Renaissance to find the beginnings of experimental hydraulics.

Leonardo da Vinci is well known for his drawings of eddy formations, profiles of free overfalls, and flows over weirs, for example. Our understanding of the velocity of flow from orifices is due to the 17th-century Italian Toricelli. Also in that century the Frenchman Pascal developed the hydraulic press, which he called the machine for multiplying forces. In the next century Daniel Bernoulli dealt with the dynamics of fluids; the well-known theorem that bears his name is concerned with the different forms of energy in a steadily moving incompressible fluid.

Toward the end of the 18th century experimenters such as Borda and Venturi worked on devices for the measurement of flow in pipes. The classical form of a contraction followed by an expansion is known as the Venturi tube. In the second half of the 19th century the Englishman Osborne Reynolds developed the criterion that today is known as the Reynolds number and that distinguishes between laminar and turbulent flows. Robert Manning, who served as president of the Irish Institution of Civil Engineers in the latter part of the 19th century, is well known for the formula for the velocity of flow in open channels, which is named after him.

Counterparts to the Venturi tube for measurement in open-channel flow were developed in the 20th century. One of these is the Parshall flume. The merging of experimental hydraulics and fluid mechanics occurred in this century, and substantial developments have taken place in universities and laboratories of the U.S. Compared with hydraulics, hydrology is a much more recent science. This is not to say that there were no concerns in earlier times about the flow of water in the natural environment. In fact, Nilometers were in use in 3000 B.C. to record the fluctuations of the Nile. However, accurate understanding of the hydrologic cycle is very recent, and the principles were not fully accepted until the 19th century.

Although the equations for surface flow were formulated by de Saint Venant in 1871, their practical solution had to wait until the 1960s, with the advent of the digital computer. Darcy's experiments with flow through sand columns, published in 1856, form the basis of our understanding of groundwater flow. The extension of Darcy's law to unsaturated flow was presented by Richards in 1931.

Recent advances in hydrology are concerned with the transport of solutes in the natural environment. In groundwater this transport takes place by advection, diffusion, and mechanical dispersion. Chemical reactions can take place between the solute and the rock or soil medium.

Of course, one very important phase of transport is that of sediments. Suspended sediment can move by convection and turbulent diffusion and is sometimes associated with the transport of pollutants. The evaluation of bed load transport is more empirical. The name of Hans Einstein, son of Albert Einstein, is associated with the transportation of bed load. Under the direction of Vito Vanoni, the ASCE published in 1975 a comprehensive manual and report on engineering practice with regard to sedimentation.

It is hoped that the placement of several of the key names in their historical context will whet the reader's interest in pursuing the following chapters.

Hauser, B. "Hydraulics"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 91

## Hydraulics

---

### 91.1 Flow Characteristics

### 91.2 Equation of Continuity

### 91.3 Pressure Characteristics

Unit Pressure • Pressure Measurement • Total Pressure • Components of Pressure • Bernoulli's Theorem

### 91.4 Effects of Pressure—Dynamic Systems

### 91.5 Pressure Loss

Head Loss—Physical Components • Compound Pipe Systems • Minor Head Loss

### 91.6 Open Channel Flow

### 91.7 Flow Measurement

Orifice Meter • Venturi Meter • Pitot Gage • Magnetic Flowmeter • Ultrasonic Meter • Positive Displacement Meter • Turbine Meter • Compound Meter • Weir • Parshall Flume

### 91.8 Centrifugal Pump

## Barbara Hauser

*Bay de Noc Community College*

Hydraulics deals with the principles that govern the behavior of liquids at rest and in motion. This is the study of the mechanics of water and its control by man. Hydraulics deals with pressurized systems and open channel flow, and includes principles of pressure and force, energy theorem, **flow** calculations and measurement, friction losses, pumps, and pumping applications.

## 91.1 Flow Characteristics

---

*Laminar flow* occurs at extremely low **velocity**; water molecules move in straight parallel lines called *laminae*, which slide upon each other as they travel, evidenced in groundwater flow; friction losses are minimal. *Turbulent flow*, normal pipe flow, occurs because of roughness encountered on the inner conduit walls. Outer layers of water are thrown into the inner layers; movement in different directions and at different velocities generates turbulence. *Steady state flow* occurs if at any one point flow and velocity are unchanging. Hydraulic calculations almost always assume steady state flow. *Uniform flow* occurs when the magnitude and direction of velocity do not change from point to point.

## 91.2 Equation of Continuity

---

At a given flow, water velocity is dependent upon the cross-sectional area of the conduit. This statement expresses the most basic hydraulic equation, the equation of continuity:

$$Q = AV \quad (91.1)$$

where  $Q$  is the flow in cfs,  $A$  is the cross-sectional area in  $\text{ft}^2$ , and  $V$  is the velocity in  $\text{ft/s}$ .

## 91.3 Pressure Characteristics

---

### Unit Pressure

Water **pressure** is due to its weight ( $62.4 \text{ lb/ft}^3$ ) and the depth of water above the point of measurement (equating pressure to depth,  $0.433 \text{ psi/ft}$  water depth, or  $2.31 \text{ ft}$  water depth/psi).

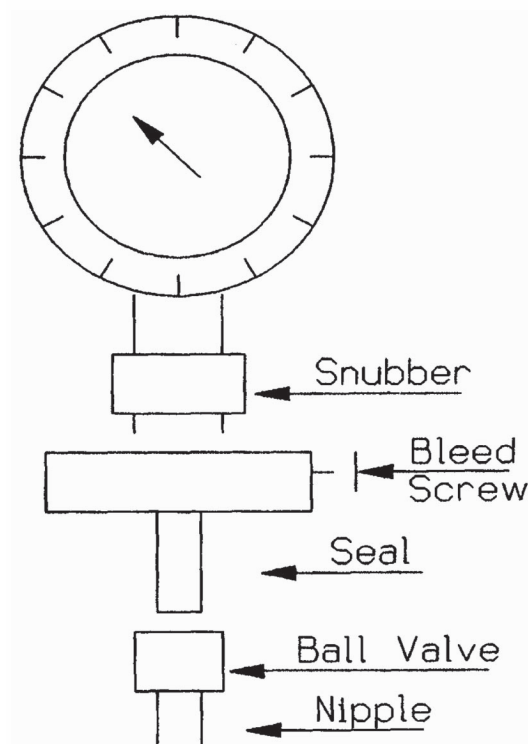
Water pressure does not affect liquid density, and *Pascal's law* states that the pressure at any one point in a static liquid is exerted with equal intensity in all directions.

### Pressure Measurement

The *piezometer* is an open-ended vertical tube inserted at the point of pressure measurement, and is impractical for measuring anything but the smallest pressures. The *manometer* is a modification of the piezometer with internal liquid fill of a higher specific gravity, bent into a U shape for easy reading. When connected to two separate sources of pressure, the unit yields a differential reading.

The *pressure gage*, called a *Bourdon tube*, may read feet, psi, or inches of mercury. It provides a direct, easy-to-read, static connection to the source of pressure (Fig. 91.1).

**Figure 91.1** Pressure gage.



The earth's atmosphere, about 200 miles deep, exerts a pressure of 14.7 psi or the equivalent of 34 ft water pressure upon the surface of the earth at sea level, the standard reference point for all pressure measurements. *Gage pressure* (psig) is the pressure read on a gage above or below atmospheric. *Absolute pressure* (psia) includes atmospheric pressure in the reading and is employed when calculating pumping suction **heads**.

## Total Pressure

**Force** is registered in pounds, and is calculated as follows:

$$\begin{aligned}\text{Force} &= \text{Pressure} \times \text{Area} & (91.2) \\ \text{lb} &= \text{lb/ft}^2 \times \text{ft}^2\end{aligned}$$

## Components of Pressure

*Pressure head* (PH) is due to the depth of the water and is measured as feet of water or registered on a pressure gage. *Velocity head* (VH) is the distance the water can move due to velocity energy; it does not register on a pressure gage, but can be captured by a pitot gage (described later) and is calculated as follows:

$$\text{VH} = \frac{V^2(\text{ft/s})^2}{2g(\text{ft/s}^2)} \quad (91.3)$$

*Elevation head* ( $Z$ ) is pressure due to elevation above the point of reference, measured as feet of water or registered on a pressure gage.

## Bernoulli's Theorem

In a fluid system employing steady state flow, the theoretical total energy is the same at every point in the path of flow, and the energies are composed of pressure head, velocity head, and elevation head ( $Z$ ). Expressed in terms of actual energy change between two points in a dynamic system where a pressure decrease or head loss (HL) occurs,

$$\text{PH}_1 + \text{VH}_1 + \neg_1 \neq \text{PH}_2 + \text{VH}_2 + \neg_2 + \text{ZHL} \quad (91.4)$$

## 91.4 Effects of Pressure<sup>3</sup>/<sub>4</sub>Dynamic Systems

---

*Water hammer* (hydraulic shock) is the momentary increase in pressure that occurs in a moving



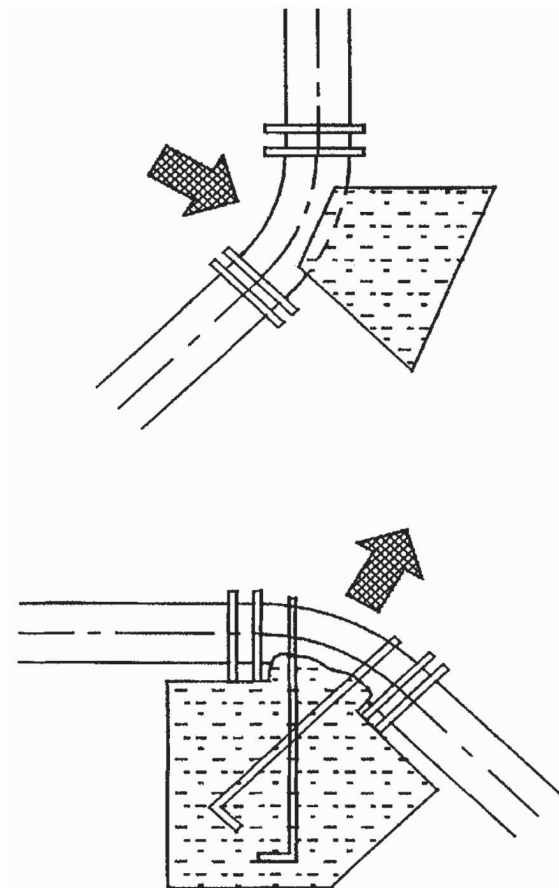
water system when there is a sudden change in direction or velocity of the water. **Suppressors** are installed where water hammer is encountered frequently in order to minimize shock and protect piping and appurtenances. *Surge*, a less severe form of hammer, is a slow-motion mass oscillation of water caused by internal pressure fluctuations in the system; it can be controlled by surge suppressors or spring-loaded pressure relief valves. *Thrust*, caused by an imbalance of pressures, is the force that water exerts on a pipeline as it turns a corner. Its intensity is directly proportional to water **momentum** and acts perpendicular to the outside corner of the pipe, affecting bends, tees, reducers, and dead ends, pushing the coupling away from both sections of pipeline. To calculate total pounds of thrust,

$$\text{Thrust} = 2TA \times \sin \frac{1}{2}\theta \quad (91.5)$$

where  $T$  is the test pressure of system in psf, plus 100 psi for hammer, and  $A$  is the cross-sectional area of fitting.

For pipes using push-on or **mechanical joints**, thrust restraint is desired. *Thrust blocks* are concrete blocks cast in place onto the pipe and around the outside corner of the turn. Block-bearing face must be large enough so that its pressure does not exceed soil-bearing strength (variable, <1000-10 000 lb/ft<sup>2</sup>, depending on soil type) (Fig. 91.2).

**Figure 91.2** Thrust blocks.



To calculate bearing face area of the thrust block:

$$\text{Area} = \frac{\text{Total thrust (lb)}}{\text{Bearing strength of soil (lb/ft}^2\text{)}} \quad (91.6)$$

In locations where it is difficult to use thrust blocks, *restrained joint pipe* is an alternative. Extra locking rings stabilize the joint under thrust conditions and transfer the load from the pipe directly to the surrounding soil.

## 91.5 Pressure Loss

*Major head loss* occurs because of friction dropping pressure along the conduit length. *Minor head loss* is caused by extra turbulence at bends, fittings, and diameter changes in the pipeline. In open channel systems, slope equates to the amount of pipe incline, as feet of drop per foot of pipe length; it is designed to be just enough to overcome friction losses so that velocity will remain constant. In closed conduit systems under pressure, the pipe is taken as horizontal unless otherwise indicated, and slope relates directly to the loss of pressure per foot of pipe:

$$\text{Slope} = \frac{\text{Head loss}}{\text{Length}} \quad (91.7)$$

## Head Loss<sup>3/4</sup>Physical Components

Interior pipe roughness is dependent upon pipe material and increases with corrosion and age, designated by the *C* factor, the roughness coefficient ([Table 91.1](#)).

**Table 91.1** Hazen-Williams Roughness Coefficient (*C*) Value

Type of Material	<i>C</i> Value
Asbestos cement	140
Brass	140
Brick sewer	100
Cast iron:	
10 yr. old	110
20 yr. old	90
Ductile iron (cement lined)	140
Concrete:	
Smooth	140
Rough	110
Copper	140
Fire hose (rubber lined)	135
Galvanized iron	120

Glass	140
Lead	130
Masonry conduit	130
Plastic	150
Steel:	
Coal-tar enamel lined	150
Unlined	140
Riveted	110
Vitrified	120

---

Length, velocity head, and diameter (inversely) also affect pressure loss. A widely used formula for flow, velocity, or head loss calculation in a closed pipe system, derived from these physical components, was developed by Hazen and Williams:

$$Q = 0.435 \times C \times d^{2.63} \times s^{0.54} \quad (91.8)$$

For field use, a **nomograph** has become popular ([Fig. 91.3](#)).

Variations of this formula are easily recognizable, and with any of them approximate values can be measured or calculated for flow, velocity, slope, or length. Accuracy is limited, however, by the roughness coefficient  $C$ , which can only be estimated based on pipe type and age and a knowledge of water quality.

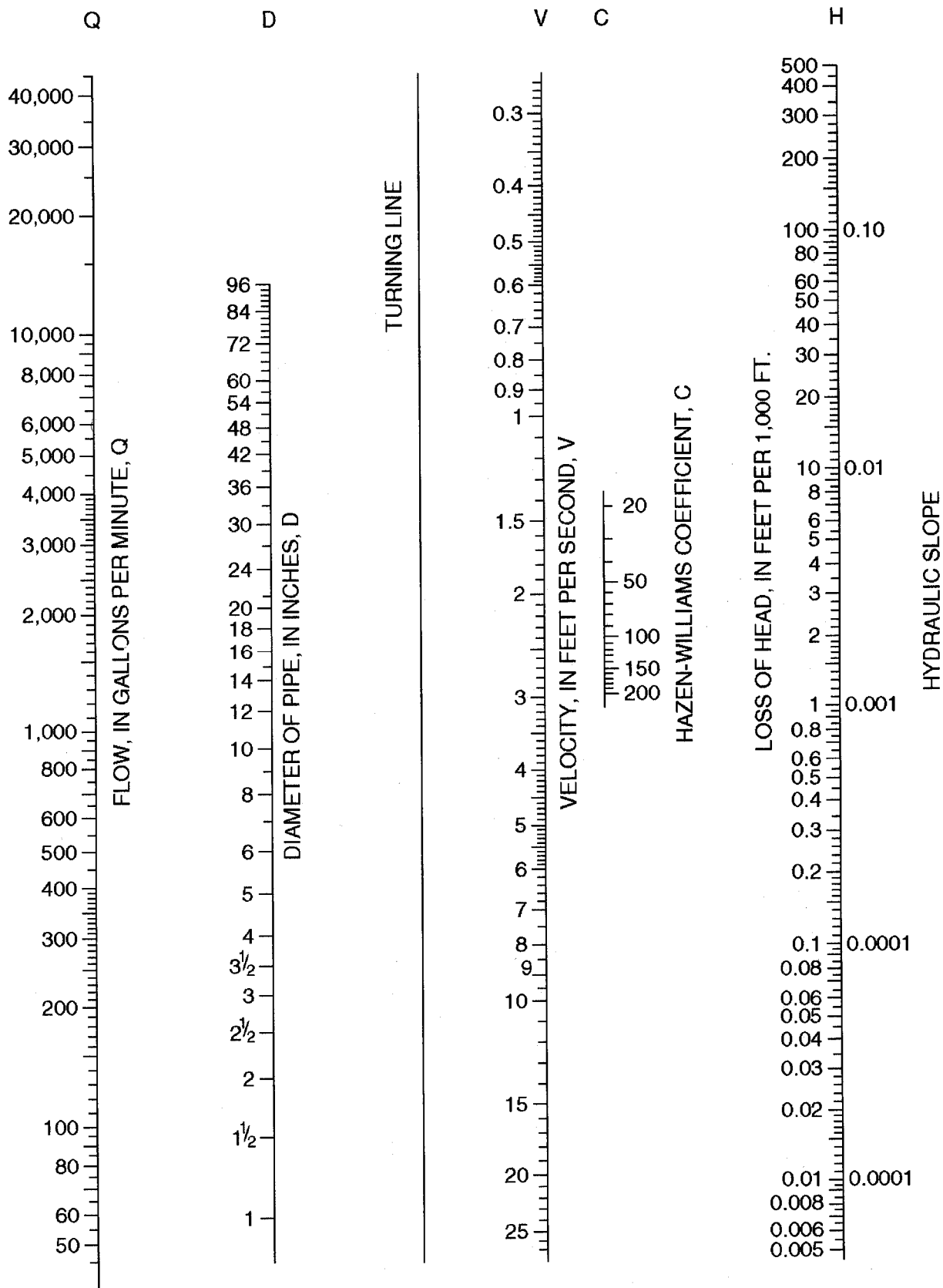
## Compound Pipe Systems

When pipes are laid in series, flow is continuous through the system and head losses in the component segments are additive. Pipes laid in parallel split flow among the components; head loss in each is identical and is the same as the total head loss. Indirect solutions may be obtained by the equivalent pipe method, creating a single pipe with head loss equivalent to that in the compound set.

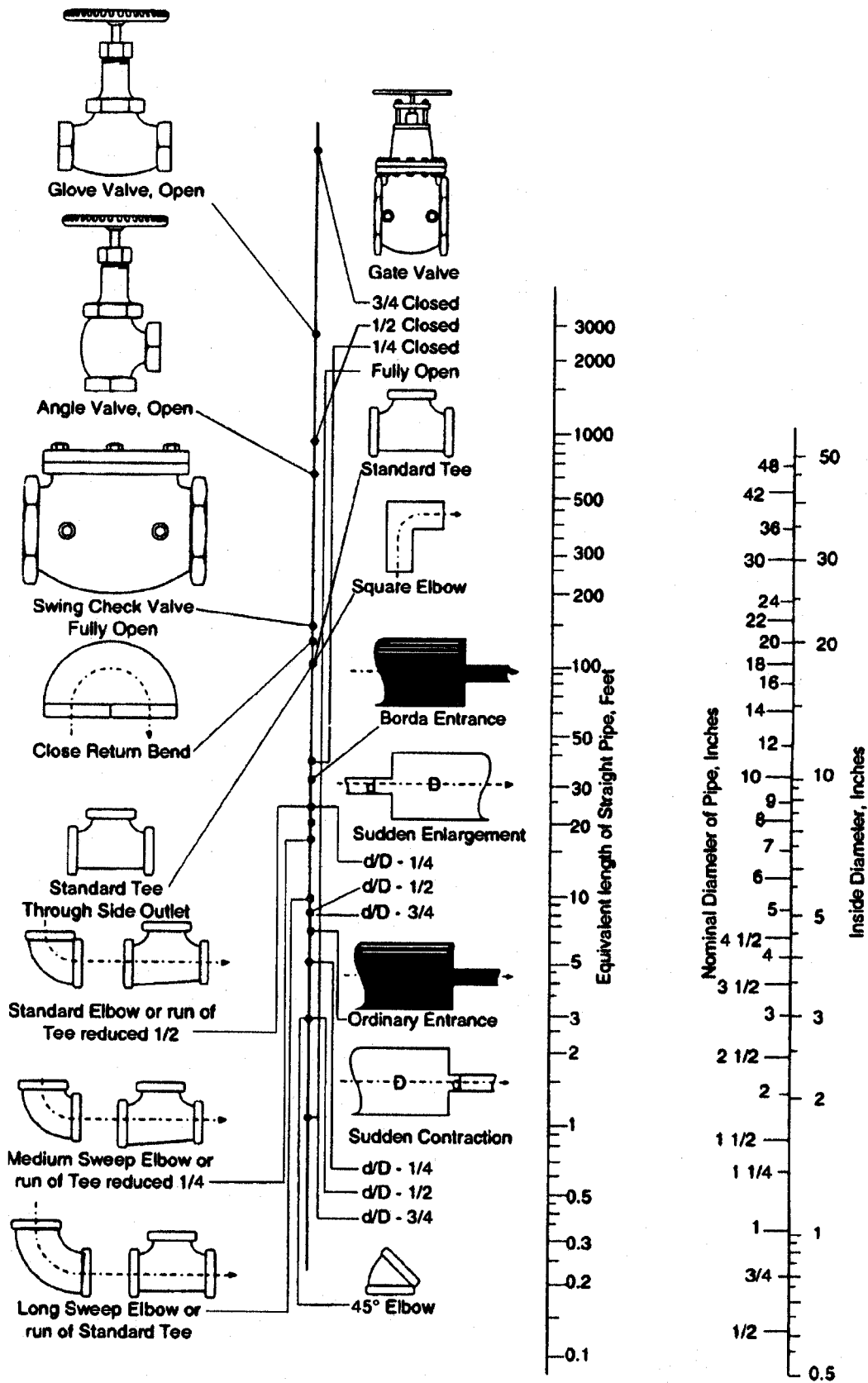
## Minor Head Loss

For determination of minor head loss, a standard nomograph is often used, with which each fitting is converted to an equivalent length of straight pipe of the same diameter. Head loss calculations using the new length of pipe will include both major and minor losses ([Fig. 91.4](#)).

Figure 91.3 Hazen-Williams alignment chart.



**Figure 91.4** Resistance of valves and fittings to flow of fluids.



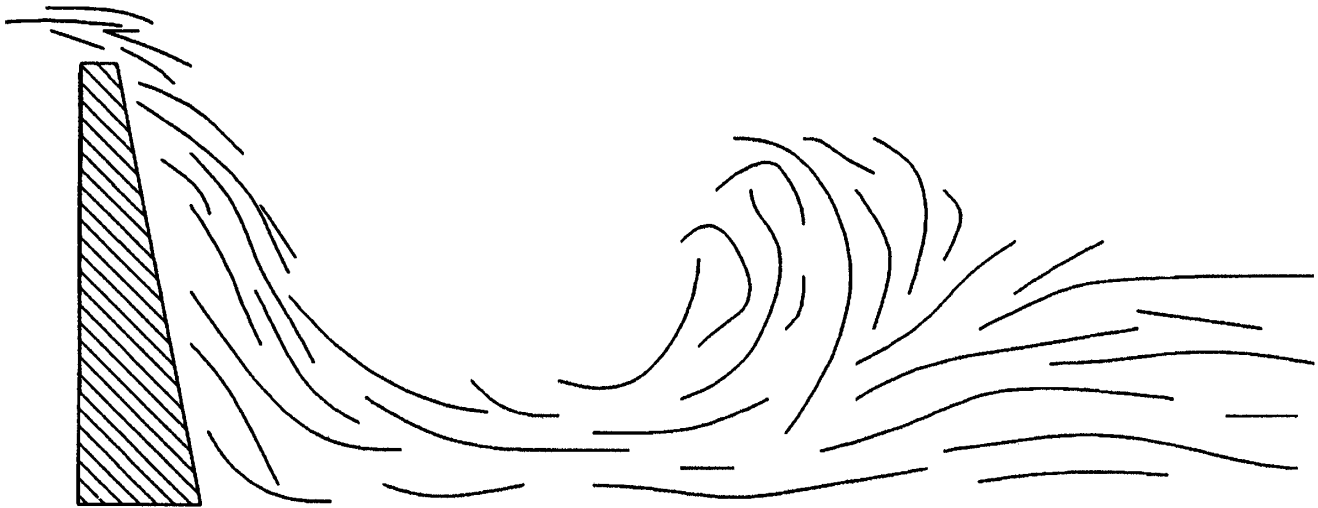
A few minor losses may be expressed in terms of velocity head ( $V^2/2g$ ). Useful in calculations where Bernoulli's theorem is in use and a velocity head value is readily at hand, this method directly converts to head loss (ordinary exit:  $HL = 1VH$  ; ordinary entrance:  $HL = 0.5VH$  ; Borda entrance:  $HL = 1VH$  ).

## 91.6 Open Channel Flow

---

In conduits where the water has a free surface exposed to atmospheric pressure, velocity head is the energy driving flow. The channel must be physically sloped enough to overcome friction losses so that velocity is maintained. In a properly sloped channel, at steady state flow, the water surface is parallel to the channel bottom and the hydraulic grade line follows. Energy loss is negated by the slope of the channel, and water depth remains constant throughout. In an open channel with a horizontal bottom, the water encounters friction and decreases in velocity, "piling up" behind; it produces the pressure head needed as a greater depth at the upstream end, and the slope of the water surface registers its progressive loss downstream. If the channel is sloped more than is necessary to overcome friction losses, velocity will increase and water depth will decrease; the steep slope creates extra velocity head. A stream bed that slopes sharply, then levels off, will carry water at a shallow depth where the slope is steep and velocity is high. Downstream, the water will be deeper, and the velocity slower (Fig. 91.5).

**Figure 91.5** Hydraulic jump.



In locations where waters of two different velocities meet, a short section of deeper water occurs; water level rises at the point of velocity change before the surface evens out again; there is extra turbulence at this point, and the water traps air and expands. Called *hydraulic jump*, the phenomenon occurs dramatically at dams and flumes, less so as shoreline waves or those created in a ship's wake, or as ripples that form when a stone is thrown into water.

*Manning's formula* is widely used for calculation of velocity or flow in open

channels:

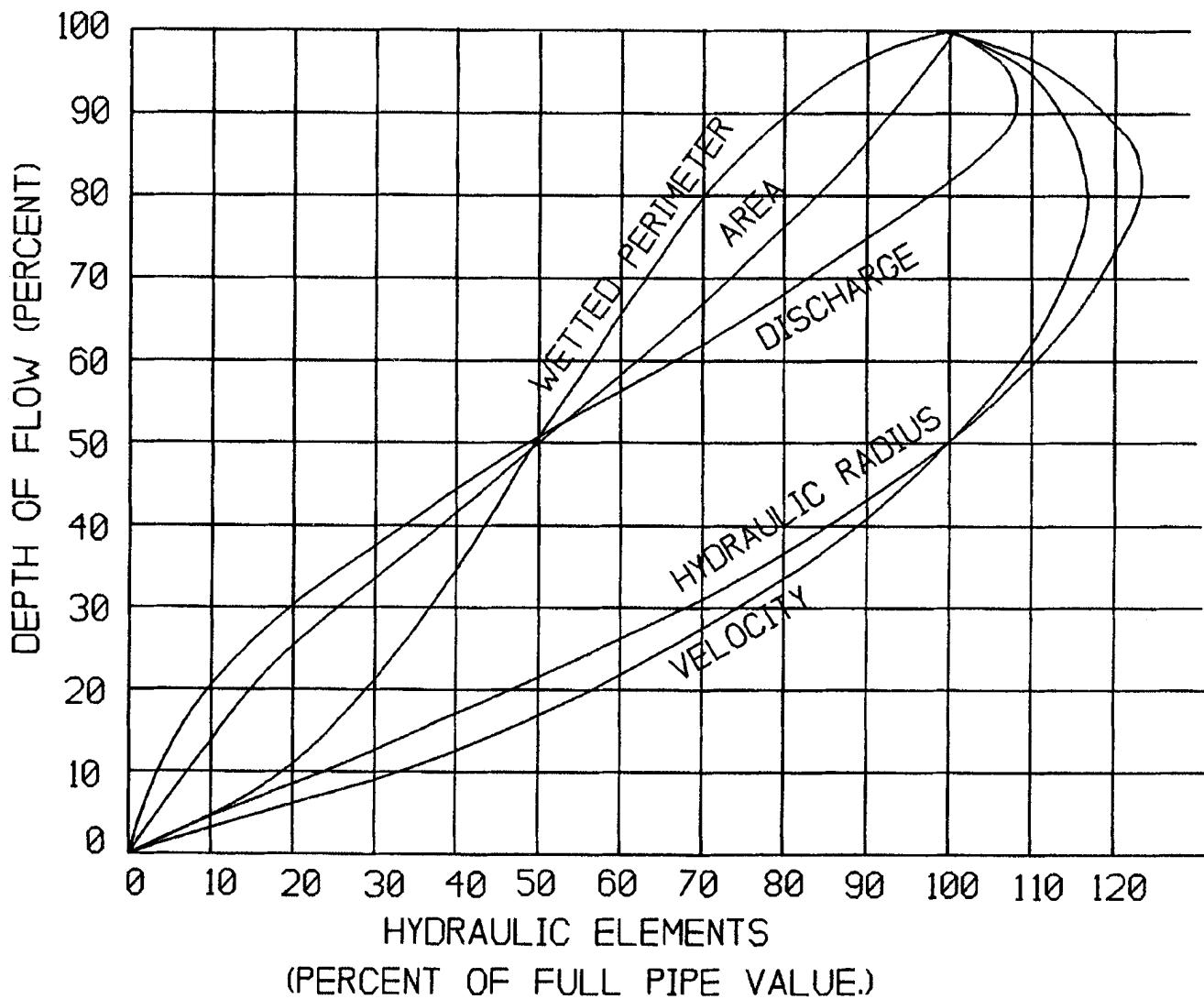
$$V = \frac{1.486}{n} \times R^{0.66} \times s^{0.5} \quad (91.9)$$

Manning's formula employs a roughness coefficient,  $n$ , which is specialized for materials of which open channels are constructed (Table 91.2). In this equation, diameter has been replaced by hydraulic radius ( $R$ ), making the formula flexible enough to adapt to all cross-sectional areas.  $R$  is a measure of the efficiency with which the conduit transmits water and is determined by dividing the cross-sectional area of the water by the wetted perimeter.

$$R = \frac{\text{Wetted area}}{\text{Wetted perimeter}} \quad (91.10)$$

For pipes less than full, wetted area and wetted perimeter are difficult to ascertain, and a hydraulic elements curve is an indirect but accurate method of obtaining the desired value based on its percentage of the full pipe value (Fig. 91.6).

**Figure 91.6** Hydraulic elements curve.



**Table 91.2** Manning Roughness Coefficient (*n*) Value

Type of Material		<i>n</i> Value
Pipe		
Cast iron:		
	Coated	0.012–0.014
	Uncoated	0.013–0.015
Wrought iron:		
	Galvanized	0.015–0.017
	Black	0.012–0.015
Steel:		
	Riveted	0.015–0.017
	Corrugated	0.021–0.026
Wood stave		0.012–0.013
Concrete		0.012–0.017
Vitrified		0.013–0.015
Clay, drainage tile		0.012–0.014
Lined Channels		
Metal:		
	Smooth semicircular	0.011–0.015
	Corrugated	0.023–0.025
Wood:		
	Planed	0.010–0.015
	Unplaned	0.011–0.015
Cement lined		0.010–0.013
Concrete		0.014–0.016
Cement rubble		0.017–0.030
Unlined Channels		
Earth:		
	Uniform	0.017–0.025
	Winding	0.023–0.030
	Stony	0.025–0.040
Rock:		
	Uniform	0.025–0.035
	Jagged	0.035–0.045
	Jagged	0.035–0.045

## 91.7 Flow Measurement

### Orifice Meter

An orifice meter is a flat steel plate with a precisely sized small diameter hole at the center that is installed between flanges in a pipeline; the pressure differential created across the orifice is measured by gages upstream and at the orifice discharge. Derived from Bernoulli's formula, calculation is based upon change in velocity head passing through the orifice:

$$Q = C_d A \sqrt{(PH_1 - PH_2) \times 2g} \quad (91.11)$$



$C_d$  = coefficient of discharge (0.6-0.9)

$A$  = area of the orifice ( $\text{ft}^2$ )

$\text{PH}_1$  = pressure head upstream

$\text{PH}_2$  = pressure head at discharge

## Venturi Meter

The Venturi meter is the most accurate and widely used closed conduit pressure differential meter—a constricted tube, with converging section, throat, and longer diverging outlet section. Gages are placed just upstream from the convergence and at the throat; flow is smooth through the unit, and head loss is minimal ( $C_d = 0.98$ ).

$$Q = C_d A \sqrt{\frac{(\text{PH}_1 - \text{PH}_2)}{1 - (d_2/d_1)^4} \times 2g} \quad (91.12)$$

## Pitot Gage

The Pitot gage is a flowmeter that relies on a direct measurement of velocity head. A pressure gage with a double sensing unit (or a differential manometer) is installed into the pipeline. One end is bent backwards into the flow to capture velocity head as well as pressure head. The differential reading is velocity head, which can be converted to flow.

## Magnetic Flowmeter

A magnetic flowmeter consists of a set of magnetic coils that surround the pipe, creating an electromagnetic field; an opposed pair of electrodes mounted at right angles register the induced voltage (converted to a current signal), which is directly proportional to the velocity of the water passing through the unit.

## Ultrasonic Meter

The transmissive type of ultrasonic meter sends ultrasonic beams through the pipe from opposite transmitter/receivers mounted at a diagonal to the flow stream; beam differential is directly proportional to water velocity. The reflective type sends a single sonic beam into the water from a transmitter mounted on the pipe; the beam bounces off solids in the water, and is picked up at a different frequency. The magnitude of frequency change is directly proportional to water velocity.

A variation sends a beam to the water surface from an overhead transmitter and relates the return time to water depth, which is convertible to flow.

## Positive Displacement Meter

Positive displacement meters are service meters suitable for residential customers; the unit consists of a measuring chamber enclosing a disk or piston; with each pulse a magnetic contact is made to a register that totalizes flow.

## Turbine Meter

Turbine meters consist of a measuring chamber with a rotor that turns in response to the velocity of the water. Large customer flows (hotels, industries) are recorded with turbine, or the more efficient "turbo" meters.

## Compound Meter

Compound meters are installed when accurate reading at both high and low flows is required for customer billing; this device consists of turbine meter on the main line and positive displacement meter on the bypass.

## Weir

The weir is the least costly open channel flow meter, a flat plate over which the flow passes. For a rectangular weir,

$$Q = 3.33 \times L \times h^{1.5} \quad (91.13)$$

where  $L$  is the width of weir and  $h$  is the head on weir. For a 90-degree V-notch weir,

$$Q = 2.5 \times h^{2.5} \quad (91.14)$$

## Parshall Flume

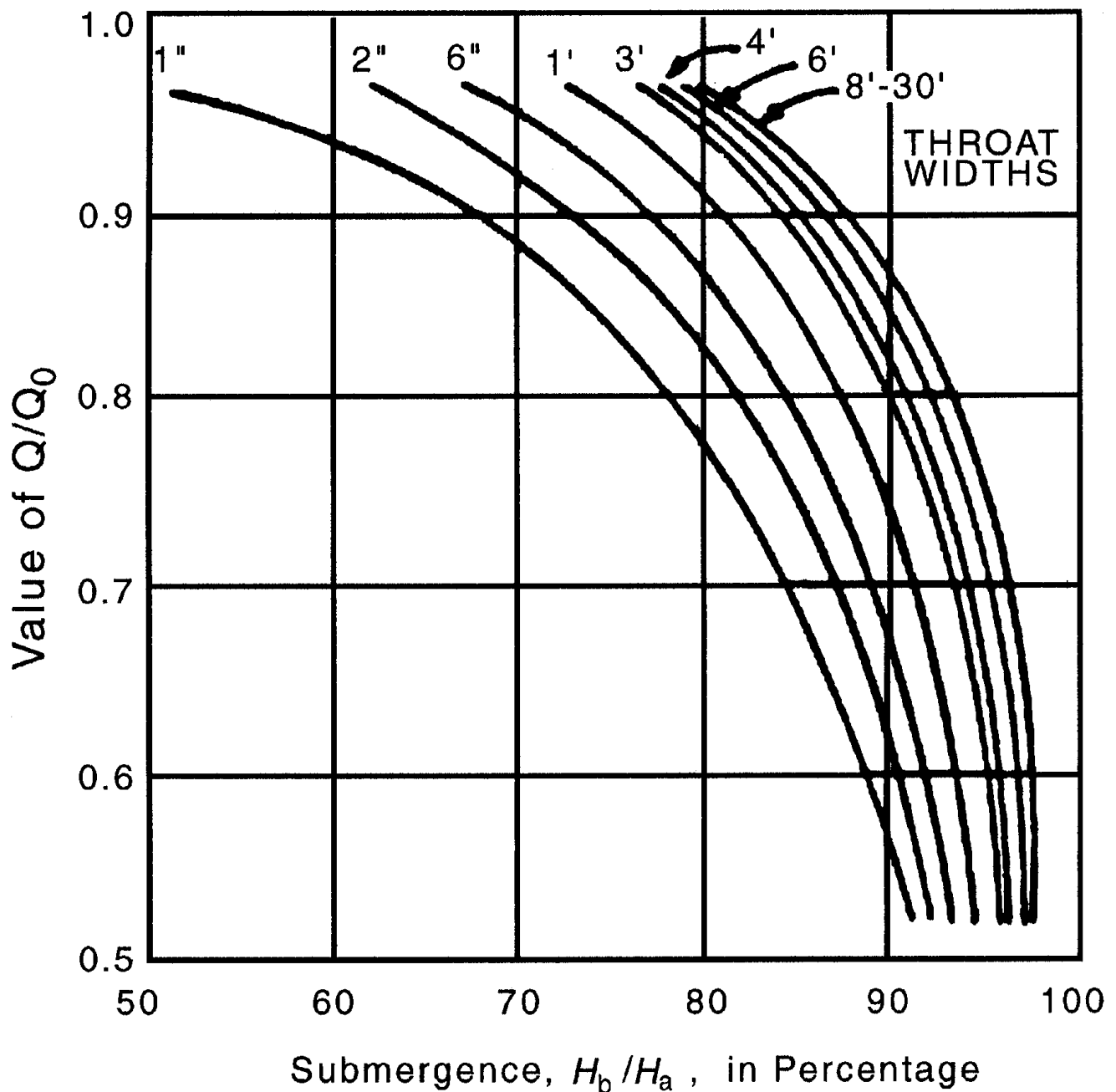
The Parshall flume is widely used for wastewater and irrigation water flow measurement due to low head loss; this device has self-cleansing capacity and ability to operate accurately over significant flow range and consists of inlet, downward inclining throat, and diverging outlet. Depth measurement is taken from a stilling well at inlet. The following formula applies to throat widths 1-8 ft and a medium range of flows:

$$Q = 4W \times H_a^{1.52} \times W^{0.026} \quad (91.15)$$

where  $W$  is the width of throat and  $H_a$  is the depth in the stilling well upstream.

Parshall flume formulas are based on low flows; at higher flows, a hydraulic jump forms at the outlet, which may submerge the throat, restricting flow. Formulas will not yield a true flow value; a stilling well at throat bottom measures downstream depth  $H_b$ , and with percent submergence the correction graph (Fig. 91.7) can be referred to.

**Figure 91.7** Corrections for Parshall flume.



## 91.8 Centrifugal Pump

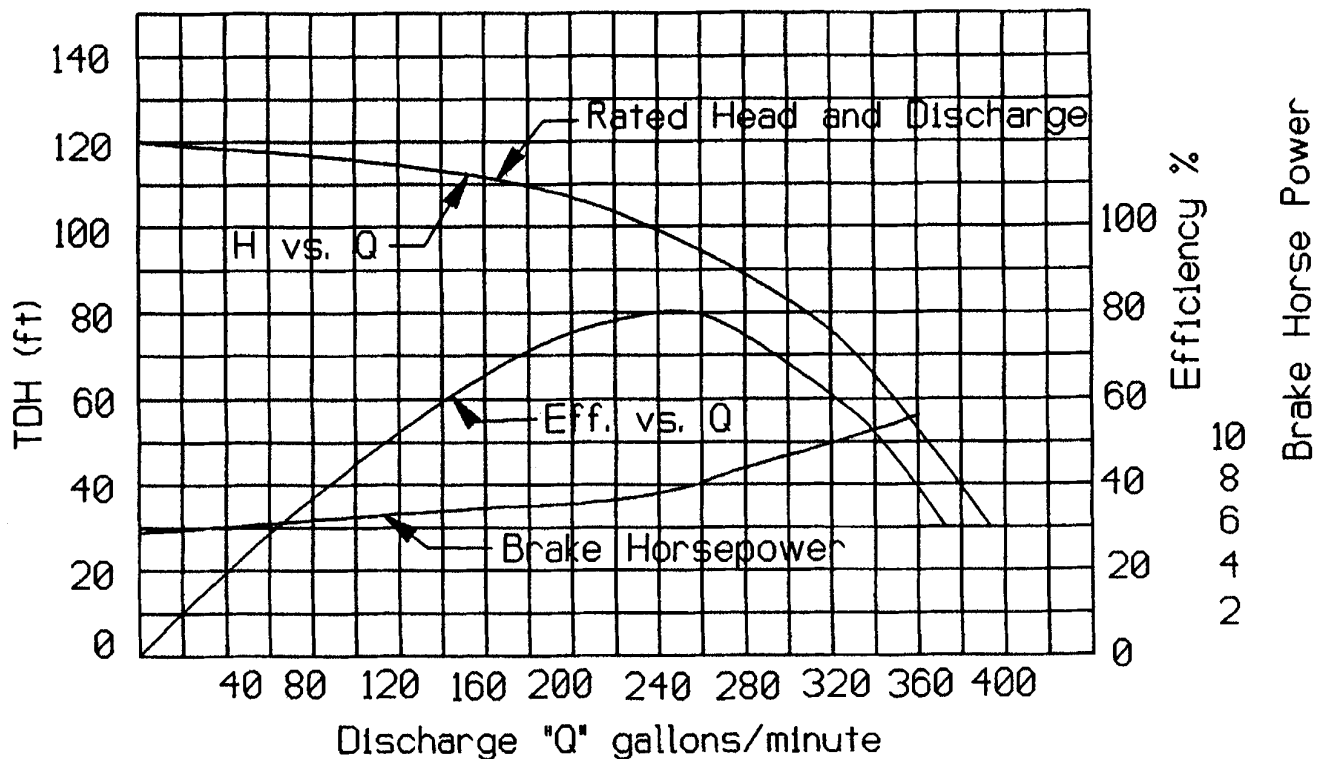
Centrifugal pumps are mechanical devices that convert other forms of energy to hydraulic energy; these pumps create the pressure needed for flow to occur. Pumping technology was limited to positive displacement and screw devices until the nineteenth century, when the centrifugal pump was developed. This device is a small, efficient unit employing a rapidly revolving impeller; high water velocity is developed, which is then converted to pressure upon exit. The pressure

developed, the horsepower required, and the resulting efficiency vary with the discharge. These values are diagrammed graphically for each pump on a pump characteristic curve (Fig. 91.8), which has been developed from the formula

$$\frac{\text{gpm} \times \text{TDH}}{3960} = \frac{\text{WHP}}{\text{Pump efficiency}} = \text{BHP} \quad (91.16)$$

where TDH is the total dynamic head (the work the pump must do in overcoming lift and losses to move the water), WHP is the water horsepower (the power needed to move the water), and BPH is the brake horsepower (the power which must be available to the pump).

**Figure 91.8** Pump characteristic curve.



Each pump is built to operate at its *design point*, the head and flow at which it achieves maximum efficiency. The point at which it does operate is dependent upon the characteristics of the system, the arrangement of the pipes and appurtenances through which the flow must pass. A centrifugal pump will perform according to its characteristic curve. Pump characteristics, however, may be changed by adjusting the pumping speed (change rpm or impeller size), and oak tree curves are designed to demonstrate characteristics using various sizes of impellers.

Pumps installed in series increase the pumping head, as in booster pumping (multiple pumps) or high-pressure or deep well pumps (multiple impellers). Pumps arranged in parallel increase the flow; the head remains that of one pump working.

For further information on incompressible fluids, refer to **Chapter 33**.

## Defining Terms

**Flow:** The quantity of water passing a point in a given unit of time (gpd, gpm, cfs).

**Force:** The total pressure registered on an entire surface area (lb).

**Head:** Pressure, registered as feet of water.

**Mechanical joint:** Bell and spigot type; has an outer follower that bolts to the flanged end.

**Momentum:** Mass multiplied by velocity; responsible for extent of hammer, thrust, surge.

**Nomograph:** A specialized chart with three or more components that may yield an answer with one or more of its dimensions unknown.

**Pressure:** Force exerted on a unit area (psf, psi).

**Suppressor:** Air chamber or open container with small orifice connection to pipeline; allows temporary exit of some water when pressures are high.

**Velocity:** Speed at which the water travels (ft/s).

## References

American Water Works Association. 1972. *Water Meters: Selection, Installation, Testing, Maintenance*. American Water Works Association, Denver, CO.

American Water Works Association. 1980. *Basic Science Concepts and Applications*. American Water Works Association, Denver, CO.

French, R. 1985. *Open Channel Hydraulics*. McGraw-Hill, New York.

Hauser, B. 1991. *Practical Hydraulics Handbook*. Lewis, Chelsea, MI.

Kanen, J. 1986. *Applied Hydraulics for Technology*. CBS College Publishing, New York.

Prasuhn, A. 1987. *Fundamentals of Hydraulic Engineering*. Holt, Rinehart & Winston, New York.

Walski, T. 1984. *Analysis of Water Distribution Systems*. Van Nostrand Reinhold, New York.

## Further Information

*The Journal of the American Water Works Association*

6666 W. Quincy Ave.

Denver, CO 80235

*Water Environment and Technology*, the journal of the Water Environment Federation

601 Wythe St.

Alexandria, VA 22314-1994

The Hydraulic Institute

712 Lakewood Center

N14600 Detroit Ave.

Cleveland, OH 44107

WATERNET, AWWA's database of the water and wastewater industries.

See AWWA's Computer Search Service, (303)347-6170.

Singh, V. P. "Hydrology"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

### 92.1 Classification of Hydrology

### 92.2 Hydrologic Cycle

### 92.3 Laws of Science

Surface Flow • Unsaturated Flow • Saturated Flow • Sediment Transport • Solute Transport • Microbial Transport • Initial and Boundary Conditions

### 92.4 Approaches to Hydrologic Problems

### 92.5 Tools for Hydrologic Analyses

### 92.6 Components of Hydrology Cycle—Deterministic Hydrology

Precipitation • Evaporation and Transpiration • Infiltration and Soil Moisture • Surface Runoff • Sediment Transport and Yield • Solute Transport • Microbial Transport • Models of Hydrologic Cycle

### 92.7 Statistical Hydrology

Empirical Analyses • Phenomenological Analyses • Stochastic Analyses

### 92.8 Hydrologic Design

## Vijay P. Singh

*Louisiana State University*

Hydrology can be defined as the science that deals with occurrence, movement, distribution, and storage of water in respect to both its quantity and quality over and below the land surface in space, time, and frequency domains. Water quantity encompasses the physical aspects, and water quality, the chemical and biological aspects. One might sum up hydrology as the study of water in all aspects at macro and higher scales.

The study of hydrology originated in the design of hydraulic works. This historical underpinning continues to dominate the scope and the range of hydrologic investigations. It is therefore no surprise that most often civil engineering is the home of hydrology. With the changing environmental landscape, however, there are signs of hydrology becoming a geophysical science in its own right.

## 92.1 Classification of Hydrology

---

It is instructive to peruse the various classifications of hydrology [Singh, 1993]. By definition, hydrology can be classified as physical hydrology, chemical hydrology, or biological hydrology; as water quantity hydrology or water quality hydrology; as surface-water hydrology or subsurface hydrology. Depending on the type of **watershed** for which the study of water is undertaken, it can be classified as agricultural hydrology, forest hydrology, urban hydrology, mountainous

hydrology, desert hydrology, wetlands hydrology, or coastal hydrology. Considering the form of water or where water occurs predominantly, this study can be classified as snow hydrology, ice and glacier hydrology, atmospheric hydrology, or lake hydrology. Depending on the particular emphasis on land phase or channel phase, it can be classified as watershed hydrology or river hydrology. Hydrology is also classified based on the tools employed for investigation of hydrologic systems. Parametric hydrology, theoretical hydrology, mathematical hydrology, statistical hydrology, probabilistic hydrology, stochastic hydrology, systems hydrology, and digital hydrology form this classification. The various classifications of hydrology are useful in that they point to its scope and the range of techniques employed in its study.

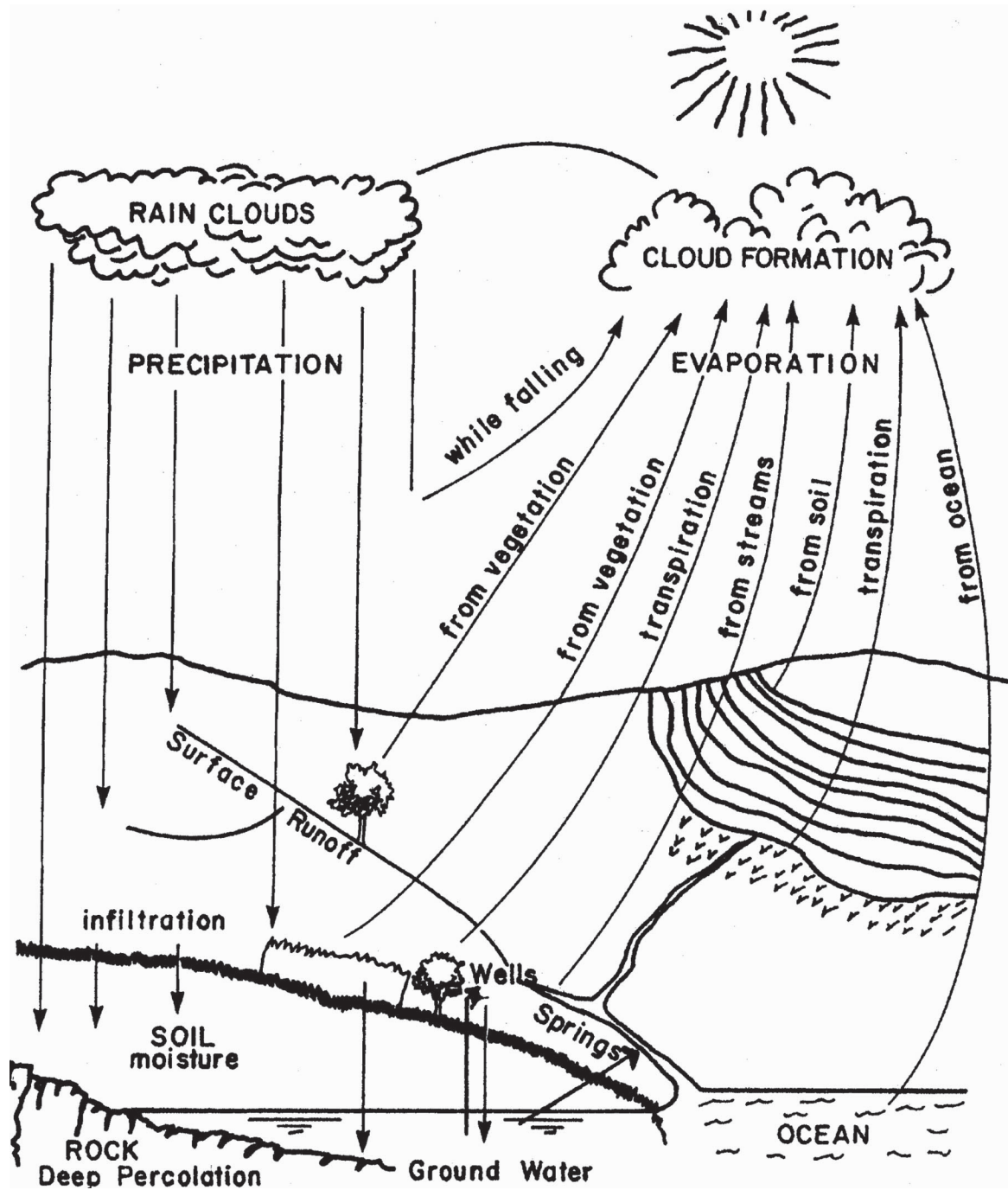
## 92.2 Hydrologic Cycle

---

Water originates in the atmosphere when water vapors are transformed into droplets forming precipitation that falls on the land surface. Part of this precipitation, which eventually returns to the atmosphere through evaporation, is intercepted during its fall by vegetative canopy, buildings, and so on. Another part, which may fall on water surfaces such as streams, lakes, ponds, and seas, may either run off, evaporate, or get stored and finally evaporate. The remainder fills in the depressions on the ground, meets the infiltrative demand of the soil, and runs off the ground to form stream flow. The infiltrated water percolates down and recharges the groundwater and may eventually become stream flow. The final destination of all streams is the ocean, so stream flows finally reach the ocean. Part of the oceanic water returns to the atmosphere through evaporation. Part of the infiltrated water as well as of the surface flow returns to the atmosphere through evapotranspiration. Thus the cyclic movement of water from the atmosphere through precipitation to the land, through stream flow to the ocean, and through evapotranspiration back to the atmosphere is designated the **hydrologic cycle**. This movement, of course, follows devious paths. The study of the hydrologic cycle can then be defined as hydrology. A sketch of the hydrologic cycle [[Ackermann \*et al.\*, 1955](#)], as shown in [Fig. 92.1](#), is meaningful in that it brings out the complexity as well as the challenge encountered in the study of water.



**Figure 92.1** A schematic of hydrologic cycle. (From Ackermann, W. C., Colman, E. A., and Ogrosky, H. O. 1955. From ocean to sky to land to ocean. In *U.S. Department of Agriculture Yearbook 1955*, pp. 41–51. USDA, Washington, DC.)



## 92.3 Laws of Science

---

The laws that govern the movement of water and the constituents it carries with it over and below the ground are the conservation of mass, momentum, and energy. The conservation of mass is expressed as a continuity equation, and that of momentum as an equation of motion. Depending on the type of flow, these equations are expressed in a variety of forms, as will be clear from the following discussion.

### Surface Flow

For simplicity, only the one-dimensional form of the governing equations using a control volume is given here. The continuity equation can be expressed as

$$\frac{\partial A}{\partial t} + \frac{\partial Q}{\partial x} = q(x, t) - i(x, t) - e(x, t) \quad (92.1)$$

the momentum equation as

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + g \frac{\partial h}{\partial x} = g(S_0 - S_f) - \frac{(q - i)(u - v)}{A} \quad (92.2)$$

and the energy equation as

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + g \frac{\partial h}{\partial x} = g(S_0 - S_f) + \frac{u - v(v/u)}{2A} q \quad (92.3)$$

where  $A$  is flow cross-sectional area,  $Q$  is discharge (volumetric rate =  $u \cdot A$ ),  $u$  is average flow velocity,  $h$  is depth of flow,  $S_0$  is bed slope,  $S_f$  is frictional slope,  $q$  is lateral inflow per unit length of flow,  $i$  is infiltration per unit length,  $e$  is evaporation rate and other abstractions per unit length,  $v$  is velocity of lateral inflow in the longitudinal direction,  $x$  is distance in the longitudinal direction, and  $t$  is time. Except for the term expressing the influence of lateral inflow or outflow, Eqs. (92.2) and (92.3) are equivalent.

Equation (92.1), in conjunction with either Eq. (92.2) or (92.3), can be employed to model **surface flows** on plains and/or in channels. Two popular approximations of Eq. (92.2) are the diffusion-wave and kinematic-wave approximations [[Lighthill and Whitham, 1955](#); [Dooge, 1973](#)], which can be expressed, respectively, as the following:

$$\frac{\partial h}{\partial x} = S_0 - S_f \quad (92.4)$$

$$S_0 = S_f \quad (92.5)$$

With use of a uniform flow formula such as Manning's or Chezy's,  $S_f$  can be expressed as

$$S_f = \beta \frac{u^2}{R^a} \quad (92.6)$$

where  $\beta = 1/C^2$  and  $a = 1$  for Chezy's equation;  $\beta = n_m^2$  and  $a = 4/3$  for Manning's equation;  $C$  is Chezy's roughness coefficient,  $n_m$  is Manning's roughness factor, and  $R$  is the hydraulic radius ( $=A/P$ ,  $P$ = wetted perimeter).

Substitution of Eq. (92.6) into Eq. (92.5) and the assumption that  $R$  and  $h$  are uniquely related leads to

$$u = \alpha h^m, \quad m > 0 \quad \text{or} \quad Q = \alpha h^n, \quad n = m + 1 \quad (92.7)$$

where  $m = 0.5$  and  $\alpha = C(S_0)^{0.5}$  for Chezy's equation, and  $m = 2/3$  and  $\alpha = (S_0)^{0.5}/n_m$  for Manning's equation.

The kinematic-wave approximation hypothesizes a unique relationship between the flux (average velocity), concentration (depth), and position. Thus, this approximation can also be expressed in forms different from Eq. (92.7), as shown by Beven [1979]. If control volume is extended to the scale of a watershed or a channel segment, then the flow variables are lumped or integrated in space and only their temporal variability is retained. Thus, integration of Eq. (92.1) in space leads to

$$\frac{dS}{dt} = Q - I(t) - f(t) - E(t) \quad (92.8)$$

where

$$\begin{aligned} S &= \int_{x_1}^{x_2} A \, dx, & Q &= Q(x_2, t), & I &= Q(x_1, t) + \int_{x_1}^{x_2} q \, dx, \\ f &= \int_{x_1}^{x_2} i \, dx, & E &= \int_{x_1}^{x_2} e \, dx \end{aligned}$$

Equation (92.8) is a volume balance or water budget equation with two unknowns,  $S$  and  $Q$ . Its solution requires another equation relating  $S$  to  $Q$ ,  $I$ , and/or other variables. A very general relation between  $S$  and  $I$  and  $Q$  is

$$S = \sum_{j=0}^M a(Q, I) \frac{d^j Q}{dt^j} + \sum_{i=0}^N b(Q, I) \frac{d^i I}{dt^i} \quad (92.9)$$

where  $a$  and  $b$  are coefficients, and  $M$  and  $N$  are some integers. A special case, involving one of the most frequently used relations in hydrology, is  $S = S(Q)$  :

$$S = KQ, \quad S = kQ^\beta \quad (92.10)$$

where  $K$  is the storage parameter (lag time for  $\beta = 1$ ), and  $k$  and  $\beta$  are parameters.

Since Eq. (92.8) is derived from Eq. (92.1), Eq. (92.10) can be derived from the momentum equation. As an example, consider Eq. (92.7) with  $n = 1$ . By multiplying both sides by  $\Delta x = x_2 - x_1$  and recalling that  $S = \Delta x \cdot h \cdot 1$  and  $Q$  is volumetric flow rate, Eq. (92.10) results immediately.

## Unsaturated Flow

In an unsaturated porous medium, part of the pore space is occupied by air, so the degree of saturation is to be taken into account in dealing with unsaturated flow. The moisture content  $\theta$  in the medium (volume of water per unit volume of porous medium) is a function of the capillary pressure  $\psi < 0$ , and likewise is the medium's hydraulic conductivity  $K(\psi)$ . The basic governing equations for **unsaturated flow** are the continuity equation and a flux law given by Darcy's equation in lieu of the momentum equation. This flux law can also be derived from energy conservation considerations. The three-dimensional continuity equation, under the assumption of incompressible water, can be written as

$$\frac{\partial q_x}{\partial x} + \frac{\partial q_y}{\partial y} + \frac{\partial q_z}{\partial z} = -\frac{\partial \theta}{\partial t} \quad (92.11)$$

and Darcy's equation as

$$q_s = -K_s(\psi) \frac{\partial h}{\partial s}, \quad s = x, y, z; \quad \vec{q} = \{q_x, q_y, q_z\} \quad (92.12)$$

where  $h$  is the hydraulic head and  $q_s$  is the flux in the  $s$  direction. Substituting Eq. (92.12) into Eq. (92.11) and recalling that  $h = \psi + z$ , one gets

$$\frac{\partial}{\partial x} \left( K_x(\psi) \frac{\partial \psi}{\partial x} \right) + \frac{\partial}{\partial y} \left( K_y(\psi) \frac{\partial \psi}{\partial y} \right) + \frac{\partial}{\partial z} \left( K_z(\psi) \frac{\partial \psi}{\partial z} + K_z(\psi) \right) = C(\psi) \frac{\partial \psi}{\partial t}$$

$$C(\psi) = \frac{\partial \theta}{\partial \psi} \quad (92.13)$$

where  $C(\psi)$  is the specific moisture capacity. This is the well-known Richards equation [Richards, 1931]. Based on simplifications of porous media properties (anisotropy and heterogeneity) and the nature of flow, a number of simpler versions can be derived. On the other hand, if the control volume is extended to a soil element, then spatially lumped equations can be derived. For example, Eq. (92.11) can be integrated over space and expressed in the form of a water balance equation as

$$\frac{dS(t)}{dt} = f_s(t) - f(t) \quad (92.14a)$$

where  $S(t)$  is the potential water storage space in the soil element,  $f_s(t)$  is the seepage rate from the element, and  $f(t)$  is the infiltration rate. If the initial storage space available in the element is  $S_0$ , then the amount of water storage at any time  $t$  is

$$W(t) = S_0 - S(t) = \int_0^t [f(t) - f_s(t)] dt \quad (92.14b)$$

which is an integral expression of continuity. Another relation in lieu of Eq. (92.12) can be expressed [Singh and Yu, 1990] as

$$f(t) = f_s(t) + \frac{a[S(t)]^m}{[S_0 - S(t)]^n} \quad (92.15)$$

where  $a$ ,  $m$ , and  $n$  are positive real constants.

## Saturated Flow

The governing equations for **saturated flow** are the continuity equation and the flux law specified by Darcy's equation. A three-dimensional form of continuity equation for incompressible flow is

$$\frac{\partial q_x}{\partial x} + \frac{\partial q_y}{\partial y} + \frac{\partial q_z}{\partial z} = -S_s \frac{\partial h}{\partial t} \quad (92.16)$$

where  $S_s$  is the specific storage for confined formations, or specific yield divided by the saturated thickness for unconfined formations. Darcy's equation can be written as

$$q_s = -K_s \frac{\partial h}{\partial s}, \quad s = x, y, z; \quad \vec{q} = \{q_x, q_y, q_z\} \quad (92.17)$$

where  $K_s$  = the saturated hydraulic conductivity in the  $s$  direction. Substitution of Eq. (92.17) into Eq. (92.16) gives the general flow equation, which specializes—depending on the simplifications of porous media properties and the nature of flow—into a number of equations, such as the Laplace equation, the diffusion equation, the Theis equation, the Poisson equation, the Boussinesq equation, and so on.

## Sediment Transport

The governing equations for transport of suspended sediment by convection and turbulent diffusion under gravity are the conservation of mass for sediment and the shallow-water equations of momentum and mass conservation for sediment-laden water. The latter two equations are Eqs. (92.1) and (92.2). The three-dimensional form of sediment mass conservation can be expressed as

$$\begin{aligned} \frac{\partial C}{\partial t} + u \frac{\partial C}{\partial x} + v \frac{\partial C}{\partial y} + w \frac{\partial C}{\partial z} = w_s \frac{\partial C}{\partial z} + \frac{\partial}{\partial x} \left( \varepsilon_x \frac{\partial C}{\partial x} \right) \\ + \frac{\partial}{\partial y} \left( \varepsilon_y \frac{\partial C}{\partial y} \right) + \frac{\partial}{\partial z} \left( \varepsilon_z \frac{\partial C}{\partial z} \right) \end{aligned} \quad (92.18)$$

where  $C$  is concentration of sediment by volume;  $u$ ,  $v$ , and  $w$  are velocity components in  $x$ ,  $y$ , and  $z$  directions;  $w_s$  is particle fall velocity; and  $\varepsilon_s$  is turbulent diffusion coefficient for sediment particle in the  $s$  direction ( $s = x, y, z$ ). A number of simplifying assumptions are often made regarding the flow, which gives rise to simplifying sediment transport models. For example, the flow is frequently assumed one-dimensional, and  $\varepsilon_s$  is considered constant or independent of the direction.

## Solute Transport

The governing equations for solute transport are the conservation of solute mass and flux laws and the shallow water equations of conservation of mass and momentum of flow containing solute. For expressing the solute mass conservation, advection, diffusion, and dispersion fluxes; adsorption and desorption; and loss and gain of solute have to be expressed. If the medium is unsaturated with moisture content  $\theta$ , then the solute-mass conservation can be expressed as

$$\frac{\partial}{\partial x} \left( \theta D_h \frac{\partial C}{\partial x} - qC \right) - \frac{\partial}{\partial t} (\theta C + \rho_s S) = \mu_w \theta C + \mu_s \rho_s S - \gamma_w \theta - \gamma_s \rho_s \quad (92.19)$$

where  $C$  is the solute concentration,  $q$  is Darcy's flux of water,  $\rho_s$  is the porous media bulk density,  $\mu_w$  is the rate constant for first-order decay in the liquid phase,  $\mu_s$  is the rate constant for first-order decay in the solid phase,  $\gamma_w$  is the rate constant for zero-order production in the liquid phase,  $\gamma_s$  is the rate constant for zero-order production in the solid phase,  $S$  is the adsorbed concentration, and  $D_h$  is the coefficient of hydrodynamic dispersion. Depending on the nature of solute and flow, a number of models can be derived from simplifications of Eq. (92.19).

## Microbial Transport

The fate of microorganisms in the subsurface environment depends on their survival and retention by soil particles. The governing equations for transport and retention of microorganisms are obtained from the conservation of microbial particles in porous media [Kommalapati *et al.*, 1991–92]. The first governing equation is for the deposited particles, which can be written as

$$\frac{\partial(\rho\sigma)}{\partial t} = K_c(\theta C) - k_d(\rho\sigma) - R_{dd} + R_{gd} \quad (92.20)$$

where  $R_{dd}$  and  $R_{gd}$  are decay and growth terms of the deposited microbes,  $\rho$  is density of microbial particles,  $\sigma$  is volume of deposited bacteria per unit volume of bulk soil,  $C$  is

concentration of suspended microbial particles per unit volume of flowing suspension, and  $\theta$  is effective porosity.

The second governing equation is the mass balance equation, including growth and decay terms:

$$\frac{\partial(\rho\sigma)}{\partial t} + \frac{\partial(\theta C)}{\partial t} = -\nabla[-\theta D \nabla C + \theta C(v_f + v_g + v_m)] + [\theta C + \rho\sigma](\mu - b) \quad (92.21)$$

where  $D$  is the coefficient of hydrodynamic dispersion,  $v_f$  is superficial longitudinal velocity of flow,  $v_g$  is settling velocity,  $v_m$  is migration velocity,  $\mu$  is specific growth term, and  $b$  is specific decay rate.

The third governing equation is derived from the mass conservation for the organic matter:

$$\frac{\partial(\rho_s S_F)}{\partial t} + \frac{\partial(\theta C_F)}{\partial t} = -\nabla(-D_e \theta \nabla C_F) + \theta v_f C_F - \frac{\mu}{Y}(\theta C + \rho\sigma) \quad (92.22)$$

where  $\rho_s$  is the bulk mass density of dry soil,  $S_F$  is the mass of adsorbed substrate per unit mass of soil particles,  $C_F$  is the mass of the substrate per unit volume,  $D_e$  is effective diffusivity coefficient, and  $Y$  is the true yield coefficient.

Depending upon the nature of flow, properties of porous media, and characteristics of microorganisms, a number of simplifications of Eqs. (92.20) through (92.22) have been made, which then form the basis of simpler models.

## Initial and Boundary Conditions

In hydrology, all three types of partial differential equations are involved. For example, Eqs. (92.1) and (92.2) are nonlinear hyperbolic; Eqs. (92.11) and (92.12) are nonlinear parabolic; and Eqs. (92.16), with unsteady state term dropped, and (92.17) are nonlinear elliptic. Initial and boundary conditions are needed to obtain a unique solution to a given problem. As an example, for surface flow the usual initial condition is one of dry surface or zero flow, the upstream boundary condition is one of zero discharge, and the downstream boundary condition is specified based on the type of flow or the existence of a control. For subcritical flow in a channel, it may be given by a known control; for supercritical flow another upstream boundary is specified. In a similar manner, conditions are to be specified for unsaturated flow, saturated flow, sediment transport, solute transport, or biological transport.

## 92.4 Approaches to Hydrologic Problems

Hydrologic systems are analyzed in one or more of three domains: (1) time, (2) space, and (3) frequency. They are also analyzed at different space-time scales. Most of the approaches developed in hydrology can be classified on the basis of domains and scales. Let us consider a hydrologic variable  $Y$  at any location  $(x, y, z)$  as a function of time. Then one can write

$$Y(t) = \overline{Y}(t) + \varepsilon(t) \quad (92.23)$$

where  $\bar{Y}(t)$  represents the mean value of  $Y$ , and  $\varepsilon$  the fluctuations around the mean. If  $Y$  is entirely deterministic, then  $\varepsilon$  vanishes and any approach for determination of  $Y$  is deterministic. If  $Y$  is entirely probabilistic, then  $Y$  is completely specified through modeling of  $\varepsilon$  and the approach is entirely probabilistic. If  $Y$  is part deterministic and part stochastic, the approach employed for determination of  $Y$  is mixed. Determination of  $\bar{Y}$  constitutes the subject matter of deterministic or mathematical hydrology, and that of  $\varepsilon$  the subject matter of statistical hydrology. All the deterministic approaches are either empirical, systems based, or physically based. Empirical approaches are based on data, systems approaches are based on the volume balance equation and a type of storage-discharge relation, and physically based approaches are based on the continuity equation together with an equation of motion or energy.

Methods for description of  $\varepsilon(t)$  are statistical, probabilistic, or stochastic. Statistical methods are mostly empirical and yield certain moment characteristics or descriptors such as mean, variance, skewness, and so on. Probabilistic approaches employ certain axioms, conceptually or physically based, and proceed to derive the probabilistic structure of  $\varepsilon$ . The stochastic approaches, on the other hand, can be likened to systems approaches and strive to describe the entire time series of  $\varepsilon$  without, however, deriving its probabilistic structure. These approaches constitute the subject matter of statistical hydrology. Frequently, the terms *statistical*, *probabilistic*, and *stochastic* are used interchangeably in hydrology. In this chapter the term *statistical* will be used to mean all three types.

## 92.5 Tools for Hydrologic Analyses

---

With continuing evolution of hydrology and its expanding role in environmental studies, a greater range of scientific, mathematical, and statistical tools are becoming increasingly important. In addition to physical, chemical, and biological training, hydrologic analyses require a good level of training in mathematics (at the level of partial differential equations and finite element method), statistics (at the level of stochastic processes, time series analysis, reliability analysis, spectral analysis, and multivariate analysis), and operations research (at the level of dynamic programming). Furthermore, to take full advantage of this knowledge, hydrologists of today have to be conversant with GIS, computer graphics, computer languages, and word processing. Methods of laboratory and field experimentation will receive increasing attention in the years to come. Hydrologic concepts and theories will have to be based more and more on what actually transpires in the field.

## 92.6 Components of Hydrology Cycle¾Deterministic Hydrology

---

### Precipitation

Precipitation forms input to hydrologic systems. It greatly varies in space and time and its space-time structure is highly random, especially at small time scales. The spatial and temporal variability of precipitation is sampled by a network of gages. These must be optimally located. An optimum number of rain gages  $N$  corresponding to an assigned percentage error  $\varepsilon$  in estimation of



mean areal rainfall can be obtained as

$$N = \left( \frac{C_v}{\varepsilon} \right)^2, \quad C_v = \frac{100S_p}{\bar{P}} \quad (92.24)$$

where  $C_v$  is the coefficient of variation of the rainfall values of the gages,  $S_p$  is the standard deviation of rainfall values, and  $\bar{P}$  is the mean of rainfall values.

Precipitation measurements are often inconsistent and incomplete. Methods for checking inconsistency are either graphical or statistical [Buishand, 1982,1984]. One of the popular methods for correcting it is the double mass curve [Singh, 1989]. Statistical methods include the von Neumann ratio test, cumulative deviations, likelihood ratio test, Bayesian tests, and run test. A multitude of methods exist for filling in the missing values, including the normal ratio method, arithmetic average, isohyetal method, the inverse distance method (IDM), ratio method, linear programming, and so on [Kruizinga and Yperlaan, 1978; Tung, 1983]. The IDM is quite popular and is based on the assumption that the dependence between any two gages is directly proportional to the inverse of some power (between 1.5 and 2.0) of the distance between them. Thus, the missing precipitation value at a gage  $x$  for a given time interval can be computed as

$$P(x) = \sum_{i=1}^m w_i P_i, \quad w_i = \frac{1/(D_i^a)}{\sum_{i=1}^m 1/(D_i^a)} \quad (92.25a)$$

where  $P_i$  is the precipitation at the  $i$  th gage,  $D_i$  is the distance between the gages  $x$  and  $i$ ,  $m$  is the number of gages used in filling (usually between three and five), and  $w_i$  is the weight assigned to the  $i$  th gage.

For hydrologic modeling, mean areal precipitation  $\bar{P}$  is often needed and is obtained in a variety of ways [Singh and Birsoy, 1975; Singh and Chowdhury, 1986], including unweighted mean, grouped area-aspect weighted mean, Thiessen polygons, individual area-altitude weighted mean, triangular area weighted mean, Myers method, isohyetal method, trend surface analysis, reciprocal distance-squared method, two-axis method, double Fourier series, modified polygon method, finite element method, analysis of variance, and kriging. These methods can be expressed as

$$\bar{P} = \sum_{i=1}^N a_i P_i \quad (92.25b)$$

where  $P_i$  is the precipitation value at the  $i$  th gage,  $N$  is the number of gages, and  $a_i$  is the weight assigned to the  $i$  th gage. The various methods differ in computation of  $a_i$  values. For example,  $a_i = 1/N$  for the arithmetic average, and  $a_i = A_i/A$  for the Thiessen polygon method, where  $A_i$  is the area of the polygon surrounding the  $i$  th gage and  $A$  is the watershed area. For monthly and yearly values all of these methods yield comparable results, but for short time intervals (e.g., hourly), isohyetal-type methods may be preferable.

An estimate of the mean areal rainfall is prone to random and systematic errors. Random errors

are caused by storm characteristics, rain gage density, and the representativeness of gages, whereas systematic errors are due to improper siting, poor exposure, and change in observer and in gage. Statistical methods for estimating the error in  $\bar{P}$  have been reported by Zawadzki [1973] and Bras and Rodriguez-Iturbe [1976].

## Evaporation and Transpiration

Evaporation is an energy exchange process through which water is transformed to vapor, and transpiration is the process by which plants transpire water. These are the only processes by which water is returned to the atmosphere to sustain the hydrologic cycle. Evaporation from a water body differs from that from land only if the latter has limited water that may be insufficient to satisfy the evaporative demand. The evaporation occurring from water bodies is referred to as *potential evaporation* (PE). If the soil has limited water and evaporative demand is not fully satisfied, then the evaporation occurs at a rate less than the potential and is called *actual evaporation* (AE).

Evaporation from a water body depends upon atmospheric conditions such as temperature, pressure, humidity, radiation, sunshine, wind velocity, and so on. As a result, a number of methods are available for estimating evaporation, some based on temperature, some on radiation, some on humidity, and some on combinations of the controlling factors [Jensen *et al.*, 1990]. Perhaps the best known is the Penman-Monteith method [Penman, 1948; Monteith, 1981]. This method combines the mass-transfer and energy balance approaches, and can be expressed as

$$E = \frac{\Delta}{\Delta + \gamma}(R_n + G) + \frac{\gamma}{\Delta + \gamma}E_a \quad (92.26)$$

where  $\Delta$  is the slope of the saturation vapor pressure curve for water,  $\gamma$  is the psychrometric constant,  $R_n$  is the net radiation,  $G$  is the sensible heat flux, and  $E_a$  is the evaporation due to water vapor saturation deficit at some height. Monteith [1981] incorporated aerodynamic and canopy resistance by modifying  $\gamma$ . Allen [1985] presents an account of several variants of the Penman method.

In addition to limiting soil moisture, evaporation from croplands is also affected by crop characteristics. In this case, evaporation from the soil and transpiration are combined to form *evapotranspiration* (ET). This phenomenon is also commonly referred to as *consumptive use*. One of the popular methods for its determination is the Jensen-Haise method [Jensen and Haise, 1963], expressed as

$$ET = C_T(T - T_x)R_s \quad (92.27)$$

where  $C_T$  is the temperature constant = 0.014 and  $T_x$  is the intercept axis = 26.5 for  $T$  in  $^{\circ}\text{F}$ ; and  $C_T = 0.025$  and  $T_x = -3$  for  $T$  in  $^{\circ}\text{C}$ .

The Blaney-Criddle method [Blaney and Criddle, 1962] is also quite popular for computing ET:

$$ET = KF = \sum kf, \quad f = \frac{TP}{100} \quad (92.28)$$

where  $ET$  is the consumptive use in inches of water during the growing season,  $K$  is the seasonal consumptive use coefficient for a crop,  $F$  is the sum of monthly consumptive use factors,  $T$  is the mean monthly air temperature in  $^{\circ}F$ ,  $P$  is the mean monthly percentage of daytime hours, and  $k$  is the monthly consumptive use coefficient.

Extraction of water by plants is limited by the availability of soil moisture. Holmes and Robertson [1959] showed that the ratio of AE to PE varies with the drying of soil and that the nature of this variation depends upon the type of soil and vegetation as well as drying rate. Thus, PE is modulated to account for soil moisture stress for determining AE [Singh and Dickinson, 1975].

## Infiltration and Soil Moisture

The process by which water enters into the soil at its surface is called *infiltration*. The subsequent downward movement of water is referred to as *percolation*. If water availability is not the limitation, then water will infiltrate the soil at the maximum rate, called *infiltration capacity*,  $f_p$ . Under water ponding, the soil's infiltration capacity declines exponentially in time. If availability of water is limited, then the infiltration rate,  $f$ , is less than the capacity rate. A number of factors affecting infiltration include soil characteristics, land use, vegetative cover, and rainfall characteristics.

A number of models for computing  $f_p$  are available. The physically based models are based on Eqs. (92.11) and (92.12), the conceptual models on Eqs. (92.14a) and (92.15), and empirical models on data. Singh and Yu [1990] showed that a number of empirical models could be derived using the systems-theoretic framework based on Eqs. (92.14a) and (92.15) and that these are actually conceptual models. Examples of some well-known conceptual models are the Philip two-term model [Philip, 1969],

$$F = At + st^{0.5} \quad (92.29)$$

the Green-Ampt model [Green and Ampt, 1911],

$$Kt = F - \eta S \ln \left( \frac{\eta S + F}{\eta S} \right) \quad (92.30)$$

and the Horton model [Horton, 1940],

$$F = f_c t + \frac{1}{k} (f_0 - f_c) [1 - \exp(-kt)] \quad (92.31)$$

where  $F$  is cumulative infiltration,  $t$  is time,  $A$  is the coefficient  $\simeq$  saturated hydraulic conductivity,  $s$  is sorptivity,  $\eta$  is wettable porosity,  $S$  is the capillary section at the setting front,  $K$  is hydraulic conductivity,  $f_0$  is the initial infiltration rate,  $f_c$  is the steady infiltration rate, and  $k$  is a constant depending upon the soil type and initial condition. By coupling these models with a rainfall event, the actual infiltration can be determined for that event.

The downward movement of water permits determination of soil moisture. This moisture is evapotranspired, so the status of soil moisture is predicted with the use of an appropriate evapotranspiration model.

## Surface Runoff

Surface runoff originates on the land surface and includes both overland flow and channel flow. Two fundamental problems in surface runoff are its time distribution for a specified rainfall event and the amount of surface runoff (also called *yield*) generated from it. A great deal of attention has historically been directed at these two problems, principally because of their ubiquitous application in the design of hydraulic works.

### Surface Runoff Hydrograph

The physically based models of the hydrograph are based on Eqs. (92.1) and (92.2) or (92.4) and (92.5). The most popular models for overland flow are based on the kinematic-wave approximation and those for channel flow on diffusion-wave approximation [Singh, 1990]. For realistic field conditions, equations are solved numerically [Liggett and Woolhiser, 1967].

The conceptual models most commonly employed in hydraulic design are based on Eqs. (92.8) and (92.9). The fundamental assumptions made in these models are that rainfall and infiltration are combined, forming excess rainfall, and that the watershed may be represented fictitiously through a network of reservoirs and/or channels, or through a geomorphologic network of average channels and cumulative overland areas [Singh, 1988]. If the watershed is assumed linear, then it is sufficient to compute the *instantaneous unit hydrograph* (IUH) for a delta-function excess rainfall, which, for a watershed represented by a linear reservoir [Dooge, 1959], is

$$h(t) = \frac{1}{k} \exp(-t/k) \quad (92.32)$$

where  $k$  is the reservoir lag time. For any excess rainfall the surface runoff hydrograph is then given by the convolution integral:

$$Q(t) = \int_0^t I(\tau)h(t-\tau)d\tau \quad (92.33)$$

If the watershed is represented by a series of  $n$  equal reservoirs each with lag time  $k$ , as done by Nash [1957], then

$$h(t) = \frac{1}{K\Gamma(n)} \left(\frac{t}{k}\right)^n \exp(-t/k) \quad (92.34)$$

where  $\Gamma(n)$  is the gamma function, with  $n$  as its argument. Convolution of this IUH with an excess rainfall yields the runoff hydrograph.

## Surface Runoff Volume

Determination of runoff volume for a given event is rather complicated, for the determination of the exact amount of infiltration and other abstractions in a watershed has been elusive. One of the popular methods for estimating storm runoff from agricultural areas is the SCS–curve number (SCS-CN) method [Soil Conservation Service, 1971]. This method is based on two main hypotheses: (1) The ratio of the actual amount of runoff to the potential amount of runoff is equal to the ratio of the actual amount of infiltration to the potential infiltration. (2) An initial amount of abstraction must be satisfied before commencement of any runoff, and that consists of interception, surface storage, and infiltration. The SCS-CN method can be written as

$$\frac{V_p - V_r - V_Q}{V_R} = \frac{V_Q}{V_p - V_r} \quad (92.35a)$$

where  $V_R$  is maximum possible retention,  $V_r$  is initial abstraction,  $V_p$  is amount of rainfall, and  $V_Q$  is amount of runoff. Note that  $V_r = aV_R$ ,  $a \simeq 0.1$  to  $0.2$ . The term  $V_R$  depends upon the characteristics of soil cover complex and antecedent soil moisture conditions. SCS expresses  $V_R$  as

$$V_r = \frac{1000}{C_N} - 10 \quad (92.35b)$$

where  $C_N$  is the curve number on a scale of 10 to 100. For computation of daily, weekly, monthly, or yearly runoff, water balance models are employed. Such models are also employed for computation of storm runoff. Reviews of such models have been presented by Sorooshian [1983] and Renard *et al.* [1982], among others.

## Sediment Transport and Yield

Sediment has been characterized as the greatest carrier of pollutant. Sediment yield generated by a storm or on a daily, monthly, or yearly basis has been modeled in a number of ways. By far the best known yield model for small watersheds is the *universal soil loss equation* (USLE) [Wischmeier and Smith, 1978], which can be written as

$$A = RKLSCP \quad (92.36)$$

where  $A$  is the soil loss per unit area, in the units of  $K$  and for the period of  $R$ ;  $R$  is the rainfall-runoff factor;  $K$  is the erodibility factor;  $L$  is the slope-length factor;  $S$  is the slope steepness factor;  $C$  is the cover and management factor; and  $P$  is the support practice factor. The values of these factors have been extensively tabulated for a wide range of soil-vegetation-land use conditions.

The conceptual models of sediment discharge are derived based on the unit hydrograph concept [Singh *et al.*, 1982; Williams, 1978; Sharma and Dickinson, 1980]. Consequently, a unit sediment graph for a watershed is defined as the graph of sediment discharge for a given duration that accumulates to 1 ton. The graph is generated by an effective sediment erosion intensity distributed

uniformly in time and in space of the watershed. Singh [1989] has presented a comprehensive discussion of conceptual sediment yield models.

There is a large volume of literature on sediment transport [Vanoni, 1975]. For comprehensive sediment transport models, the entire transport continuum, comprising detachment, deposition, degradation, suspension, and bed-load transport, has to be taken into account. Many of these processes can be modeled using the kinematic-wave theory.

## Solute Transport

The sources of water pollution can be distinguished as environmental, domestic, industrial, and agricultural. Depending upon the nature of the solute, transport process in a hydrologic environment may entail advection, diffusion, dispersion, adsorption, desorption, decay of contaminants, chemical reactions, solubilization, precipitation, volatilization, particulate transport, and miscibility. The mechanisms for transport are advection, dispersion, diffusion, solute-solid interaction, chemical reactions, and decay phenomena. The process by which solute is transported by the bulk motion of flowing water is called *advection*. When a solute moves, it tends to spread out from its advection path. This spreading, called *hydrodynamic dispersion*, comprises mechanical dispersion and molecular diffusion. Mechanical dispersion is caused entirely by mixing during the motion of the fluid. In molecular diffusion, molecular constituents move under the influence of their thermal kinetic energy in the direction of their concentration gradient. Many solutes react with soil through the process of adsorption. This reaction results in partitioning of the solute into the mobile solution phase and the immobile soil surface phase. The reverse of adsorption is the process of dislodgement of chemicals from soil, called *desorption*.

The fundamental law governing the solute transport is the law of conservation of solute mass, which for one-dimensional flow through unsaturated porous media is given by Eq. (92.19). Based on the consideration of sources and sinks, four groups of models can be identified: (1) models with no production and no decay terms ( $\gamma = \mu = 0$ ), (2) models with zero production only ( $\gamma \neq 0$ ,  $\mu = 0$ ), (3) models with first-order decay only ( $\mu \neq 0$ ,  $\gamma = 0$ ), and (4) models with zero-order production and first-order decay. If transport mechanisms constitute the basis of model classification, then they can be distinguished as kinematic-wave models and advection-dispersion models.

## Microbial Transport

Contaminated soils pose one of the most serious threats to surface and groundwater quality. In situ bioremediation is being touted as a viable technology to remedy this problem. Fundamental to development of this technology is the modeling of microbial transport. When microbes are injected into the subsurface environment to augment degradation, one of the problems faced is the limited capacity to transport and disperse bacteria by the soil through the zone of soil contamination [Jackson *et al.*, 1994]. The retention of bacteria by the soil matrix restricts transport of bacteria and is controlled by straining and adsorption. The models for bacterial transport are derived from Eqs. (92.20) through (92.22). Kommalapati *et al.* (1991–92) numerically solved the first two equations by assuming the absence of substrate for biological growth. In a similar manner,

simplifications of these equations lead to a variety of simple models.

## Models of Hydrologic Cycle

The models of hydrologic cycle are also the stream flow simulation models or the watershed hydrology models. These models are either event based or continuous-time. A large number of simulation models of both types are available. Singh [1989] has provided a short review of these models.

## 92.7 Statistical Hydrology

---

Statistical analyses in hydrology have been developed along three lines, as mentioned previously. The following sections present short discussions of each.

### Empirical Analyses

These analyses may involve determination of moments, regression and correlation, or entropy for given hydrologic data. If  $x_i, i = 1, 2, \dots, N$  represents a set of observations, then the mean  $\bar{x}$ , the variance  $S_x^2$ , and the skewness  $G$  are given, respectively, as

$$\begin{aligned}\bar{x} &= \frac{1}{N} \sum_{i=1}^N x_i; \\ S_1^2 &= \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2; \quad (92.37) \\ G &= \frac{1}{(n-1)(n-2)} \sum_{i=1}^N (x_i - \bar{x})^3\end{aligned}$$

The correlation coefficient  $r$  between two data sets of data  $X = \{x_1, x_2, \dots, x_N\}$  and  $Y = \{y_1, y_2, y_3, \dots, y_N\}$  is

$$r = \frac{\text{cov}(x, y)}{[\text{var}(x) \text{var}(y)]^{1/2}} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{N S_x S_y} = \frac{\sum_{i=1}^N x_i y_i - N \bar{x} \bar{y}}{N S_x S_y} \quad (92.38)$$

where  $\text{cov}(x, y)$  is the covariance of  $x$  and  $y$ ,  $\text{var}(\bullet)$  is the variance of  $(\bullet)$ ,  $S_x$  is the standard deviation of  $x$ , and  $S_y$  is the standard deviation of  $y$ . The coefficient of determination is the square of  $r$  and gives the amount of variability explained by the relationship between  $Y$  and  $X$ .

The regression analysis permits establishment of a relationship between the dependent variable  $Y$  and the set of independent variables  $X_i, i = 1, 2, \dots, M$  as

$$Y = f(X_i, i = 2, \dots, M; a_i, i = 0, 1, 2, \dots, M) \quad (92.39)$$



where  $a_i$  values are parameters. The parameters appearing in Eq. (92.39) are estimated using the least squares method.

Another empirical analysis frequently employed in hydrology involves finding a frequency distribution that best fits the given set of data. This can be done in three ways: (1) graphically, (2) testing the fit of an assumed distribution, and (3) applying the principle of maximum entropy subject to some information about the hydrologic system the data represent [Singh and Fiorentino, 1992]. A number of distributions have been used in hydrology and entropy has been used to derive a number of these distributions. Other examples of empirical analyses are pattern recognition-based analyses for in-filling of missing records, stream-flow forecasting, and so on [Unny, 1982; Panu, 1992].

## Phenomenological Analyses

These analyses make certain postulates about the behavior of a random variable and employ them to derive the probability distribution of the random variable. Examples of such analyses are the extreme-value analysis based on the random number of random variables [Todorovic, 1982], the point process theory of rainfall [Waymire *et al.*, 1984], the storage theory of reservoirs [Phatarford, 1976], and so on. As an example, probabilistic models of floods employ such assumptions as independence, stationarity, Markov property, and so on that concern the stochastic structure of the family of the random variables involved. A random number of random variable (RNRV) model considers flood exceedances above a threshold as a sequence of independent random variables and the number of such exceedances within the selected time interval, say a year, also as a random variable. The choice of probability distributions or stochastic processes for these random variables leads to a number of RNRV-based flood models [Todorovic, 1982].

## Stochastic Analyses

Stochastic analyses involve constructing the entire time series of a variable without explicitly knowing the probabilistic structure of the variable. Techniques based on time series analysis, spectral analysis, and so on belong to this class. As an example, consider a random variable  $X$  that is observed in a sequential manner as  $X_1, X_2, \dots$ , where the subscript may represent intervals of time, distance, and so forth. When the interval is time, this sequence is a time series and is stochastic. The set of random variables  $X_1, X_2, \dots$ , associated with its underlying probability distribution is a stochastic process. A time series model has a certain mathematical form and a set of parameters. A comprehensive discussion of time series models in hydrology is given by Salas *et al.* [1980].

## 92.8 Hydrologic Design

---

The type of project determines the need for a particular type of hydrologic information. Hydrologic design of many water resources projects is based on either a peak discharge or a complete discharge hydrograph. The selection of a design flood for water resources projects such as dams and spillways involves selecting safety criteria and estimating the flood that meets these criteria.



Depending upon the size of a water resources project, three types of design floods are recognized: (1) probable maximum flood (PMF), (2) standard project flood (SPF), and (3) frequency-based flood (FBF). Design of large dams is often based on the concepts of probable maximum precipitation (PMP) and probable maximum flood (PMF). These are either determined deterministically or statistically. Urban drainage is designed using design hydrographs and an acceptable level of risk. Reservoirs are designed using either the storage theory or flow duration curves. Methods of estimating a design flood can be distinguished, depending upon the data requirements, as rainfall-runoff methods, frequency-based methods, and risk-based methods.

## Defining Terms

**Hydrologic cycle:** Denotes the endless movement of water between atmospheric, earth, and oceanic systems through the processes of precipitation, evaporation, infiltration, and stream flow.

**Saturated flow:** Includes the flow of water occurring in the saturated geologic formations.

**Surface flow:** Includes flow of water occurring over the land.

**Unsaturated flow:** Includes the flow of water occurring in the vadose zone.

**Watershed:** Denotes the drainage area draining into an outlet; there is no flow across this area's boundaries. The magnitude of this area varies with the position of the outlet.

## References

- Ackermann, W. C., Colman, E. A., and Ogrosky, H. O. 1955. From ocean to sky to land to ocean. In *U.S. Department of Agriculture Yearbook 1955*, pp. 41–51. USDA, Washington, DC.
- Allen, R. G. 1985. A Penman formula for all seasons. *Journal of Irrigation and Drainage Engineering, ASCE*. 112(4):348–368.
- Beven, K. 1979. On the generalized kinematic routing method. *Water Resources Research*. 15(5): 1238–1242.
- Blaney, H. F. and Criddle, W. D. 1962. *Determining Consumptive Use and Irrigation Water Requirements*. Tech. Bull. 1275. USDA, Washington, DC.
- Bras, R. L. and Rodriguez-Iturbe, I. 1976. Evaluation of mean square error involved in approximating the areal average of a rainfall event by a discrete summation. *Water Resources Research*. 12(2):181–183.
- Buishand, T. A. 1982. Some methods for testing the homogeneity of rainfall records. *Journal of Hydrology*. 58:11–27.
- Buishand, T. A. 1984. Testing for detecting a shift in the mean of hydrological time series. *Journal of Hydrology*. 73:51–69.
- Dooge, J. C. I. 1959. A general theory of the unit hydrograph. *Journal of Geophysical Research*. 64(2):241–256.
- Dooge, J. C. I. 1973. *Linear Theory of Hydrologic Systems*. Tech. Bull. 1468. USDA, Agricultural Research Service, Washington, DC.
- Green, W. H. and Ampt, C. A. 1911. Studies on soil physics. 1. Flow of air and water through soils. *Journal of Agricultural Sciences*. 4:1–24.

- Holmes, R. M. and Robertson, G. W. 1959. A modulated soil moisture budget. *Monthly Weather Review*. 87(3):1–5.
- Horton, R. E. 1940. An approach toward a physical interpretation of infiltration capacity. *Soil Science Society of America Proceedings*. 5:399–417.
- Jackson, A., Roy, D., and Breitenbeck, G. 1994. Transport of a bacterial suspension through a soil matrix using water and an anionic surfactant. *Water Research*. 28(4):943–949.
- Jensen, M. E., Burman, R. D., and Allen, R. G. (Eds.) 1990. *Evaporation and Irrigation Water Requirements*. ASCE Manuals and Reports on Engineering Practice No. 70, 332 p. ASCE, New York.
- Jensen, M. E. and Haise, H. R. 1963. Estimating evapotranspiration from solar radiation. *Journal of Irrigation and Drainage Division, ASCE*. 89:15–41.
- Kommalapati, R. R., Wang, G. T., Roy, D., and Adrian, D. D. 1991–92. Transport and retention of microorganisms in porous media: Comparison of numerical techniques and parameter estimation. *Journal of Environmental Systems*. 21(2):121–142.
- Kruizinga, S. and Yperlaan, G. J. 1978. Spatial interpolation of daily totals of rainfall. *Journal of Hydrology*. 36:65–73.
- Liggett, J. A. and Woolhiser, D. A. 1967. Finite-difference solutions of the shallow water equations. *Journal of the Engineering Mechanics Division, ASCE*. 93(EMZ):39–71.
- Lighthill, M. J. and Whitham, G. B. 1955. On kinematic waves: 1. Flood movement in long rivers. *Proceedings of the Royal Society of London*. Series A. Vol. 229, pp. 281–316.
- Monteith, J. L. 1981. Evaporation and environment. *Symposium of Society for Experimental Biology*. 19:205–235.
- Nash, J. E. 1957. The form of the instantaneous unit hydrograph. *IAHS*. 45(3):114–121.
- Panu, U. S. 1992. Application of some entropic measures in hydrologic data infilling procedures. In *Entropy and Energy Dissipation in Water Resources*, ed. V. P. Singh and M. Fiorentino, pp. 175–192. Kluwer Academic, Dordrecht, The Netherlands.
- Penman, H. L. 1948. Natural evaporation from open water, bare soil and grass. *Proceedings of the Royal Society of London*. Series A. 193:120–145.
- Phatarford, R. M. 1976. Some aspects of stochastic reservoir theory. *Journal of Hydrology*. 30: 199–217.
- Philip, J. R. 1969. Theory of infiltration. In *Advances in Hydrosience, Vol. 5*, ed. V. T. Chow, pp. 215–296. Academic Press, New York.
- Ponce, V. M. 1986. Diffusion wave modeling of catchment dynamics. *Journal of Hydraulic Engineering*. 112(8):716–727.
- Renard, K. G., Rawls, W. J., and Fogel, M. M. 1982. Currently available models. In *Hydrologic Modeling of Agricultural Watershed*, ed. C. T. Haan, pp. 507–522. ASAE Monograph No. 5. American Society of Agricultural Engineers, St. Joseph, MI.
- Richards, L. A. 1931. Capillary conduction of liquids through porous mediums. *Physics*. 1:318–333.
- Salas, J. D., Delleur, J. W., Yevjevich, V., and Lane, W. L. 1980. *Applied Modeling of Hydrologic Time Series*. Water Resources Publications, Littleton, CO.
- Sharma, T. C. and Dickinson, W. T. 1980. System model of daily sediment yield. *Water Resources*

- Research*. 16(3):501–506.
- Singh, V. P. 1988. *Hydrologic Systems: Vol. 1. Rainfall-Runoff Modeling*. Prentice Hall, Englewood Cliffs, NJ.
- Singh, V. P. 1989. *Hydrologic Systems: Vol. 2. Watershed Modeling*. Prentice Hall, Englewood Cliffs, NJ.
- Singh, V. P. 1990. Hydraulic considerations for water resources modeling. *V. U. B. Hydrologie* 17, 280 pp. Vrije Universiteit Brussel, Brussels, Belgium.
- Singh, V. P. 1993. *Elementary Hydrology*. Prentice Hall, Englewood Cliffs, NJ.
- Singh, V. P., Baniukiewicz, A., and Chen, V. J. 1982. An instantaneous unit sediment graph study for small upland watersheds. In *Modeling Components of Hydrologic Cycle*, ed. V. P. Singh, pp. 534–554. Water Resources Publications, Littleton, CO.
- Singh, V. P. and Birsoy, Y. K. 1975. Comparison of the methods of estimating mean areal rainfall. *Nordic Hydrology*, 6(4):222–241.
- Singh, V. P. and Chowdhury, P. K. 1986. Comparing some methods of estimating mean areal rainfall. *Water Resources Bulletin*. 22(2):275–282.
- Singh, V. P. and Dickinson, W. T. 1975. An analytical method to determine daily soil moisture. *Proceedings of the Second World Congress on Water Resources*. IV:355–365. New Delhi, India.
- Singh, V. P. and Fiorentino, M. (Eds.) 1992. *Entropy and Energy Dissipation in Water Resources*. Kluwer Academic, Dordrecht, The Netherlands.
- Singh, V. P. and Yu, F. X. 1990. Derivation of infiltration equation using systems approach. *Journal of Irrigation and Drainage Engineering*. 116(6):837–858.
- Soil Conservation Service. 1971. Hydrology. In *SCS National Engineering Handbook* (section 4). USDA, Washington, DC.
- Sorooshian, S. 1983. Surface water hydrology: On line estimation. *Reviews of Geophysics and Space Physics*. 21(3):706–721.
- Todorovic, P. 1982. Stochastic modeling of floods. In *Rainfall-Runoff Relationship*, ed. V. P. Singh, pp. 597–650. Water Resources, Littleton, CO.
- Tung, Y. K. 1983. Point rainfall estimation for a mountainous region. *Journal of Hydraulic Engineering, ASCE*. 109(10):1386–1393.
- Unny, T. E. 1982. Pattern analysis for hydrologic modeling. In *Statistical Analysis of Rainfall and Runoff*, ed. V. P. Singh, pp. 349–387. Water Resources Publications, Littleton, CO.
- Vanoni, V. (Ed.) 1975. *Sedimentation Engineering*. ASCE Manual and Reports on Engineering Practice No. 54. ASCE, New York.
- Waymire, E., Gupta, V. K., and Rodriguez-Iturbe, I. 1984. A spectral theory of rainfall intensity at the meso- $\beta$  scale. *Water Resources Research*. 20:1454–1465.
- Williams, J. R. 1978. A sediment graph model based on an instantaneous unit sediment graph. *Water Resources Research*. 14(4):659–664.
- Wischmeier, W. H. and Smith, D. D. 1978. *Predicting Rainfall Erosion Losses: A Guide to Conservation Planning*. Agricultural Handbook No. 537. Science and Education Administration, USDA, Washington, DC.
- Zawadzki, I. I. 1973. Errors and fluctuations of raingage estimates of areal rainfall. *Journal of*

*Hydrology*. 18:243–255.

### **Further Information**

*Advances in Water Resources*. Published quarterly by Elsevier Science Publishers.

*Hydrological Sciences Journal*. Published quarterly by International Association of Hydrological Sciences.

*Journal of Hydrology*. Published monthly by Elsevier Science Publishers.

Maidment, D. R. (Ed.) 1992. *Handbook of Hydrology*. McGraw-Hill, New York.

*Water Resources Research*. Published monthly by American Geophysical Union.

Richardson, E. V. "Sedimentation"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

- 93.1 Fluvial Geomorphology
- 93.2 Sediment Properties
- 93.3 Beginning of Motion
- 93.4 Sediment Yield
- 93.5 Bed Forms
- 93.6 Sediment Transport
- 93.7 Reservoir Sedimentation

## E. V. Richardson

*Ayres Associates Fort Collins, Colorado*

Sedimentation in water resources engineering requires knowledge of and the ability to determine the yield of sediment from the land surface, the transport of sediment by streams, erosion and erosion control, and reservoir sedimentation. This project requires an understanding of **fluvial geomorphology**, sediment properties, hydraulics of open channel flow, river mechanics, and **bed forms** in alluvial channels.

## 93.1 Fluvial Geomorphology

---

Fluvial geomorphology is the science dealing with the profiles and planforms of streams and rivers. Rivers are dynamic—always changing their shape, position, and other morphological characteristics—with long- and short-term changes in water or sediment discharge, climate, tectonic and humanity's activities. A local change in a river or stream, such as construction of a dam or bridge, causes modifications of a river both up- and downstream. In a dam, changes in water and sediment discharge can cause degradation downstream, and backwater can cause aggradation upstream. The fluvial geomorphology of a stream determines its mode of sediment transport and its response to changes in climate, tectonic activity, man's activity, and changes in sediment and water discharge.

Geomorphic factors affecting stream morphology include stream size, flow habit (ephemeral, perennial but flashy, or perennial), bed and bank material, valley setting, flood plains, natural levies, apparent incision, channel boundaries, vegetation, stream planform, variability of width, and development of bars [Vanoni, 1975; Schumm, 1977; Richardson *et al.*, 1990; and Simons and Senturk, 1992]. Stream planform is broadly classified as **meandering, straight, braided, or anabranching**. These broad classifications have been subclassified by geomorphologists and engineers, but for sedimentation studies these broad classifications normally will be

sufficient.

Lane, studying the relationship between **slope**,  $S$ , **discharge**,  $Q$ , and channel planform, observed that, when  $SQ^{1/4}$  [Richardson *et al.*, 1990]. Also, when  $SQ^{1/4} > 0.01$ , a stream had a braided planform. A stream with a value between the two could have either a meandering or braided planform. A meandering stream with slope increased such that the value of  $SQ^{1/4}$  is larger than 0.0017 might become a braided stream. Conversely, a braided stream that has its discharge or slope decreased such that the value of  $SQ^{1/4}$  is smaller than 0.01 might change to a meandering planform.

Lane also proposed the following qualitative relationship to predict the response of a stream or river to changes in water discharge,  $Q$ , slope,  $S$ , sediment discharge,  $Q_s$ , and **median diameter** of the bed material,  $D_{50}$  [Schumm, 1972]:

$$QS \sim Q_s D_{50} \quad (93.1)$$

This relationship shows that, with a decrease in  $Q$  and no changes in the qualities on the right, there is an increase in the slope. The increase in slope results from deposition of  $Q_s$ . Changes in other qualities would induce a response to keep a balance between the left and right sides of the relationship. Richardson *et al.* [1990], Simons and Senturk [1992], and Lane [Schumm, 1972] give examples of the uses of this relationship and also expand on the effect of additional variables on the response of streams to change. Leopold and Maddock, Mackin, and others [Schumm, 1972, 1977] give additional insight into the response of streams to change and the importance of fluvial geomorphology in sedimentation.

## 93.2 Sediment Properties

---

Sediment properties of value in sedimentation are the physical size, fall velocity and density of the particles, and bulk properties of the bed and bank material and sediment deposits. Sediments are composed of clay (0.000 2 to 0.004 mm), silt (0.004 to 0.062 mm), sand (0.062 to 0.2 mm), gravel (2.0 to 64 mm), cobble (64 to 250 mm), or boulder (250 to 4000 mm) material. Bulk properties are described by the size frequency distributions, specific weight, and porosity. Size distributions are determined by sieving, visual accumulation tube analysis, pebble count, and pipette methods. Size distributions are usually expressed as a percentage finer than a given size in the distribution. Also of major importance is the viscosity of suspensions with large concentrations of silts and clays. Sediment properties and methods to determine them are discussed in detail by Brown [1950], Vanoni [1975], Richardson *et al.* [1990], and Simons and Senturk [1992].

## 93.3 Beginning of Motion

---

Knowledge of the point at which fluid forces are large enough to move sediment particles is important in sediment transport, erosion, and design of riprap. Shields [Brown, 1950; Vanoni, 1975; and Simons and Senturk, 1992] experimentally determined a relationship at beginning of motion between the ratio of the **critical shear stress**  $\tau_c$  to move a particle and its submerged

weight expressed as  $(S_s - 1)\gamma D$  and the **shear velocity**, particle size Reynolds number  $[(gRS)^{0.5} D/\nu]$ , where  $S_s$  is the specific gravity of the particle,  $\gamma$  is the unit weight of water,  $g$  is the acceleration of gravity,  $R$  is the hydraulic radius,  $D$  is the particle size, and  $\nu$  is the kinematic viscosity. Lane, Fortier and Scobey, Keown, and others [Brown, 1950; Vanoni, 1975; Richardson *et al.*, 1990; Simons and Senturk, 1992] give values for the critical shear or critical velocity for the beginning of motion of silts, clays, sand, and coarser particles. Equations for determining **shear stress** on a boundary are given in the above references. [Neil Richardson *et al.*, 1993] gives the following equation for the critical velocity at the beginning of motion of a particle:

$$V_c = 1.58[(S_s - 1)gD]^{1/2} (y/D)^{1/6} \quad (93.2)$$

where  $D$  = particle size (ft or m),  $g$  = acceleration of gravity (ft/s<sup>2</sup> or m/s<sup>2</sup>),  $V_c$  = critical velocity above which bed material of size  $D$  and smaller will be transported (ft/s or m/s),  $S_s$  = specific gravity of bed material particles, and  $y$  = depth of flow (ft or m). When  $S_s$  equals 2.65—a typical value for sand—Eq. (93.2) in metric units reduces to

$$V_c = 6.36y^{1/6} D^{1/3} \quad (93.3)$$

## 93.4 Sediment Yield

---

Determining the amount of erosion (*sediment yield*) from the land has great significance in water resources engineering. Erosion is classified into overland processes and stream processes. Overland processes include sheet wash, rilling, and gullying. Stream processes include bed and bank erosion and will be described later.

The *universal soil loss equation* (USLE), developed by the U.S. Department of Agriculture, is the most widely used regression equation for predicting sediment yield from overland flow [Vanoni, 1975; Simons and Senturk, 1992]. The five major factors used to determine the average annual soil loss  $A$ , in U.S. tons per acre, are the rainfall factor  $R$ , in inches per hour; soil erodibility factor  $k$ , in tons per acre per unit of rainfall factor  $R$ ; topography factors  $S$  for slope (land gradient) and  $L$  for length of slope; cropping and management factor  $C$ ; and erosion control practices factor  $P$ .

## 93.5 Bed Forms

---

Flow in sand and medium gravel alluvial channels ( $D_{50}$  from 0.062 to 16 mm) is divided into lower and upper flow regime, separated by a transition zone on the basis of the bed form, resistance to flow, and **bed material discharge**. In the lower flow regime the bed forms in natural channels are dunes, with large resistance to flow (Manning  $n$  ranges from 0.02 to 0.04) and low bed material discharge (concentrations ranging from 200 to 2000 parts per million by weight). In the upper flow regime the bed forms are plane beds or antidunes, with low resistance to flow



(Manning  $n$  ranging from 0.012 to 0.020) and large bed material discharge (concentrations of 1000 ppm and larger). In the transition between the two regimes the bed form, resistance to flow, and bed material discharge ranges between characteristics of the two. The bed form in an alluvial stream depends on the discharge, slope, depth of flow, size of bed material, and the viscosity of the fluid [Vanoni, 1975; Richardson *et al.*, 1990; and Simons and Senturk, 1992].

Alluvial streams with steep slopes may flow in the upper flow regime all the time, whereas streams with lower slopes change from lower to upper flow regimes depending upon the discharge, bed material size, and fluid viscosity. This tendency results in streams that at low flow are in the lower flow regime and at high flow are in the upper flow regimes. These streams can have a dune bed form in the summer and washed-out dunes, plane beds, or antidunes in the fall and winter. An example is the Missouri River along the border between Nebraska and Iowa. In the summer, when the water temperature is high (70 to 80°F), the bed form is in the lower flow regime and the Manning  $n$  is 0.020; in the fall, water temperatures are much lower (35 to 65°F) and the bed form is in the transition and Manning  $n$  is 0.015 with, consequently, a lower depth and higher velocity for the same discharge (U.S. Corps of Engineers, 1969). These changes in flow regime with discharge and temperature produces discontinuous or shifting rating curves for many sand channel streams and affects the determination of discharge, depths of flow, velocity, and bed material discharge in these streams.

## 93.6 Sediment Transport

---

The quantity of sediment transported by a stream consists of the **bed material discharge** and the **fine material discharge (washload discharge)** from the watershed and banks. The bed material discharge is determined by the flow variables and fluid and sediment properties and is, with some degree of accuracy, subject to calculation. The fine sediment discharge depends on availability of fine sediment, is not transported at the capacity of the stream to transport it, is not functionally related to measurable hydraulic variables, and must be measured.

Bed material and fine material are transported in suspension by the turbulence of the stream (**suspended sediment discharge**) or by rolling along in contact with the bed (**contact sediment discharge**). Laursen [Richardson *et al.*, 1990, 1993] determined that, when the ratio of the shear velocity  $[(gRS)^{0.5}]$  to the fall velocity ( $\omega$ ) of the bed material is less than 0.5, the bed material discharge is mostly contact bed material discharge; ratios between 0.5 and 2.0 suggest some suspended bed material discharge, and those larger than 2.0 suggest mostly suspended bed material discharge. When the bed material discharge is mostly in contact with the bed, the Meyer-Peter-Muller equation is normally used to determine bed material discharge. Suspended bed material discharge is measured or calculated by integrating the velocity and sediment concentration at a point through the vertical and across the stream. Suspended fine sediment discharge is measured by integrating the velocity and sediment concentration at a point through the vertical and across the stream using depth integrating samplers. To calculate suspended bed material discharge, an equation for the distribution of sediment particles in the vertical (developed by Rouse) is combined with a velocity distribution equation and integrated through the depth. Various investigators then combined the suspended bed material discharge with the contact bed material discharge to develop an equation to determine the total bed material discharge. To determine the

total sediment discharge of a stream or river, the fine sediment discharge, which has to be measured, is added to the measured or calculated bed material discharge.

The Meyer-Peter-Muller, Einstein, modified Einstein (developed by Colby and Hembree), and other equations for calculating total bed material discharge are given by Brown [1950], Vanoni [1975], Richardson *et al.* [1990], and Simons and Senturk [1992].

Colby [1964] developed a very useful graphical method of determining total bed material discharge. The method is given in Figs. 93.1 and 93.2 and Eqs. (93.4) and (93.5):

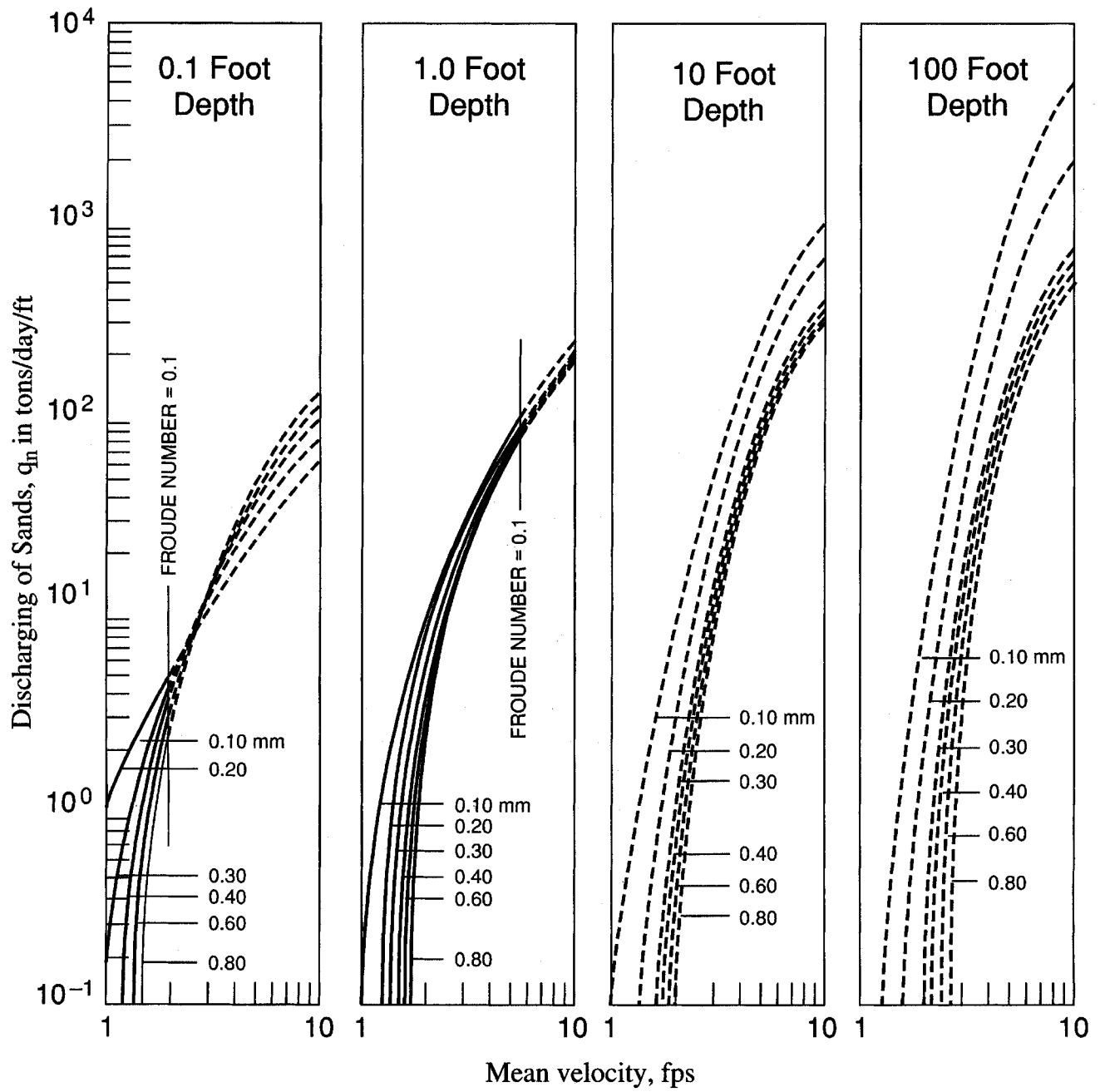
$$q_T = [1 + (K_1 K_2 - 1) K_3] q_n \quad (93.4)$$

$$Q_s = W q_T \quad (93.5)$$

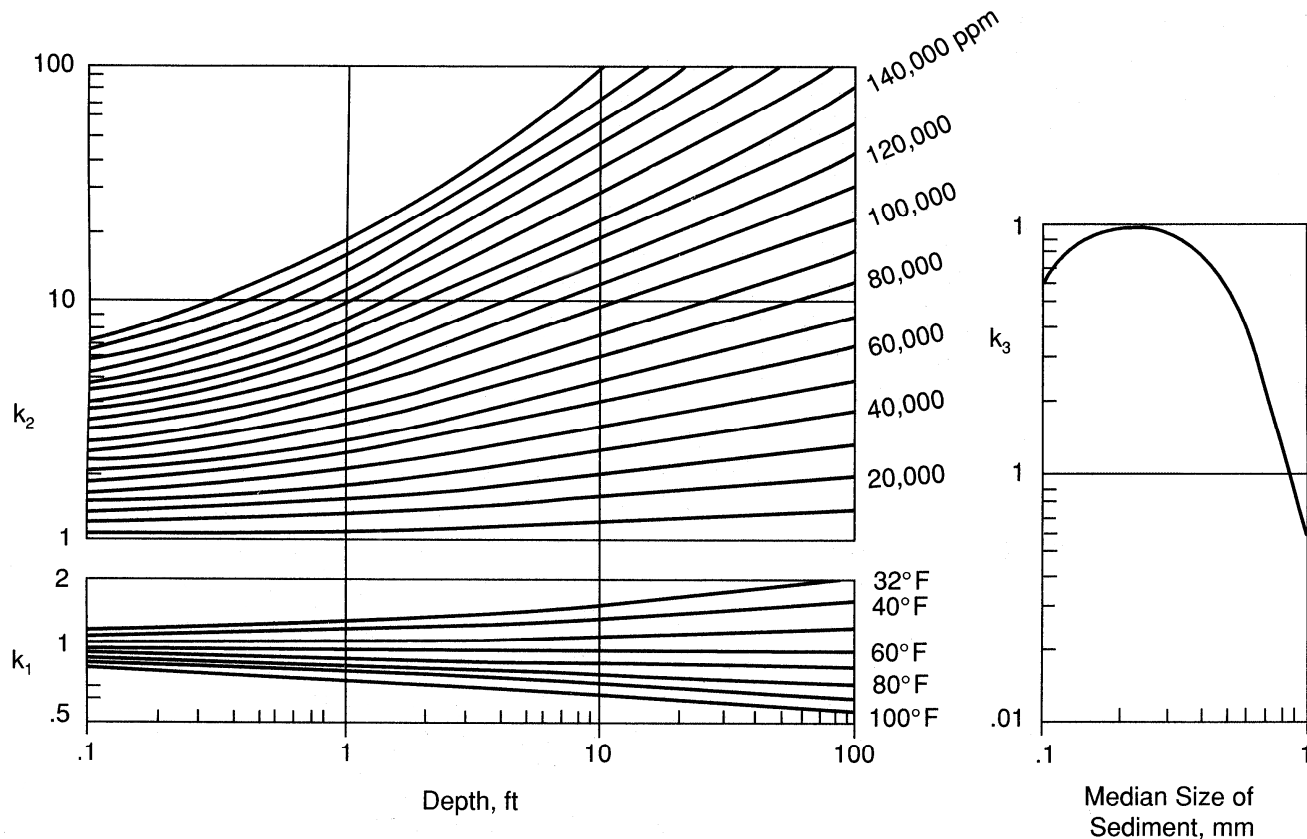
where  $q_T$  is total bed material discharge per unit width (U.S. tons/day/ft), the  $K_s$  are correction coefficients determined from Fig. 93.2,  $q_n$  is the discharge of bed material determined from Fig. 93.1, and  $Q_s$  (U.S. tons/day) is the total bed material discharge for a channel of width  $W$  (ft).

**Figure 93.1** Relation of discharge of sands to mean velocity for six median sizes of bed sands, four depths of flow, and a water temperature of 60°F. (Source: Colby, B. R. 1964. *Discharge of Sands and Mean-Velocity in Sand-Bed Streams*. U.S. Geological Survey Professional Paper 462-A. U.S. Geological Survey, Washington DC.)

**Figure 93.1**



**Figure 93.2** Colby's correction curves for temperature and fine sediment (Source: Colby, B. R. 1964. *Discharge of Sands and Mean-Velocity in Sand-Bed Streams*. U.S. Geological Survey Professional Paper 462-A. U.S. Geological Survey, Washington, DC.)



The uncorrected total bed material discharge ( $q_n$ ) is determined from Fig. 93.1 for a given velocity, median diameter of the bed material, and two depths that bracket the desired depth by interpolating on a logarithmic graph of depth versus  $q_n$  to obtain the bed material discharge per unit width. The corrected bed material discharge per unit width ( $q_T$ ) is determined using Eq. (93.4) and Fig. 93.2. The total bed material discharge is then obtained using Eq. (93.5).

The modified Einstein method determines the **total bed material discharge** and **total sediment discharge** from measurements of the suspended sediment discharge to determine the **unmeasured sediment discharge** and, thus, is probably the most accurate of the methods used to calculate the total bed material discharge and the total sediment discharge of a stream. The unmeasured sediment discharge is composed of the contact bed material discharge and the suspended sediment discharge in the unmeasured zone. Suspended sediment samplers, except in special cases, only measure within a set distance to the bed (0.2 to 0.4 ft, depending on the sampler). Suspended sediment samplers that are used in very turbulent streams or in specially constructed turbulence flumes (all the bed material is in suspension, no contact discharge) or that traverse the total depth as at the nape of a weir measure the total sediment discharge of a stream. In fact, this method was used to obtain data for the development of the modified Einstein method and other equations. Suspended sediment samplers and measurement methods are described by Vanoni [1975] and Simons and Senturk [1992]. Mathematical computer models have been developed to determine total bed material discharge using currently available equations. Models are described by

Richardson, et al. [1990] and Simons and Senturk [1992].

## 93.7 Reservoir Sedimentation

---

The rate of depletion of reservoir storage from storage of sediment depends on (1) the volume of the reservoir in relation to inflow, (2) sediment inflow, (3) reservoir trap efficiencies, and (4) the specific weight (density) of the sediment deposits. The distribution of the sediment in the reservoir can sometimes be important. Sedimentation damages a reservoir over time if it decreases storage volume to such an extent that it no longer can serve its design function.

Reservoirs with large storage volume in relation to average annual inflow of water have a lower rate of storage loss due to sedimentation than reservoirs that do not—even with the same trap efficiency. Lake Mead on the Colorado River and Lake Nasser on the Nile River have storage volumes that are about double their average annual inflow. Their trap efficiencies are almost 100%. Their rate of sediment depletion of storage is so low that their useful life expectancy is measured in thousands of years. The reservoir behind Tarbella Dam on the Indus River has a storage volume that is less than 20% of the average annual flow. Its trap efficiency is also about 100%. Its rate of storage depletion is so large that its useful life as a storage reservoir is less than 100 years. All three reservoirs have approximately the same annual sediment inflow. Of interest is the fact that the reservoir behind Old Aswan Dam on the Nile (downstream of Lake Nasser), with a storage-to-inflow ratio of less than 3%, has an infinite useful life. The reservoir stored water for over 50 years before Lake Nasser was created by the High Aswan Dam. Its trap efficiency is close to zero because (1) it has under-sluices across the total dam width, and (2) most of the annual flood passes through the reservoir, with only the tail of the flood stored.

The sediment inflow into a reservoir can be estimated by (1) use of the recorded annual measured total sediment discharge of the stream, (2) computation by flow duration–sediment rating curve method, and (3) estimating the sediment yield from the watershed.

The use of recorded annual measured total sediment discharge of a stream is limited to those sites that have a historical record based on frequent sampling to establish a reliable estimate of the sediment inflow into the reservoir. Normally, the record of sediment inflow is less than the stream flow record and is only suspended sediment discharge. However, with studies, the suspended sediment record can be adjusted to obtain the total sediment discharge using methods described in the previous section.

The flow duration–sediment rating curve method, which extends the available total sediment discharge record to the historical stream flow record, is the most desirable. In this method a sediment rating curve is made by relating the daily sediment discharge (normally expressed in tons per day) to the daily discharge. There will be a large scatter in the data, some of which may be seasonal, but with study a reliable set of curves can be developed. A flow duration curve is also prepared for the entire stream flow record. From these two sets of relationships the average annual sediment discharge is determined [Vanoni, 1975; Simons and Senturk, 1992]. The sediment rating curve may be developed using measured suspended sediment discharge, measured total sediment discharge (suspended sediment plus unmeasured sediment discharge), or a measured suspended sediment discharge corrected for the unmeasured sediment discharge.

The sediment yield from a watershed is estimated using the universal soil loss equation given earlier. This value is then used with the appropriate trap efficiency to determine sedimentation rates in a reservoir. This method is often used for small reservoirs.

Trap efficiency is the measure of the percentage of the sediment inflow retained (trapped) in the reservoir. It depends on the velocity of flow through the reservoir as well as sediment size. The velocity of flow depends on the size of the reservoir, type of dam, and operating procedures. Reservoirs that are formed from large embankments (soil, concrete, or rock), with large storage volumes and over-year storage, and outlets that normally discharge less flow than the incoming floods have large trap efficiencies (close to 100%). These reservoirs may have some sediment removed by density currents or large velocities when the reservoir is low, but these amounts are ignored in consideration of the other approximations and uncertainties in the determination of the sediment inflow. Small reservoirs with large velocities or reservoirs with under-sluices that can pass large inflows will have low trap efficiencies. The Old Aswan Dam described earlier is an example of the latter. Operation of the dam to have a low pool with high velocities part of each year will decrease trap efficiency.

To determine trap efficiency, Churchill—from a study of TVA reservoirs—developed a relationship between the percentage of incoming silt passing through the reservoir and the ratio of period of retention divided by mean velocity; Brune developed a relationship between percentage of sediment trapped and a ratio of reservoir capacity divided by the mean annual inflow [Vanoni, 1975; Simons and Senturk, 1992]. In using these relations, engineering judgment must be used to determine trap efficiency for a particular reservoir.

The specific weight of the sediment deposited in a reservoir is needed to convert to volume the estimate of sediment deposited in the reservoir, which is normally given in terms of weight. The specific weight of the sediment deposits in the reservoir increases with time as they consolidate. Coarse sediments (sand and gravels) will consolidate faster than the finer silts and clays and will reach their ultimate weight faster. Also, sediments consolidate faster if they are not always submerged. Vanoni [1975] and Simons and Senturk [1992] present methods for estimating the specific weight developed by Lane and Koelzer, which takes into account the type of sediment and degree of submergence.

The distribution of sediment in reservoirs depends on the size composition of the sediment flowing into a reservoir and the management of the outflow. Coarse sediments are deposited in the upper reaches of the reservoir and the finer sediments farther down. If the water storage in the reservoir is managed so as to have a low water level part of the year, then the coarser materials are moved farther down into the reservoir. Coarse sediment deposits at the upper end of the reservoir can increase the backwater effects upstream. Vanoni [1975] and Simons and Senturk [1992] discuss the location of sediment deposits and present methods to evaluate the location.

Sediment deposits in reservoirs are measured using sonic sounders and surveying techniques to monitor the loss of storage with time. Vanoni [1975] describes methods for conducting these surveys.

## Defining Terms

**Anabranch:** A stream whose flow is divided at normal or lower discharges by large, relatively permanent islands. The channels are more permanent and more widely and distinctly separated than those of a braided stream.

**Bed form:** A relief feature on the bed of a stream, such as dunes, plane bed, or antidunes. Also called *bed configuration*.

**Bed material:** Material found on the bed of a stream. May be transported in contact with the bed or in suspension.

**Bed material discharge:** The part of the total sediment discharge of a stream that is composed of grain sizes found in the bed and is equal to the transport capability of the flow.

**Braided stream:** A stream whose flow is divided at normal and low flow into several channels by bars, sandbars, or islands. The bars and islands change with time, sometimes with each runoff event. A braided stream has the aspect of a single large channel, with several subordinate channels at normal and low flows.

**Contact sediment discharge:** Sediment that is transported in a stream by rolling, sliding, or skipping along in contact with the bed. Also called *bed load* or *contact load*.

**Critical shear stress:** The minimum amount of shear (force) exerted by the flow on a particle or group of particles that is required to initiate particle motion.

**Discharge:** Time rate of the movement of a quantity of water or sediment passing a given cross section of a stream or river.

**Fine sediment discharge (washload discharge):** That part of the total sediment discharge of a stream that is not found in appreciable quantities in the stream bed. Normally, the fine sediment discharge in a sand bed stream is composed of particles finer than sand (0.062 mm). In coarse gravel, cobble, or boulder bed stream silts, clays and sand could be fine sediment or washload discharge.

**Flow duration curve:** A graph indicating the percentage of time a given discharge is exceeded.

**Fluvial:** Related to stream or rivers.

**Geomorphology:** The branch of physiography and geology that deals with the general configuration (form) of the earth's surface and the changes that take place as the result of the forces of nature.

**Hydraulic radius:** The cross-sectional area of a stream, divided by its wetted perimeter. Equals the depth of flow when the width is larger than ten times depth.

**Meandering stream:** A stream with sinuous S-shaped flow pattern.

**Median diameter:** The particle diameter at which 50% of a sample's particles are coarser and 50% are finer (D<sub>50</sub>).

**Sediment yield:** The total sediment outflow from a unit of land (field, watershed, or drainage area) at a point of reference and per unit of time.

**Shear stress, tractive force:** The force or drag on the channel boundaries that is caused by the flowing water. For uniform flow, shear stress is equal to the unit weight of water times the hydraulic radius times the slope. Usually expressed as force per unit area.

**Shear velocity:** The square root of the shear stress divided by the mass density of water, in units of

velocity.

**Slope:** Fall per unit length of the channel bottom, water surface, or energy grade line.

**Suspended sediment:** Sediment particles that are suspended in the flow by the turbulence of the stream.

**Suspended sediment discharge:** The quantity of suspended sediment passing through a stream cross section per unit of time.

**Total bed material discharge:** The sum of the suspended sediment bed material discharge and the contact sediment (bed-load) discharge.

**Total sediment discharge:** The sum of the suspended sediment discharge and the contact sediment (bed-load) discharge, the sum of the bed material discharge and the wash-load discharge, or the sum of the measured sediment discharge and the unmeasured sediment discharge.

**Unmeasured sediment discharge:** The sediment discharge that is not measured by suspended sediment samplers. It consists of the suspended sediment discharge in the unsampled zone and the contact sediment discharge. Suspended sediment samplers normally do not measure to the bed of a stream.

## References

- Brown, C. B. 1950. Sediment transportation. In *Engineering Hydraulics*, ed. H. Rouse, p. 769–858. John Wiley & Sons, New York.
- Colby, B. R. 1964. *Discharge of Sands and Mean-Velocity in Sand-Bed Streams*. U.S. Geological Survey Professional Paper 462-A. U.S. Geological Survey, Washington, DC.
- Richardson, E. V., Harrison, L. J., Richardson, J. R., and Davis, S. R. 1993. *Evaluating Scour at Bridges*. Pub. No. FHWA-IP-90-017. FHWA, McLean, VA.
- Richardson, E. V., Simons, D. B., and Julien, P. Y. 1990. *Highways in the River Environment*. Pub. No. FHWA-HI-90-016. FHWA, Washington, DC.
- Schumm, S. A. (Ed.). 1972. *River Morphology*. Benchmark Papers in Geology. Dowden, Hutchinson & Ross, Stroudsburg, PA.
- Schumm, S. A. 1977. *The Fluvial System*. John Wiley & Sons, New York.
- Simons, D. B. and Senturk, F. 1992. *Sediment Transport Technology*. Water Resources Publications, Littleton, CO.
- U.S. Corp. of Engineers. 1969. *Missouri River Channel Regime Studies*. MRD Sed. Series No. 13.B. U.S. Corps of Engineers, Omaha, NE.
- Vanoni, V. A. (Ed.). 1975. *Sedimentation Engineering*. ASCE Manual No. 54. ASCE, New York.

## Further Information

The *ASCE Journal of Hydraulic Engineering, Transactions* and the annual publication titled *Hydraulic Engineering* report advances in sedimentation. For subscription information contact: ASCE, 345 E. 47th St. New York, NY 10164-0749. Phone: (800)548-2723.

The U.S. Geological Survey publishes data on the sediment discharge of streams in the U.S., techniques for measuring sediment discharge, and recent advances in the science of sedimentation.



Catalogs of their publications are available from USGS Map Distribution, Box 25286, MS 306, Federal Center, Denver, CO 80225.

The U.S. Army Corps of Engineers publishes reports on their research in sedimentation. These reports focus on studies of a specific river problem as well as the general subject of sedimentation. The library at the Waterways Experiment Station, Vicksburg, MS 39180, maintains copies of most Corps publications.

American Geophysical Union (AGU) is an excellent source for additional information on sedimentation. They can be contacted at AGU, 2000 Florida Ave. NW, Washington, DC. 20009.

The Water Resources Bulletin of AWRA contains papers on sedimentation. Their address is AWRA, 5410 Grosvenor Lane, Suite 220, Bethesda, MD 20814-2192.

Agencies of the federal government such as ARS and SCS (U.S. Department of Agriculture) and FEMA and USBR (U.S. Department of the Interior) have extensive information on sedimentation of both specific and general nature. Access to their publications can be made through their local offices or offices in Washington, DC.

Ziener, R. E. "Linear Systems and Models"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



This video camcorder provides a good example of the interaction of a number of subsystems which, taken together as a system, are capable of performing the necessary tasks of recording and playing back video imagery. One important component is the autofocus subsystem, which automatically adjusts the focus in the center of the scene via a feedback system. Digital signal processing is used in another subsystem to improve picture clarity and to provide special effects such as fading. The tape drive uses a combination of electrical and mechanical subsystems to maintain a constant speed of the tape past the record and playback heads. These are examples of only a few of the subsystems which enable this amazing technological marvel to provide high-quality video pictures in a small, easy-to-use, lightweight package. (Photo courtesy of RCA.)

# XVI

## Linear Systems and Models

---

**Rodger E. Ziemer**

*University of Colorado, Colorado Springs*

**94    Transfer Functions and Laplace Transforms**    *C. N. Dorny*

Transfer Functions • The Laplace Transformation • Transform Properties • Transformation and Solution of a System Equation

**95    Block Diagrams**    *A. P. Sage*

Elements of the Block Diagram • Block Diagram Reduction • Summary

**96    Signal Flow Analysis**    *A. D. Kraus*

The Signal Flow Graph • Transmission Gain • Signal Flow Graph Algebra • The Mason Gain Rule

**97    Linear State-Space Models**    *B. D. Schimel and W. J. Grantham*

State-Space Models • Linearization • Linear System Representations • Transforming System Representations

**98    Frequency Response**    *P. Neudorfer and P. Gehlen*

Frequency Response Plotting • A Comparison of Methods

**99    Convolution Integral**    *R. E. Ziemer*

Fundamentals • Properties of the Convolution Operation • Applications of the Convolution Integral • Two-Dimensional Convolution • Time-Varying System Analysis

**100   Stability Analysis**    *R. T. Stefani*

Response Components • Internal (Asymptotic) and External (BIBO) Stability • Unstable and Marginally Stable Responses • Structural Integrity

**101    $z$  Transform and Digital Systems**    *R. Johansson*

The  $z$  Transform • Digital Systems and Discretized Data • The Transfer Function • Digital Systems Described by Difference Equations (ARMAX Models) • Prediction and Reconstruction • The Kalman Filter

A SYSTEM CAN BE DEFINED as a combination and interconnection of several components or subsystems to perform a desired task. Systems can be mechanical, electrical, thermal, acoustic, or a combination of these as well as other possibilities. For example, the automatic focusing feature of a video camera incorporates an optical sensor subsystem coupled electronically to an electromechanical control subsystem for changing the focus of the lens subsystem.

A linear system is one for which the principle of superposition holds, that is, the response to a linear combination of two separate inputs yields the same result as the linear combination of the outputs due to each input applied separately. It is rather fortunate that many systems are well modeled as linear systems, because this simplifies their analysis considerably. For the most part, this section deals with linear systems and their models. It consists of eight chapters that deal with various aspects of system analysis and synthesis.

**Chapter 94**, on transfer functions and the Laplace transform, provides the cornerstone for

system analysis. The Laplace transform provides a one-to-one correspondence between a signal, or time function, and a function of a complex frequency variable,  $s$ , called the Laplace transform of the signal, which in turn converts the differential equation describing a system to an algebraic expression. The ratio of the Laplace transform of the output to the Laplace transform of the input of a fixed linear system (i.e., one whose properties do not vary with time) is the transfer function of the system. Many properties of the system can be deduced from the transfer function, and it in fact allows the design or synthesis of systems to perform desired tasks. **Chapter 95**, on block diagrams, provides a pictorial representation of a fixed linear system in terms of its subsystems. Manipulation of subsystems via the block diagram of the system allows much flexibility and power in analysis and synthesis of systems. **Chapter 96**, on signal flow graphics, provides an alternative to the block diagram of a system; the signal flow graph is more compact but still incorporates the convenience of a block diagram. In **Chapter 97**, on linear state-space models, a representation for a system is provided that is equally applicable to nonlinear and linear systems, is especially suited for time domain analysis, and allows multiple-input, multiple-output systems to be represented as conveniently as single-input, single-output systems. Furthermore, the state equations for a linear time-varying system are no more complex than those of a fixed linear system, and their numerical solution via digital computer can proceed with equal facility. **Chapter 98**, on frequency response, provides a description of a fixed linear system based on its response to a steady state unit-amplitude sinusoid in terms of the output sinusoid's amplitude (amplitude response) and phase (phase response). These two functions of frequency provide much insight into the response of a system to other inputs and its stability (i.e., its response to a bounded input). **Chapter 99** describes the convolution integral, its properties, and its applications. Among other applications, the convolution integral provides a means for determining the output of a linear system in response to a wide range of input signals. In **Chapter 100**, a treatment of stability analysis for linear systems is given. Various types of stability are defined and relationships between them are given. The final chapter in this section provides an introduction to discrete-time systems. The introduction to the  $z$  transform given in this chapter provides the basis for transform analysis of such systems much as the Laplace transform provides an analysis tool for continuous-time systems.

This section provides a thorough overview of linear systems, their mathematical modeling, and appropriate tools for analyzing them as summarized by eight experts in the field. >

Dorny, C. N. "Transfer Functions and Laplace Transforms"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Transfer Functions and Laplace Transforms

---

- 94.1 Transfer Functions
- 94.2 The Laplace Transformation
- 94.3 Transform Properties
- 94.4 Transformation and Solution of a System Equation

**C. Nelson Dorny**

*University of Pennsylvania*

We perceive a system primarily through its behavior. Therefore, our mental image of a system usually includes representative response **signals**. The *step response*, the behavior when we suddenly turn on the system, is such a system-characterizing signal. We should view the step response as a description of the system. The *impulse response* is another description of the system. For a system represented by linear differential equations, the unit-step response is the integral of the unit-impulse response.

Let us represent time differentiation ( $d/dt$ ) by the *time-derivative operator*,  $p$ . Then we can denote the time derivative of a signal  $y$  by  $py$ , its second derivative by  $p^2y$ , its integral with respect to time by  $(1/p)y$ , and so on. This *operator notation* simplifies the expressions for differential equations. We shall use the expression *system equations* to mean a set of differential equations that determines fully the behaviors of the dependent variables that appear in those equations. We can reduce a *linear* set of system equations to a single **input-output system equation** by eliminating all but one dependent variable from the set. The *transfer function* associated with that dependent variable is a mathematical expression that contains all the essential information embodied in the system differential equation.

The Laplace transformation converts signals (functions of time) to functions of a *complex-frequency variable*,  $s = \sigma + j\omega$ . There is a one-to-one correspondence between a signal and its Laplace transform. We can retrieve the time function by inverse transformation. Laplace transformation produces images that have some properties that are more convenient than those of the original signals. In particular, time differentiating a signal corresponds to multiplying its Laplace transform by the complex-frequency variable  $s$ . Hence, the transformation converts linear constant-coefficient differential equations to linear algebraic equations. Such simplifications of time-domain operations make Laplace transformation useful.

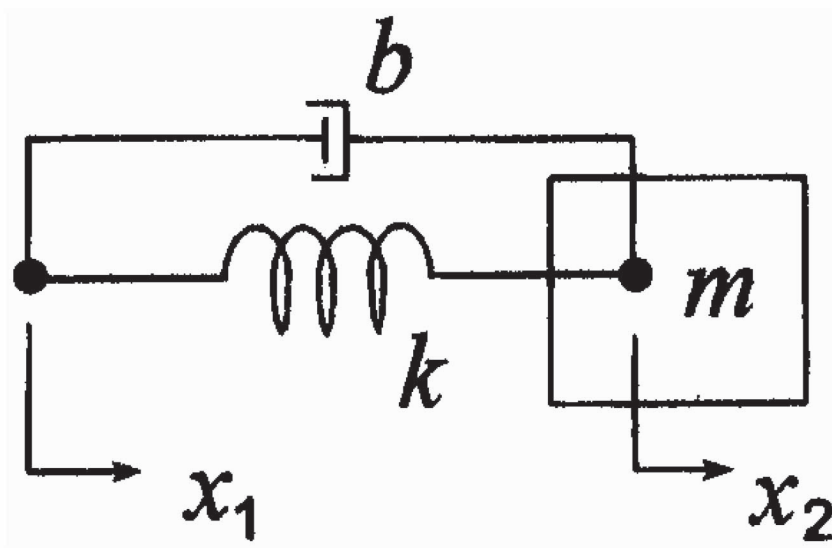
The Laplace transformation also converts the impulse response of a system variable to the transfer function for that variable. As a consequence, we can view the differential equation that

represents a linear system as an expression of the response of that system to an impulsive input.

## 94.1 Transfer Functions

The node displacements  $x_1$  and  $x_2$  and the compressive forces  $f_1$  and  $f_2$  within the branches of the lumped model of Fig. 94.1 are related to each other by the spring equation, the damper equation, and the balance of forces at node 2. The spring equation is  $f_2 = k(x_1 - x_2)$ . The equation for the damper is  $f_1 = b(\dot{x}_1 - \dot{x}_2)$ . The balance of forces requires that  $f_1 + f_2 = m\ddot{x}_2$ . These equations describe fully the behavior of the system if the spring and mass are unenergized. (If the mass were moving and/or the spring were compressed, we would have to express separately their initial energy states to describe fully the future relations among the variables.)

**Figure 94.1** The lumped model of a mechanical system.



Eliminate  $f_1$  and  $f_2$  from the equations to obtain the operational equation:

$$(mp^2 + bp + k)x_2 = (bp + k)x_1 \quad (94.1)$$

This differential equation describes fully the **zero-state** relation between  $x_1$  and  $x_2$ . Rearrange Eq. (94.1) to form the ratio



$$\frac{x_2}{x_1} = \frac{bp + k}{mp^2 + bp + k} \quad (94.2)$$

We call Eq. (94.2) the *transfer function* from  $x_1$  to  $x_2$ . The transfer function focuses attention on the mathematical operations that characterize the behavioral relationships rather than on the particular natures of the variables. [Note that the transfer function from  $v_1$  to  $v_2$ , where  $v_1 = px_1$  and  $v_2 = px_2$ , is the same as the transfer function given by Eq. (94.2).]

In general, suppose that  $y_1$  and  $y_2$  are two variables related (in operator notation) by the linear differential equation

$$y_2 = G(p)y_1 \quad (94.3)$$

We formally define the *transfer function* from  $y_1$  to  $y_2$  by

$$G(p) = \left. \frac{y_2}{y_1} \right|_{\text{ZS}} \quad (94.4)$$

where the notation ZS means **zero state**. If  $y_1$  is an independent variable, then  $G(p)$  is the *input-output transfer function* for the variable  $y_2$  and accounts fully for its behavior owing to the **input signal**  $y_1$ . We can determine from that transfer function the behavior of the system for any source waveform and any initial state.

## 94.2 The Laplace Transformation

---

The *one-sided Laplace transformation*,  $\mathcal{L}$ , is an integral operator that converts a signal  $f(t)$  to a complex-valued function  $F(s)$  in the following fashion:

$$\mathcal{L}[f(t)] \equiv F(s) \triangleq \int_{0^-}^{\infty} f(t)e^{-st} dt \quad (94.5)$$

We refer to the transformed function  $F(s)$  as the *Laplace transform* of the signal  $f(t)$ . Picture the lower limit  $0^-$  of the integral as a *specific* instant prior to but infinitesimally close to  $t = 0$ . It is customary to use a lowercase symbol ( $f$ ) to represent a signal waveform and an uppercase symbol ( $F$ ) to represent its Laplace transform. [Although we speak here of time signals, there is nothing in Eq. (94.5) that requires  $f(t)$  to be a function of time. The transformation can be applied to functions of any quantity  $t$ .]

We shall use the Laplace transformation to transform the signals of **time-invariant** linear systems. The behavior of such a system for  $t \geq 0$  depends only on the input signal for  $t \geq 0$  and on the prior **state** of the output variable (at  $t = 0^-$ ). Hence, it does not matter that the Laplace transformation ignores  $f(t)$  for  $t < 0^-$ .

The process of finding the time function  $f(t)$  that corresponds to a particular Laplace transform  $F(s)$  is called *inverse Laplace transformation*, and is denoted by  $\mathcal{L}^{-1}$ . We also call  $f(t)$  the

*inverse Laplace transform* of  $F(s)$ . Since the one-sided Laplace transformation ignores  $t < 0^-$ ,  $F(s)$  contains no information about  $f(t)$  for  $t < 0^-$ . Therefore, inverse Laplace transformation cannot reconstruct  $f(t)$  for  $t < 0^-$ . We shall treat all signals as if they are defined only for  $t \geq 0^-$ . Then there is a one-to-one relation between  $f(t)$  and  $F(s)$ .

To illustrate the Laplace transformation, we find the Laplace transform of the decaying exponential,  $f(t) = e^{-\alpha t}$ ,  $t \geq 0^-$ . The transform is

$$\begin{aligned} F(s) &= \int_{0^-}^{\infty} e^{-\alpha t} e^{-st} dt = \left. \frac{e^{-(s+\alpha)t}}{-(s+\alpha)} \right|_{0^-}^{\infty} \\ &= \left. \frac{e^{-(\sigma+\alpha)t} e^{-j\omega t}}{-(s+\alpha)} \right|_{0^-}^{\infty} = \frac{1}{s+\alpha} \quad \text{for } \operatorname{Re}[s] > -\alpha \end{aligned} \quad (94.6)$$

We must require  $\sigma > -\alpha$ , where  $\sigma$  is the real part of  $s$ , in order that the real-exponent factor converge to zero at the upper limit. (The magnitude of the complex-exponent factor remains 1 for all  $t$ .) Therefore, the Laplace transform of the decaying exponential is defined only for  $\operatorname{Re}[s] > -\alpha$ . This restriction on the domain of  $F$  in the complex  $s$  plane is comparable to the restriction  $t \geq 0^-$  on the domain of  $f$ .


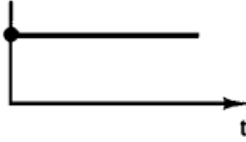

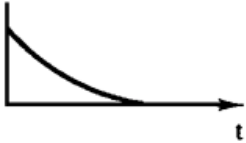
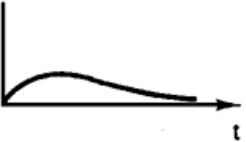

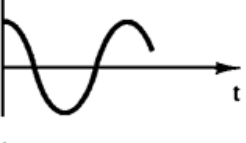
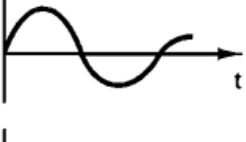
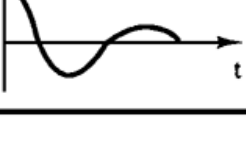
The significant features of the complex-frequency function  $1/(s+\alpha)$  are the existence of a single pole and the location of that pole,  $s = -\alpha$  rad/s. [The pole defines the left boundary of that region of the complex  $s$  plane over which the transform  $1/(s+\alpha)$  is defined.] The significant features of the corresponding time function are the fact of decay and the rate of decay, with the exponent  $-\alpha$  rad/s. There are clear parallels between the features of  $f(t)$  and  $F(s)$ . We should think of the whole complex-valued function  $F$  as representing the whole time waveform  $f$ .

As a second transformation example, let  $f(t) = \delta(t)$ , the unit impulse, essentially a unit-area pulse of very short duration. It acts at  $t = 0$ —barely within the lower limit of the Laplace integral. It has value zero at  $t = 0^-$ . (Because we use  $0^-$  as the lower limit of the defining integral, it does not matter whether the impulse straddles  $t = 0$  or begins to rise at  $t = 0$ .) The impulse is nonzero only for  $t \approx 0$ , where  $e^{-st} \approx 1$ . Therefore, the Laplace transform is

$$\Delta(s) = \int_{0^-}^{\infty} \delta(t) e^{-st} dt \approx \int_{0^-}^{\infty} \delta(t) (1) dt = 1 \quad (94.7)$$

It is not necessary to derive the Laplace transform for each signal that we use in the study of systems. [Table 94.1](#) gives the transforms for some signal waveforms that are common in dynamic systems.

**Table 94.1** Laplace Transform Pairs

	$f(t) = \mathcal{L}^{-1}[F(s)], t \geq 0^-$	$F(s) = \mathcal{L}[f(t)]$
1. Unit impulse $\delta(t)$		1
2. Unit step $u_s(t)$		$\frac{1}{s}$
3. $t^n, \quad n = 1, 2, \dots$		$\frac{n!}{s^{n+1}}$
4. $e^{-\alpha t}$		$\frac{1}{s + \alpha}$
5. $t^n e^{-\alpha t}, \quad n = 1, 2, \dots$		$\frac{n!}{(s + \alpha)^{n+1}}$
6. $\sin(\omega_0 t)$		$\frac{\omega_0}{s^2 + \omega_0^2}$
7. $\cos(\omega_0 t)$		$\frac{s}{s^2 + \omega_0^2}$
8. $e^{-\alpha t} \sin(\omega_d t)$		$\frac{\omega_d}{(s + \alpha)^2 + \omega_d^2}$
9. $e^{-\alpha t} \cos(\omega_d t)$		$\frac{s + \alpha}{(s + \alpha)^2 + \omega_d^2}$

Source: Dorny, C. N. 1993. *Understanding Dynamic Systems*, p. 412. Prentice Hall, Englewood Cliffs, NJ. With permission.

## 94.3 Transform Properties

A number of useful properties of the Laplace transformation  $\mathcal{L}$  are summarized in [Table 94.2](#). According to the derivative property, the multiplier  $s$  acts precisely like the time-derivative operator, but in the domain of Laplace-transformed signals. When we Laplace transform the equation for an energy-storage element such as a mass or a spring, the derivative property automatically incorporates the prior energy state of the element—essentially the value of the variable at  $t = 0^-$ . When we Laplace transform the input-output system equation for a particular system variable, the derivative property automatically incorporates the whole prior system state. As a consequence, we can find the solution to the system equation without having to determine the initial conditions (at  $t = 0^+$ )—a considerable simplification of the solution process.

**Table 94.2** Properties of the Laplace Transformation,  $\mathcal{L}$

1. Magnification	$\mathcal{L}[af(t)] = aF(s)$
2. Addition	$\mathcal{L}[f_1(t) + f_2(t)] = F_1(s) + F_2(s)$
3. Derivative	$\mathcal{L}[\dot{f}(t)] = sF(s) - f(0^-)$
4. Derivatives	$\mathcal{L}[\ddot{f}(t)] = s^2F(s) - sf(0^-) - \dot{f}(0^-)$
5. Integral	$\mathcal{L}\left[\int_{0^-}^t f(t) dt\right] = \frac{F(s)}{s}$
6. Convolution	$\mathcal{L}\left[\int_0^t f_1(\lambda)f_2(t-\lambda) d\lambda\right] = F_1(s)F_2(s)$
7. Initial value	$f(0^+) = \lim_{t \rightarrow 0^+} f(t) = \lim_{s \rightarrow \infty} sF(s)$
8. Final value	$f(\infty) = \lim_{t \rightarrow \infty} f(t) = \lim_{s \rightarrow 0} sF(s) \quad \text{if finite}$
9. Definite integral	$\int_0^\infty f(t) dt = \lim_{s \rightarrow 0} F(s) \quad \text{if finite}$
10. Exponential decay	$\mathcal{L}[e^{-\alpha t} f(t)] = F(s + \alpha)$
11. Delay	$\mathcal{L}[f(t - t_0)u_s(t - t_0)] = e^{-t_0 s} F(s) \quad \text{for } t_0 \geq 0$
12. Time multiplication	$\mathcal{L}[tf(t)] = -\frac{dF(s)}{ds}$
13. Time division	$\mathcal{L}\left[\frac{f(t)}{t}\right] = \int_s^\infty F(s) ds$
14. Time scaling	$\mathcal{L}[f(at)] = \frac{F(s/a)}{a}$

Source: Dorny, C. N. 1993. *Understanding Dynamic Systems*, p. 413. Prentice Hall, Englewood Cliffs, NJ. With permission.

Since  $F(s)$  contains all information about  $f(t)$  for  $t \geq 0^-$ , it is possible to find some features of the signal  $f(t)$  from the transform  $F(s)$  without performing an inverse Laplace transformation.

Properties 7 through 9 of [Table 94.2](#) provide three of these features—namely, the initial value ( $t \rightarrow 0^+$ ), the final value ( $t \rightarrow \infty$ ), and the area under the waveform. The remaining properties in the table show the effect on the transform of various changes in the signal waveform.

The usual approach to finding inverse transforms is to use a table of transform pairs. That table might be stored in a software package such as CC, MATLAB, MAPLE, and so on. [Table 94.1](#) demonstrates that transforms of typical system signals are ratios of polynomials in  $s$ . A ratio of polynomials can be decomposed into a sum of *simple* polynomial fractions—a process referred to as *partial fraction expansion*. Hence, the inversion process can be accomplished by a computer program that incorporates a brief table of transforms.

## 94.4 Transformation and Solution of a System Equation

Suppose that an independent external source applies a specific velocity pattern  $v_1(t)$  to node 1 of [Fig. 94.1](#). To obtain the input-output system equation that relates the velocity  $v_2$  of node 2 to the input signal  $v_1$ , multiply Eq. (94.2) by  $p$  and substitute  $v_1$  for  $px_1$  and  $v_2$  for  $px_2$ . The result is

$$(mp^2 + bp + k)v_2 = (bp + k)v_1 \quad (94.8)$$

The two sides of Eq. (94.8) are identical functions of time. Therefore, the Laplace transforms of the two sides of Eq. (94.8) are equal. Since the Laplace transformation is linear (properties 1 and 2 of [Table 94.2](#)), and since the coefficients of the differential equation are constants, the Laplace transform can be applied separately to the individual terms of each side. The result is

$$\begin{aligned} m[s^2V_2(s) - sv_2(0^-) - \dot{v}_2(0^-)] + b[sV_2(s) - v_2(0^-)] + kV_2(s) \\ = b[sV_1(s) - v_1(0^-)] + kV_1(s) \end{aligned} \quad (94.9)$$

where the derivative properties of the Laplace transformation (properties 3 and 4 of [Table 94.2](#)) introduce the prior values  $v_1(0^-)$ ,  $v_2(0^-)$ , and  $\dot{v}_2(0^-)$  into the equation. According to Eq. (94.9), to fully determine the transform  $V_2(s)$  of the behavior  $v_2(t)$ , we must specify these prior values and also  $V_1(s)$ . It can be shown that specifying the three prior values is equivalent to specifying the energy states of the spring and mass.

Let us assume that the independent source applies the constant velocity  $v_1(t) = v_c$  beginning at  $t = 0$ . The corresponding transform, by item 3 of [Table 94.1](#) and property 1 of [Table 94.2](#), is  $V_1(s) = v_c/s$ . Substitute the transform  $V_1(s)$  into Eq. (94.9) and solve for

$$V_2(s) = \frac{(bs + k)v_c + ms\dot{v}_2(0^-) + bs[v_2(0^-) - v_1(0^-)] + ms^2v_2(0^-)}{s(ms^2 + bs + k)} \quad (94.10)$$

We could find the output signal waveform  $v_2(t)$  as a function of the model parameters  $m, k, b$ , the source-signal parameter  $v_c$ , and the prior state information  $v_1(0^-)$ ,  $v_2(0^-)$ , and  $\dot{v}_2(0^-)$ , but the

expression for the solution would be messy. Instead, we complete the solution process for specific numbers:  $m = 2$  kg,  $b = 4$  N · s/m,  $k = 10$  N/m,  $\dot{v}_2(0^-) = 0$  m/s<sup>2</sup>,  $v_1(0^-) = 0$  m/s,  $v_2(0^-) = -1$  m/s, and  $v_c = 1$  m/s. The partial-fraction expansion of the transform and the inverse transform, both obtained by a commercial computer program, are

$$V_2(s) = \frac{1}{s} - \frac{2s + 2}{(s + 1)^2 + 2^2} \quad (94.11)$$

$$v_2(t) = 1 - 2e^{-t} \cos(2t), \quad \text{for } t \geq 0 \quad (94.12)$$

We can take Laplace transforms of the system equations at any stage in their development. We can even write the equations directly in terms of transformed variables if we wish. The process of eliminating variables can be carried out as well in one notation as in another. For example, the operator  $G(p)$  in Eq. (94.3) represents a ratio of polynomials in the time-derivative operator  $p$ . Therefore, Laplace transforming the differential Eq. (94.3) introduces the prior values of various derivatives of  $y_1$  and  $y_2$ . If the prior values of all these derivatives are zero, then the Laplace-transformed equation is

$$Y_2(s) = G(s)Y_1(s) \quad (94.13)$$

where the operator  $p$  in Eq. (94.3) is replaced by the complex-frequency variable  $s$  in Eq. (94.13). It is appropriate, therefore, to define the transfer function directly in terms of Laplace-transformed signals:

$$G(s) = \left. \frac{Y_2(s)}{Y_1(s)} \right|_{\text{PV}=0} \quad (94.14)$$

where  $Y_1(s)$  and  $Y_2(s)$  are the Laplace transforms of the signals  $y_1(t)$  and  $y_2(t)$ , and the notation  $\text{PV} = 0$  means that the prior values (at  $t = 0^-$ ) of  $y_1(t)$  and  $y_2(t)$  and the various derivatives mentioned above in connection with Eq. (94.13) are set to zero. The *frequency domain* definition [Eq. (94.14)] is equivalent to the *time domain* definition [Eq. (94.4)].

Suppose that the input signal  $y_1(t)$  is the unit impulse  $\delta(t)$ . Then the response signal  $y_2(t)$  is the unit-impulse response of the system. Since the Laplace transform of the unit impulse is  $Y_1(s) = \Delta(s) = 1$  by entry 2 of [Table 94.1](#), Eq. (94.13) shows that the Laplace transform  $Y_2(s)$  of the unit-impulse response is identical to the zero-state transfer function (expressed in the transform domain).

The transfer function for a linear system has two interpretations. Both interpretations characterize the system. In the frequency domain, the transfer function  $G(s)$  is the multiplier that produces the response—by multiplying the source-signal transform, as in Eq. (94.13). In the time domain, we use a representative response signal—the impulse response—to characterize the system. The transfer function  $G(s)$  is the Laplace transform of that characteristic response.

## Defining Terms

**Input:** An independent variable.

**Input-output system equation:** A differential equation that describes the behavior of a single dependent variable.

**Output:** A dependent variable.

**Signal:** An observable variable; a quantity that reveals the behavior of a system.

**State:** The state of an  $n$ th-order linear system corresponds to the values of a dependent variable and its derivatives.

**Time invariant:** A system that can be represented by differential equations with constant coefficients.

**Zero state:** A condition in which no energy is stored or in which all variables have the value zero.

## References

Franklin, G. F., Powell, J. D., and Emami-Naeini, A. 1994. *Feedback Control of Dynamic Systems*, 3rd ed. Addison Wesley, Reading, MA.

Kuo, B. C. 1991. *Automatic Control Systems*, 6th ed. Prentice Hall, Englewood Cliffs, NJ.

Nise, N. S. 1992. *Control Systems Engineering*. Benjamin Cummings, Redwood City, CA.

## Further Information

A thorough mathematical treatment of Laplace transforms is presented in *Advanced Engineering Mathematics*, by C. Ray Wylie and Louis C. Barrett. *Understanding Dynamic Systems*, by C. Nelson Dorny, applies transfer functions and related concepts in a variety of contexts. The following journals publish papers that use transfer functions and Laplace transforms:

*IEEE Transactions on automatic control*. Published monthly by the Institute of Electrical and Electronics Engineers.

*IEEE Transactions on Systems, man, and Cybernetics*. Published bimonthly

*Journal of Dynamic Systems, Measurement, and Control*. Published quarterly by the American Society of Mechanical Engineers.

Sage, A. P. "Block Diagrams"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



[95.1 Elements of the Block Diagram](#)[95.2 Block Diagram Reduction](#)[95.3 Summary](#)**Andrew P. Sage***George Mason University*

It is often very helpful to have a structural or pictorial representation that describes the flows of information and physical quantities in a system. In this chapter, we will examine two essentially equivalent ways of accomplishing this goal. First, we will examine the use of linear system block diagrams. This is a very traditional and conventional approach of the control systems engineer. We shall show that this method leads to the same transfer function determination formulas as does the signal flow graph method in the concluding portion of the chapter.

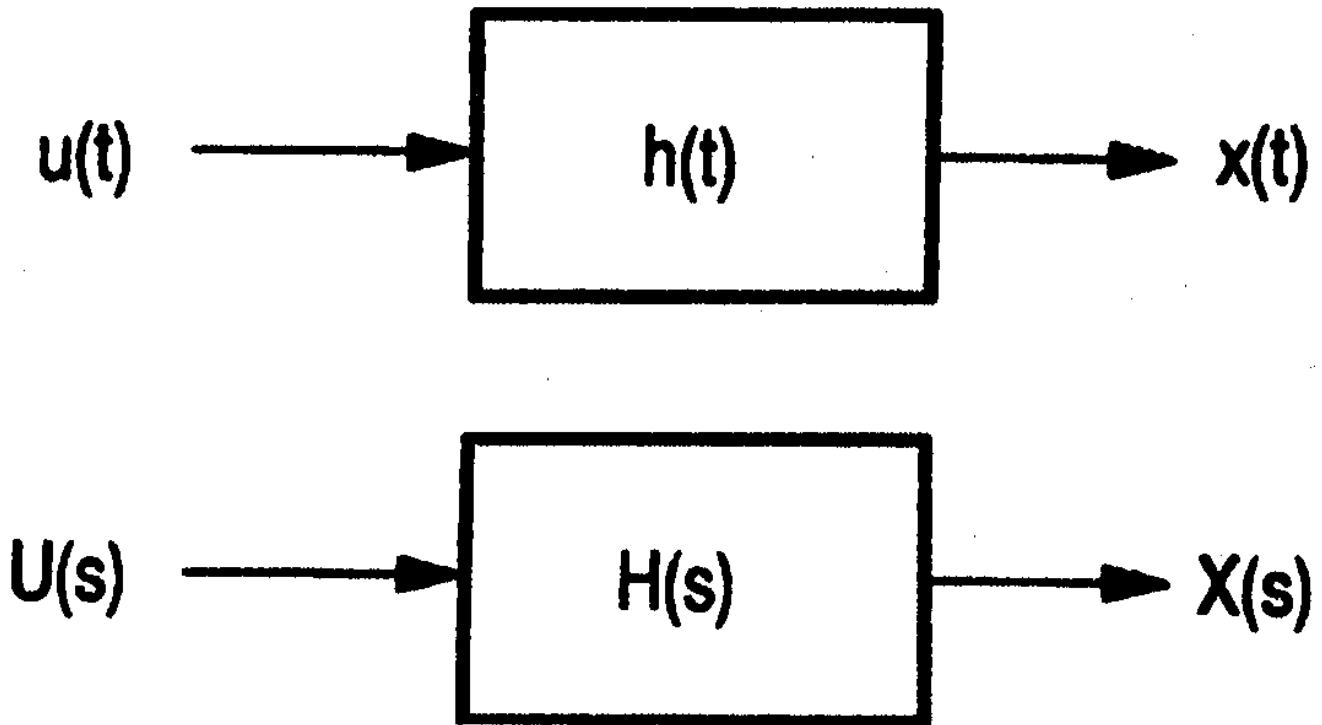
---

**95.1 Elements of the Block Diagram**

---

In previous chapters of this handbook, the Laplace transform of a system impulse response  $h(t)$  was defined as the system transfer function  $H(s)$ . It is often convenient to display this as a simple block diagram, as in [Fig. 95.1](#). Although it is proper to call either the time domain or frequency domain representations by the name **block diagram**, we will usually reserve the use of this term for the frequency domain representation. The block diagram may be a single-input, single-output block diagram that represents only the input-output behavior of the system. In this case we have a functional block diagram. On the other hand, the block diagram may be very detailed, such as to represent the entire physical structure of the systems. In either case the concepts developed here are applicable. Functional block diagram representations of input-output behavior are much more common than structural block diagrams in most control system design application studies. Structural representations that correspond to control system architectures are quite useful also. Generally, these have been converted to functional representations before control systems design efforts begin.

**Figure 95.1** Linear systems and block diagram representations.



**Linear systems** are often constructed by interconnecting subsystems. It will be of considerable value therefore to have a convenient method to cope with two (or more) coupled subsystems as indicated in Fig. 95.2. We must be very careful to describe precisely what we mean by the interconnection shown in this figure. The easiest way to accomplish this is to use a differential equation representation of the two subsystems. We assume that

$$\sum_{i=0}^{n_1} a_{1i} \frac{d^i x_1}{dt^i} = \sum_{i=0}^{m_1} b_{1i} \frac{d^i u_1}{dt^i} \quad (95.1)$$

describes the first system. The second system is assumed to be described by a similar relation, as given by

$$\sum_{i=0}^{n_2} a_{2i} \frac{d^i x_2}{dt^i} = \sum_{i=0}^{m_2} b_{2i} \frac{d^i u_2}{dt^i} \quad (95.2)$$

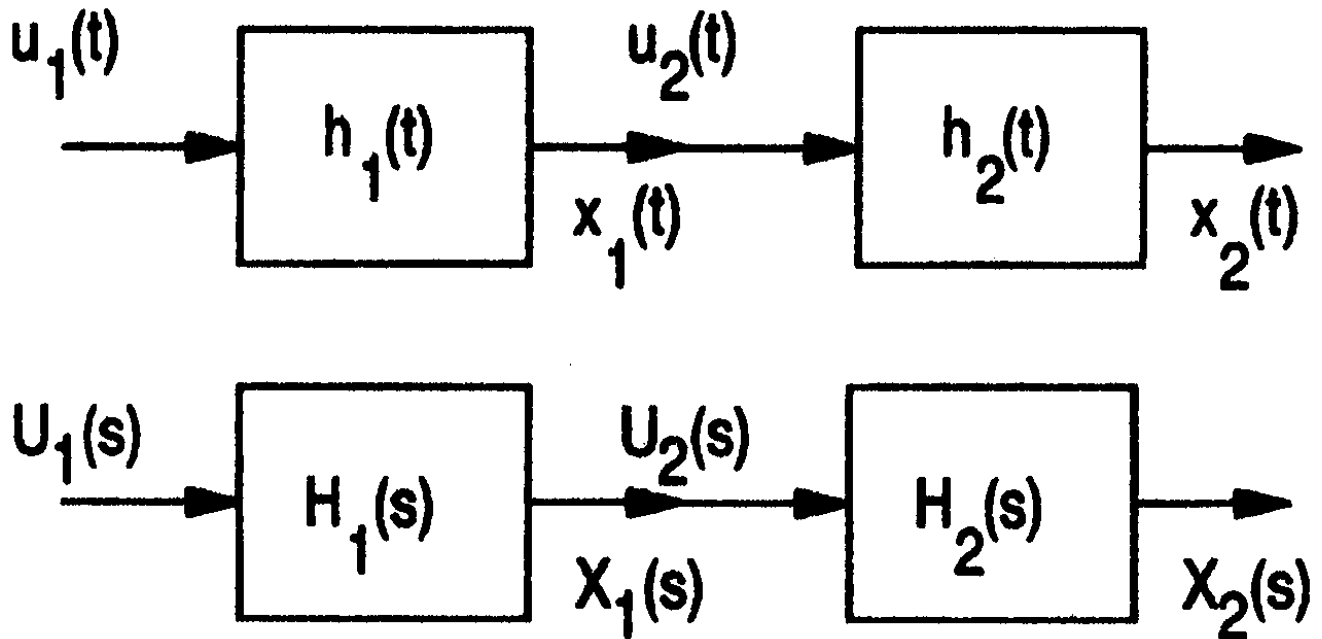
The systems are assumed to be interconnected such that the output from the first system serves as the input to the second system. This requires that

$$u_2 = x_1 \quad (95.3)$$

There are some physical requirements involving such factors as nonloading considerations that must be imposed to ensure that the output,  $x_1$ , of the first differential equation can be input to the

second differential equation,  $u_2$ , without altering the physical characteristics—and hence the differential equation—of the first system. Thus, there are a number of modeling assumptions involved, and loading considerations must be appreciated when attempting to cascade two systems together.

**Figure 95.2** Linear systems and block diagram representations for two cascaded system elements.



There are several ways in which we could describe this coupled system. We could use the impulse response characterization to write the output relations as

$$x_1(t) = \int_0^{\infty} h_1(\tau_1) u_1(t - \tau_1) d\tau_1 \quad (95.4)$$

$$x_2(t) = \int_0^{\infty} h_2(\tau_2) u_2(t - \tau_2) d\tau_2 \quad (95.5)$$

$$u_2(t) = x_1(t) \quad (95.6)$$

We can combine these three relations into a single relation that expresses the input-output dependency of the coupled system:

$$x_2(t) = \int_0^{\infty} \int_0^{\infty} h_2(\tau_2) h_1(\tau_1) u_1(t - \tau_1 - \tau_2) d\tau_1 d\tau_2 \quad (95.7)$$

Generally, this is a difficult integral to evaluate and shows very complex time domain coupling

between subsystems 1 and 2. Fortunately, the Laplace transform of this equation yields a very simple result, as we will now demonstrate. We use the fundamental definition of the Laplace transform,

$$X(s) = L[x(t)] = \int_0^{\infty} x(t)e^{-st} dt \quad (95.8)$$

and obtain, after a modest amount of relatively routine manipulation,

$$X_2(s) = H_1(s)H_2(s)U_1(s) \quad (95.9)$$

The transfer function of the interconnected system is just the ratio  $X_2(s)/U_1(s)$ . This ratio is simply the product of the two transfer functions, a very simple and desirable relationship:

$$H(s) = \frac{X_2(s)}{U_1(s)} = H_1(s)H_2(s) \quad (95.10)$$

A somewhat simpler proof of this product relationship can be obtained by determining the transfer function of the individual subsystems. We have, from Eqs. (95.4) through (95.6),

$$X_1(s) = \frac{\sum_{i=0}^{m_1} b_{1i} s^i}{\sum_{i=0}^{n_1} a_{1i} s^i} = \frac{Q_1(s)}{P_1(s)} = H_1(s) \quad (95.11)$$

$$X_2(s) = \frac{\sum_{i=0}^{m_2} b_{2i} s^i}{\sum_{i=0}^{n_2} a_{2i} s^i} = \frac{Q_2(s)}{P_2(s)} = H_2(s) \quad (95.12)$$

$$U_2(s) = X_1(s) \quad (95.13)$$

These are easily combined to yield Eq. (95.10).

We have just developed a very useful method to determine the transfer function of the two interconnected systems. It is also of interest to obtain the impulse response of an interconnected system. For the system considered here we can obtain impulse response from Eq. (95.10) by letting  $u_1(t)$  be an impulse occurring at  $t = 0$ . For  $u_1(t) = \delta(t)$ , we obtain from Eq. (95.10)

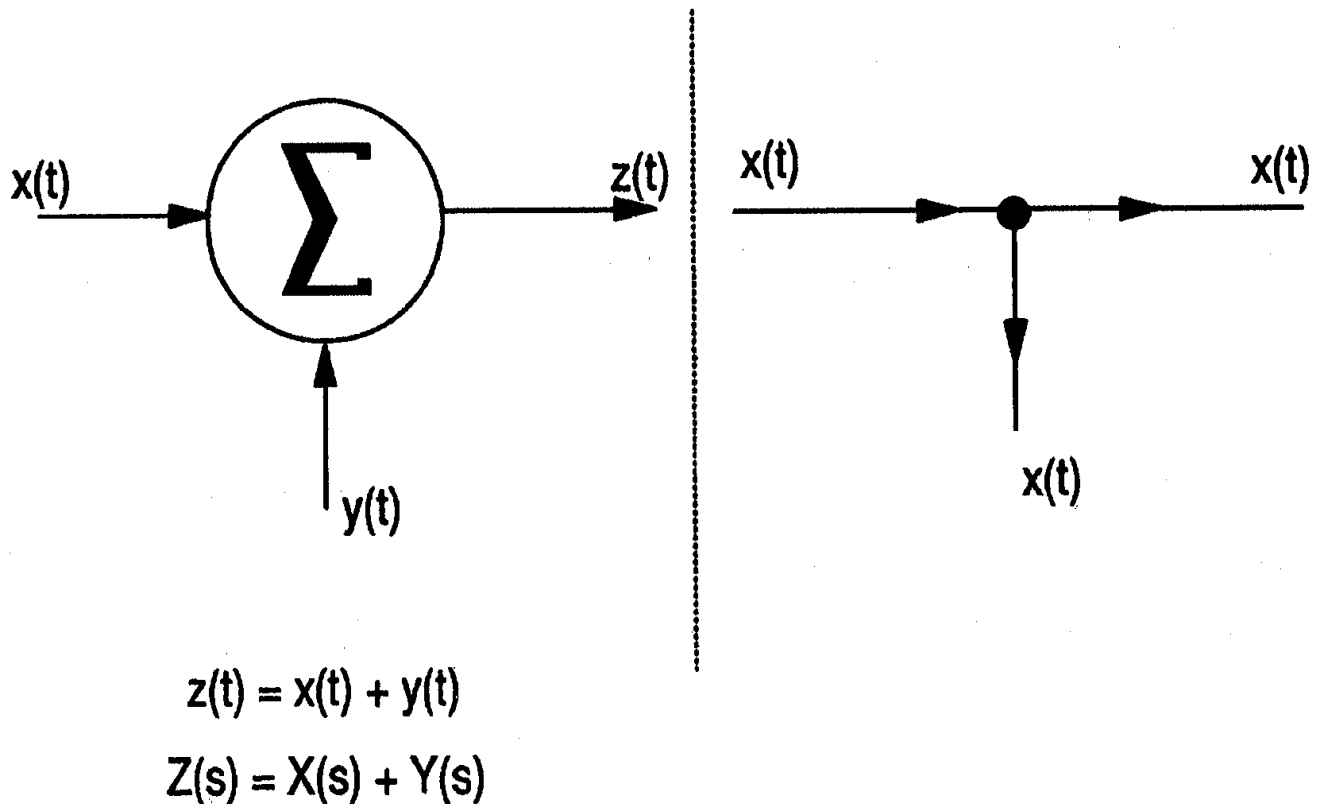
$$h(t) = \int_0^{\infty} h_2(\tau_2)h_1(t - \tau_2)d\tau_2 \quad (95.14)$$

and we see that the impulse response of the interconnected system is just the convolution of the impulse response of the individual systems. It will generally be simpler to obtain the transfer functions of systems such as these and manipulate these functions rather than convolve impulse responses.

Two other elements are needed in order to complete our description of the fundamental linear block diagram elements. These are the summation ("summer") and pick-off point, which are

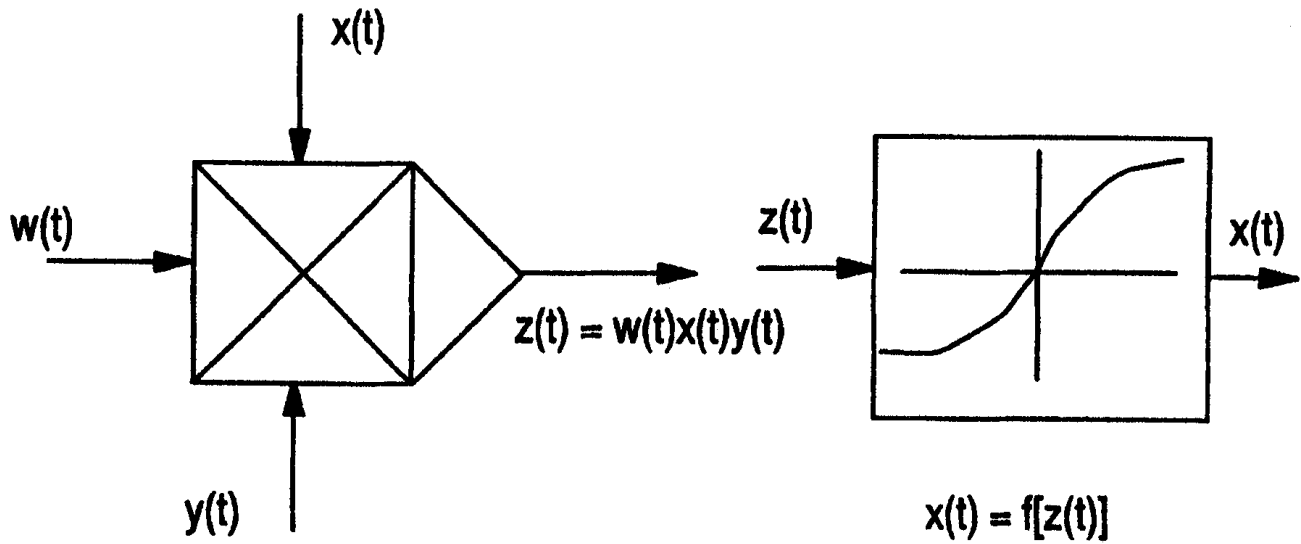
depicted in Fig. 95.3. We usually place signs alongside a summer to denote whether inputs to the summer are added or subtracted. If no signs are shown, all inputs are generally summed. As was the case with interconnection of subsystems, it is assumed that the physical model is such that no physical loading of the system occurs through use of these elements.

**Figure 95.3** Summation and pick-off symbols.



We could devise elements for multiplication and other nonlinear elements such as those shown in Fig. 95.4. The block diagram approach is primarily useful for **linear systems**, however, which is our primary concern here. Since this chapter is primarily concerned with linear systems, we shall not explore nonlinear elements in any detail here.

**Figure 95.4** Representation of nonlinear elements.



## 95.2 Block Diagram Reduction

A linear system block diagram is a representation of oriented lines and transfer function boxes, each of which is labeled with a variable and interconnect using summers and pick-off points. More often than not there will be one or more feedback loops in a typical linear system block diagram. One very simple diagram of this sort, representative of a linear-positioning servomechanism, is shown in Fig. 95.5. The presence of a closed loop whereby we can start at point  $A(s)$  and move through the system in the feed-forward position and return to point  $A(s)$  is noted. The term *closed loop system* is used to denote systems of this type in which there are feedback elements such that we may go through a complete path in the forward direction and return to the starting point. We may combine the forward transfer functions  $G_3(s)$  and  $G_4(s)$  and redraw the block diagram that represents the transfer function of this system. This step results in Fig. 95.6, which shows a single feedback loop. Often, this is called **major loop feedback**. Then we may combine the two elements in this reduced block diagram such that we obtain the single-block representation shown in Fig. 95.7. We use the elements in Fig. 95.6 and write the simple algebraic equations, or transfer function relations,

$$X(s) = G_2(s)A(s) \quad (95.15)$$

$$A(s) = U(s) - G_1(s)X(s) \quad (95.16)$$

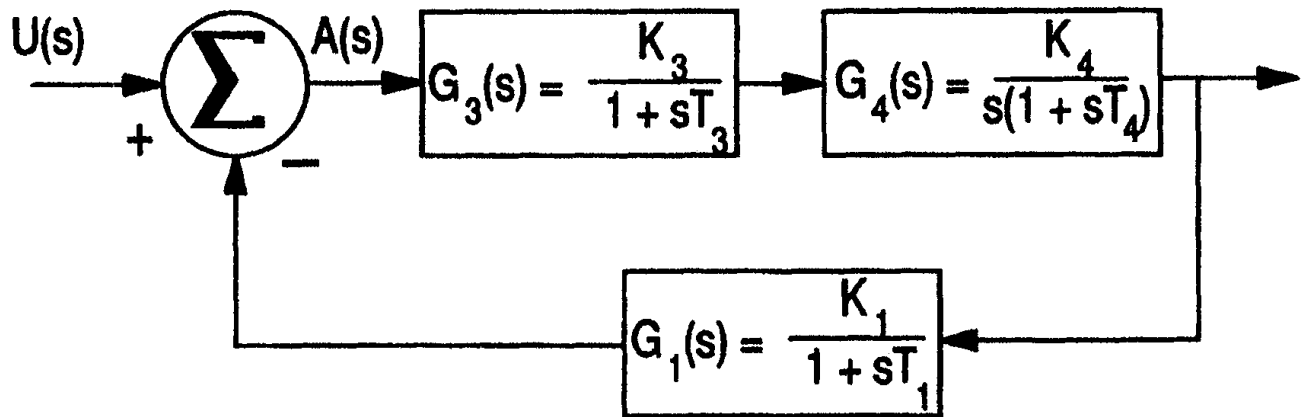
which we may solve for the desired input-output transfer relation as

$$\frac{X(s)}{U(s)} = H(s) = \frac{G_2(s)}{1 + G_1(s)G_2(s)} \quad (95.17)$$

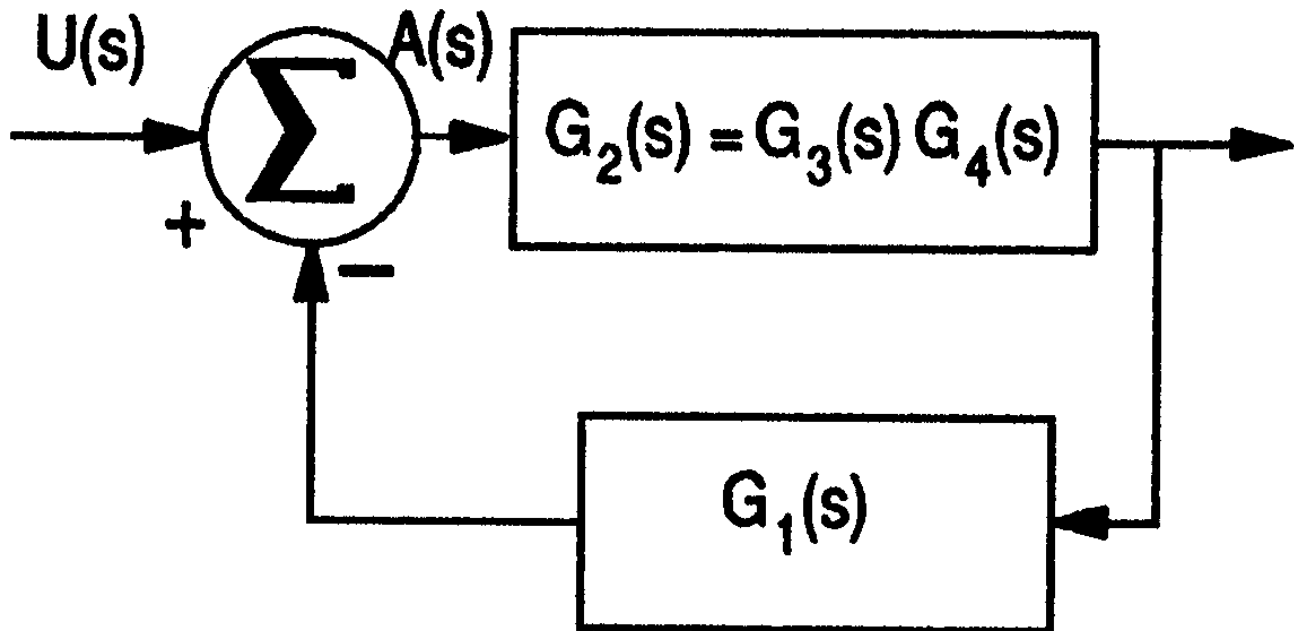
This simple relation is a very important one. We may restate it as follows: *The input-output*

transfer function for a single closed loop system is the forward transfer function divided by one plus the closed loop transfer function.

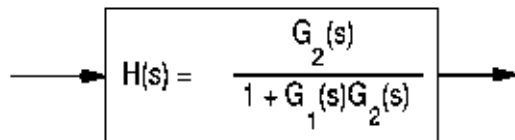
**Figure 95.5** Initial block diagram of simple positioning systems.



**Figure 95.6** Partially reduced block diagram.

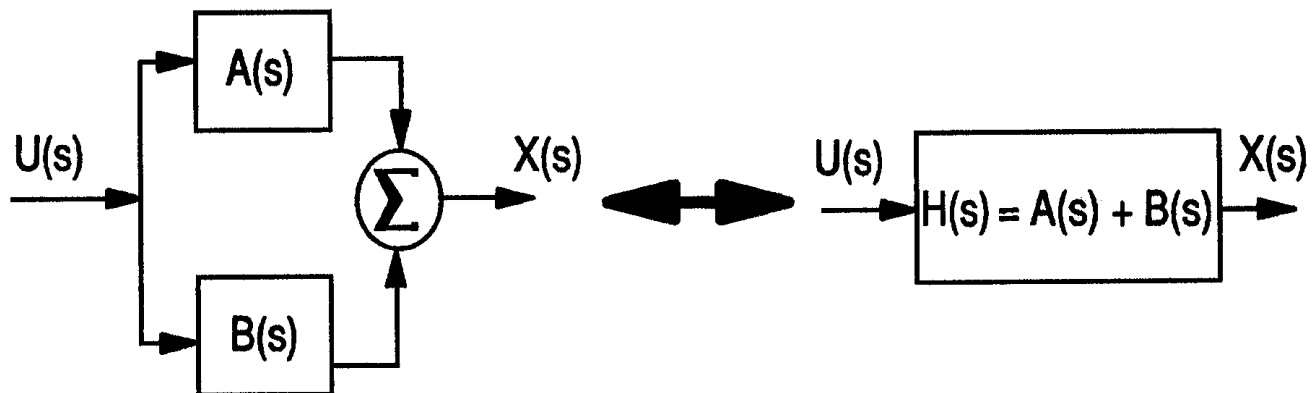


**Figure 95.7** Single-block equivalent of initial feedback system.

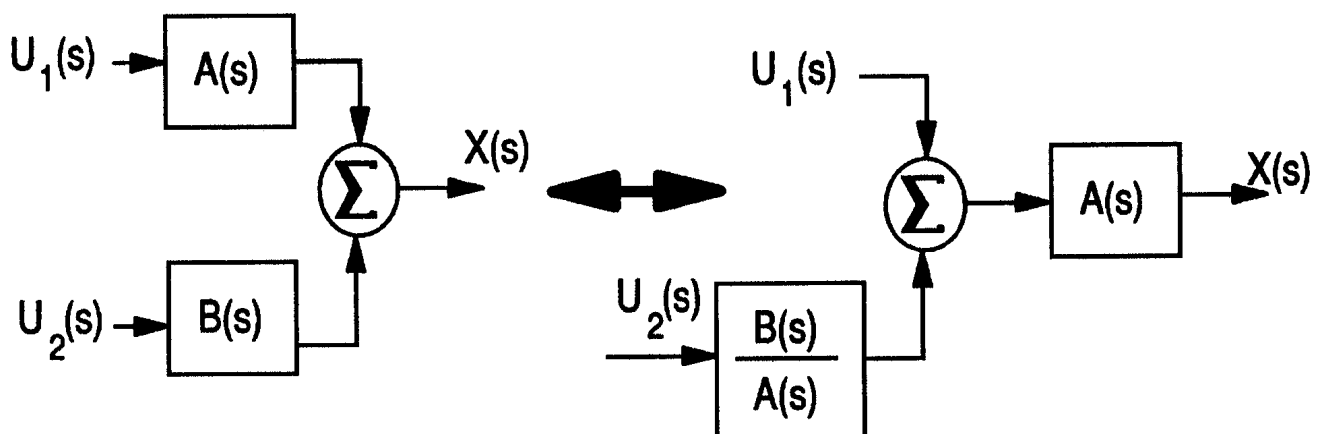


Three other rules for block diagram manipulation are helpful: a rule for combining transfer functions in parallel, and two rules for moving summing junctions and takeoff points around transfer functions. Figures 95.8 through 95.12 illustrate these rules as well as the rules for cascading transfer functions and feedback loop reduction that we have already obtained.

**Figure 95.8** Combining parallel transfer functions.

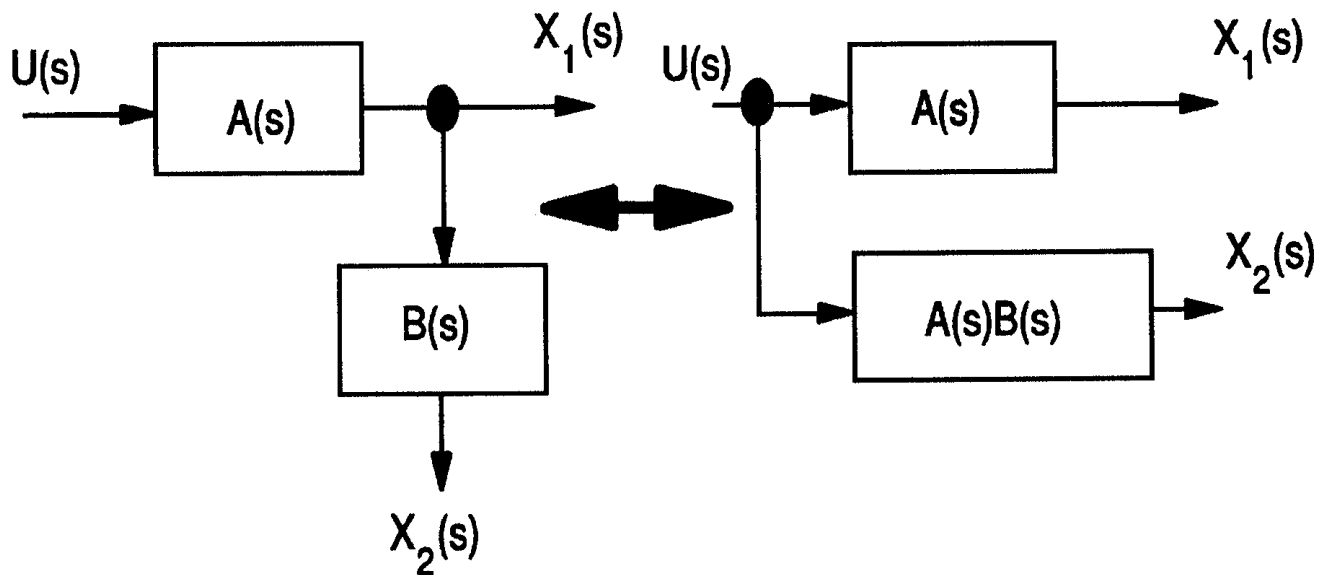


**Figure 95.9** Moving a summing junction.

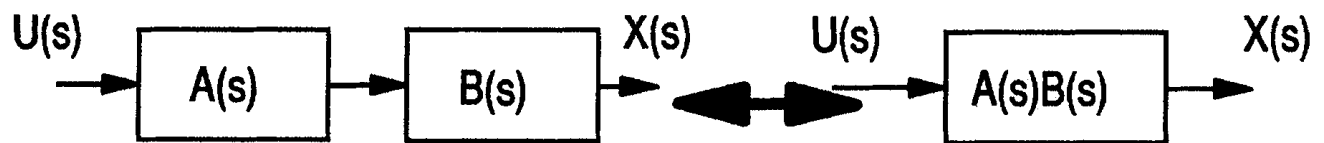




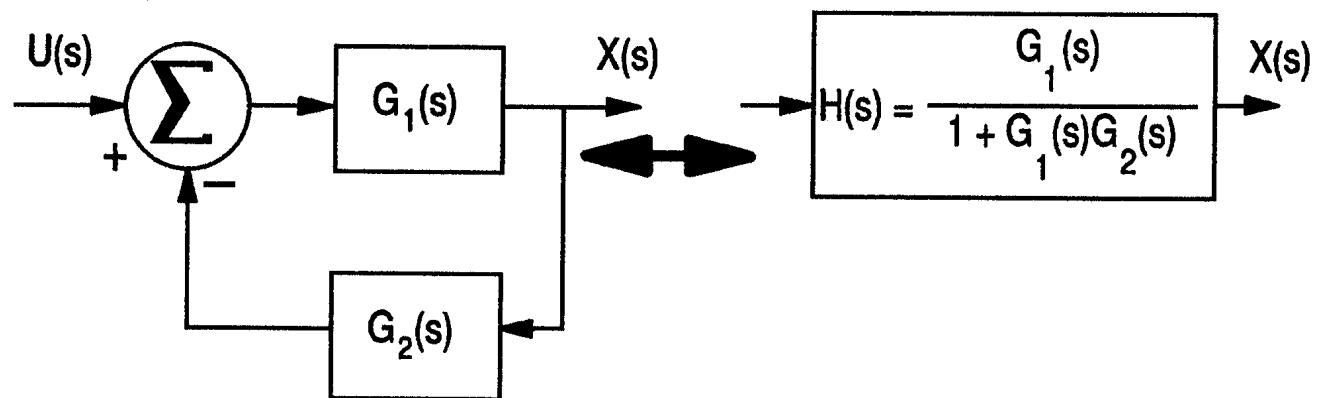
**Figure 95.10** Moving a takeoff point.



**Figure 95.11** Combining cascaded transfer functions.



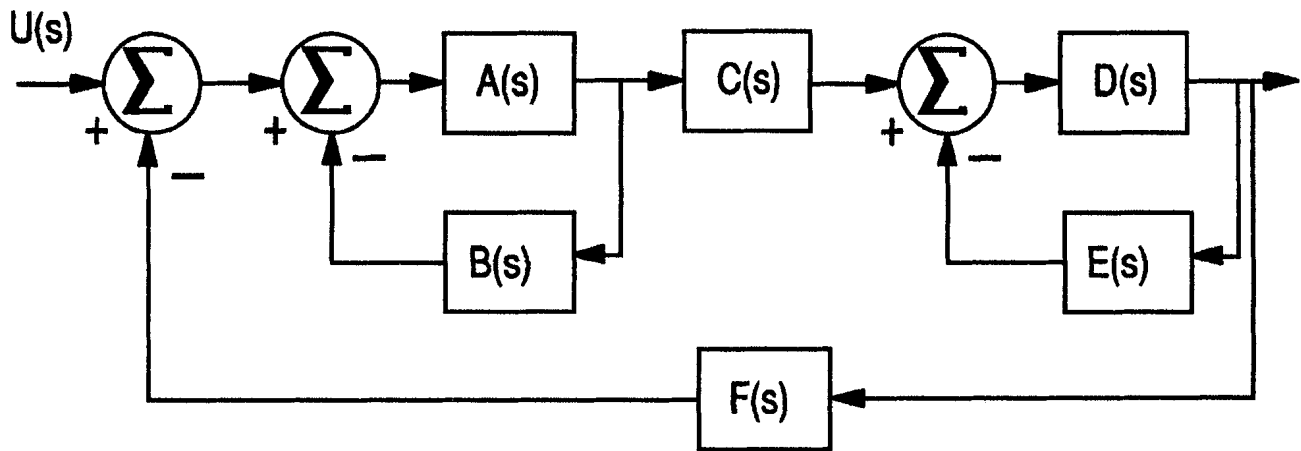
**Figure 95.12** Reduction of single-loop feedback system.



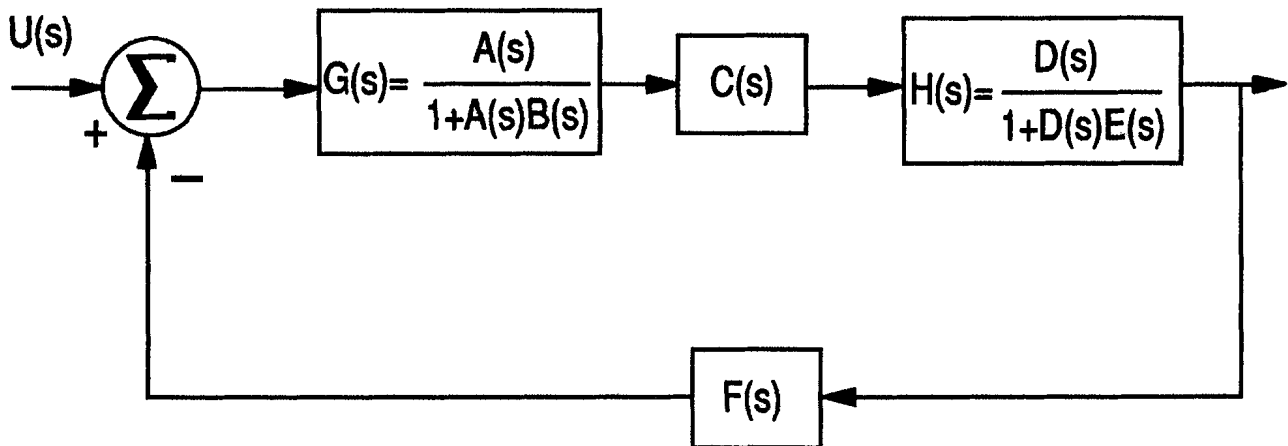
We could now proceed to establish general formulas for transfer functions of complex multiple-loop feedback systems. To motivate this, we first consider several examples that establish some preliminary ground rules and then present the more complex general rule. In each case we apply the rules established in Figs. 95.8 through 95.12.

**Example 95.1.** Consider the three-loop feedback system shown in Fig. 95.13. We want to obtain the input-output transfer function. Each of the single-loop feedback systems may be reduced to yield Fig. 95.14 by use of the rule for obtaining a single block from a feedback loop, as in Fig. 95.12. The cascade rule of Fig. 95.11 then allows us to represent the system as in Fig. 95.15. Use of the simple feedback rule of Fig. 95.12 allows us to obtain the final single-block transfer function of Fig. 95.16. Through this process we have reduced the complex three-loop system to an equivalent transfer function. This is the end of the example as far as block diagram reduction is concerned.

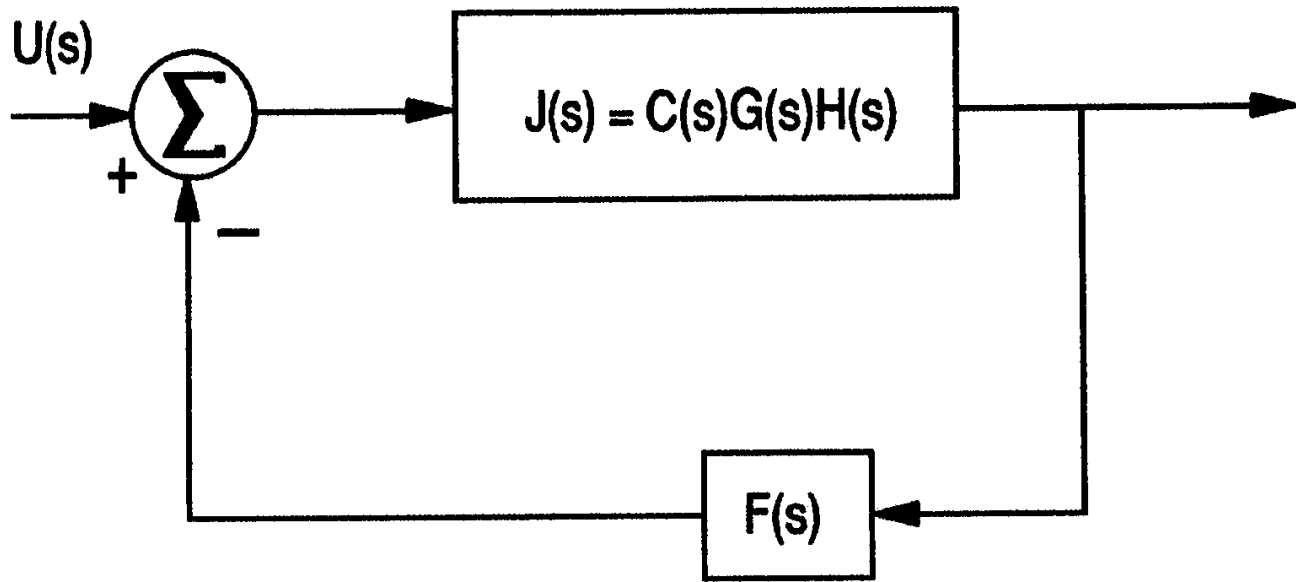
**Figure 95.13** Original feedback system.



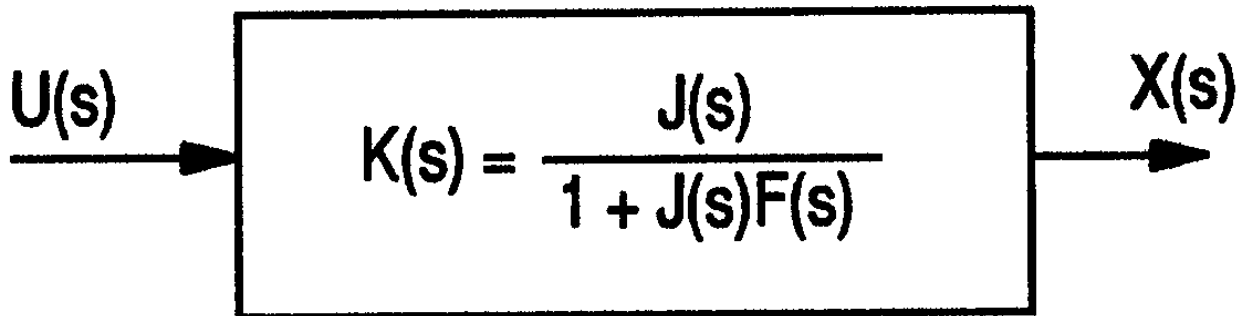
**Figure 95.14** System after reduction of two feedback loops.



**Figure 95.15** Feedback control system after application of cascade transfer function rule.



**Figure 95.16** Final single-blockdiagram.



Before stating a general formula for multiple-loop feedback systems, it is first desirable to give some further interpretation of the overall system transfer function, which is

$$H(s) = \frac{ACD}{1 + AB + DE + ACDF + ABDE} \quad (95.18)$$

In this expression we have deleted the  $(s)$  arguments for convenience. Notice that the numerator of this expression is just the forward transfer function from  $U(s)$  to  $X(s)$ . The denominator contains one minus the sum of the closed loop transfer functions, whose sum is  $AB + DE + ACDF$ . Also in the denominator is the term  $ABDE$ , which is the product of two of the closed loop transfer functions. We naturally ask why there are no products with the third loop transfer function,  $ACDF$ . A little examination reveals that such a circumstance arises because this loop contains elements in common with other closed loops. Thus it seems that a plausible rule, which certainly

works for this example, is that the overall input-output transfer function is the forward transfer function divided by one plus the sum of the loop gains plus the product of loop gains taken two at a time with nontouching nodes or common paths or elements.

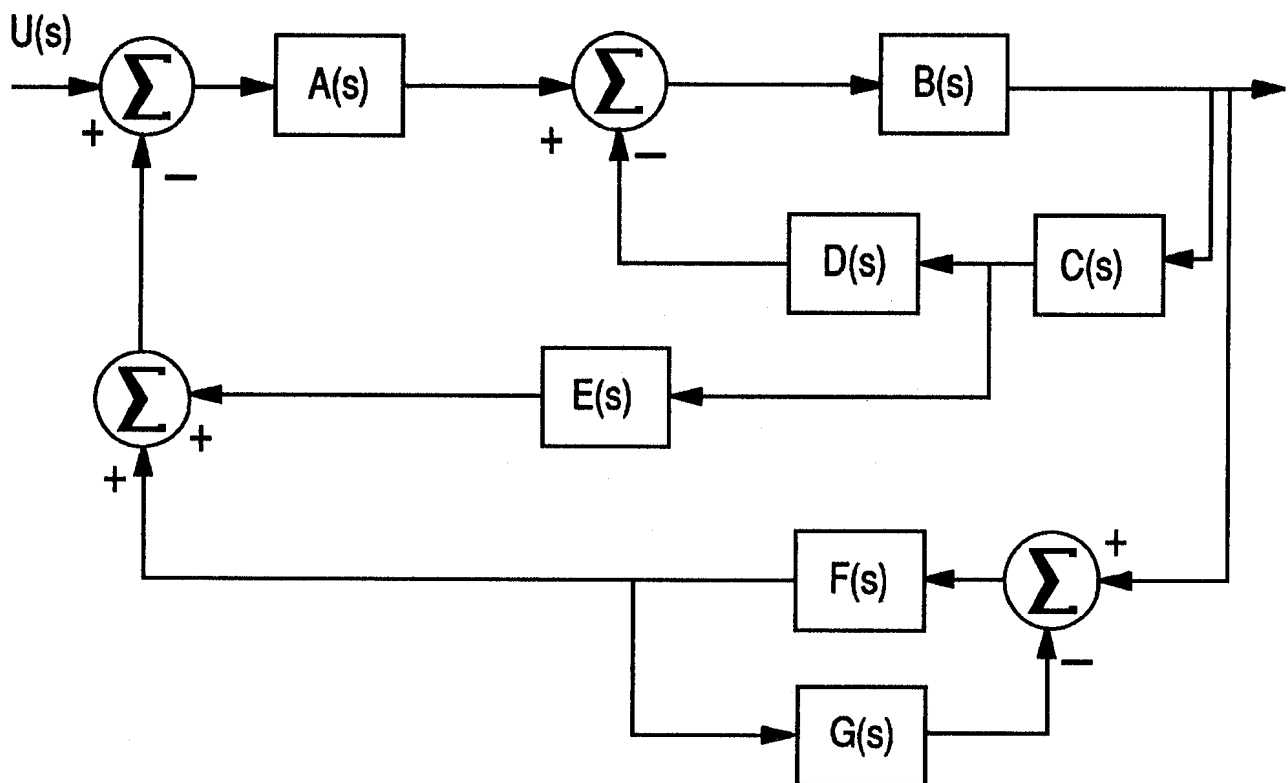
This observation turns out to be a correct rule for this system but not a complete rule, as we see in our next example, in which there is a minor loop feedback loop.

**Example 95.2.** We now consider block diagram reduction for the system shown in Fig. 95.17, in which there is a minor feedback loop. The system shown is reduced by applying our block diagram reduction rules. Solution is left as an exercise for the interested reader; it serves as a useful check on understanding material covered up to this point. The final transfer function is determined, in terms of the basic transfer functions of Fig. 95.17, as the expression

$$H(s) = \frac{X(s)}{U(s)} = \frac{AB(1 + FG)}{1 + BCD + ABCE + ABF + FG + BCDFG + ABCEFG} \quad (95.19)$$

Inspection of this result indicates that the numerator of this expression is different from that which would have been computed using the rule established in the previous example. Evaluation, through induction, reveals that the numerator here is multiplied by all terms in the denominator that do not have a path element in common with the forward transfer function contained in the numerator.

**Figure 95.17** System to be reduced by block diagram reduction.



This is still not quite the most general statement of the transfer function formula. Extension of these examples to cases in which three or more closed loops have no paths in common would show that the general input-output transfer function formula is

$$H(s) = \frac{X(s)}{U(s)} = \sum_k \frac{G_k(s)\Delta_k(s)}{\Delta(s)} \quad (95.20)$$

where

$G_k(s)$  is the  $k$ th forward transfer function through the system from  $u(t)$  to  $x(t)$ . No feedback loops can be contained in any  $G_k(s)$ .

$\Delta(s) = 1 -$  (sum of all individual closed loop transfer functions)  $+$  (sum of products of all closed loop transfer functions taken two at a time with no common paths or elements)  $-$  (sum of products of all closed loop transfer functions taken three at a time with no common paths or elements)  $+$   $\cdots$ .

$\Delta_k(s)$  = all terms in  $\Delta(s)$  that do not have elements or paths in common with an element or path in  $G_k(s)$ .

The summation is taken over all forward transfer function paths in the system.

The inputs and outputs,  $u(t)$  and  $z(t)$ , must be true inputs and outputs and not just nodes within a more complex feedback system.

**Example 95.3.** This example uses the general transfer function formula of Eq. (95.20) to evaluate the input-output transfer function of the system shown in Fig. 95.18. Direct application of Eq. (95.20) yields

$$\Delta = 1 + A + B + ABC + D + AB + AD + BD + ABCD + ABD$$

$$G_k = AB$$

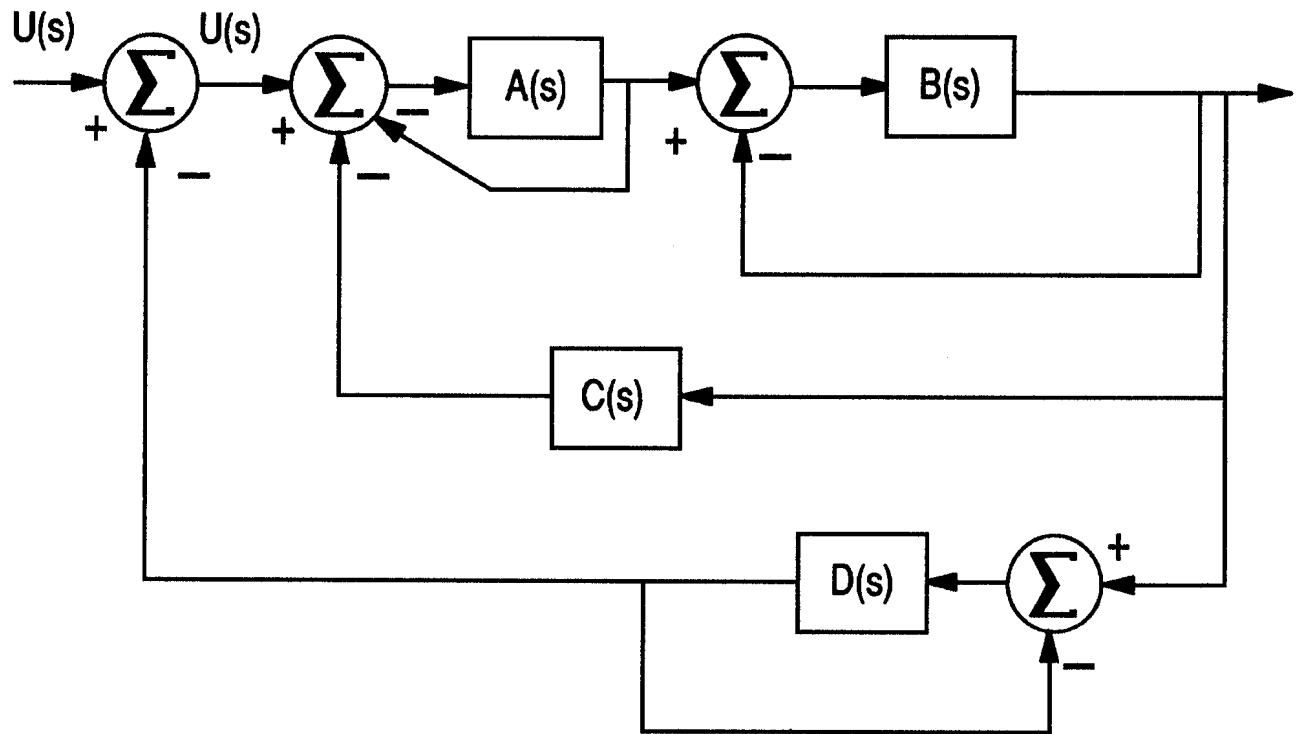
$$\Delta_k = 1 + D$$

and so we have

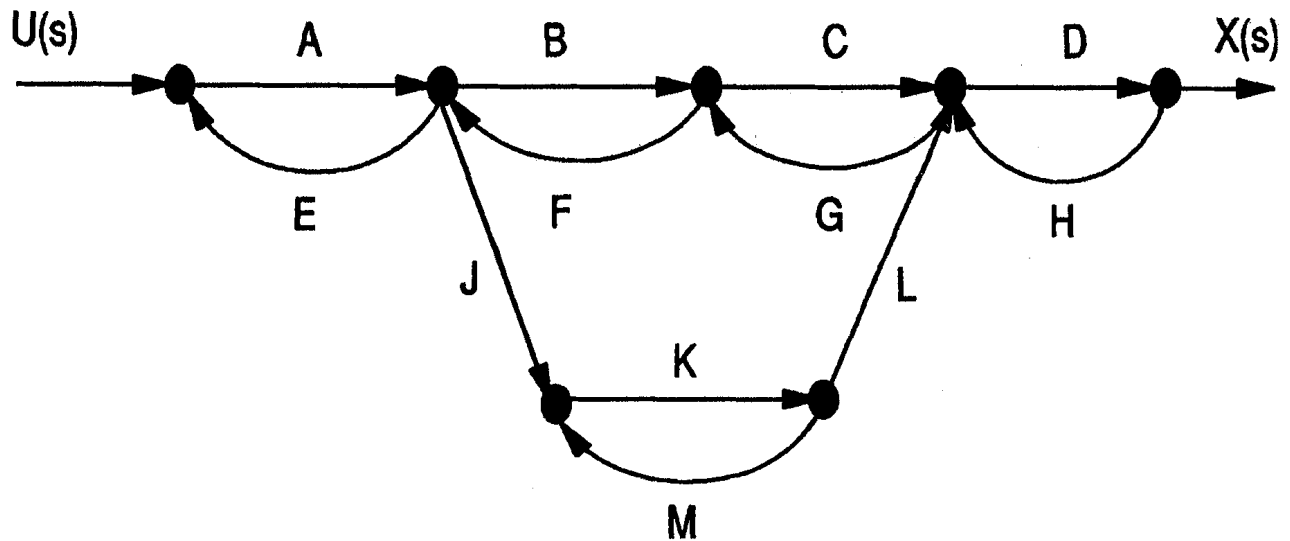
$$\begin{aligned} H(s) &= \frac{X(s)}{U(s)} = \frac{G_k \Delta_k}{\Delta} \\ &= \frac{AB(1 + D)}{1 + A + B + ABC + D + AB + AD + BD + ABCD + ABD} \end{aligned}$$

It would be of interest to obtain this transfer function from block diagram reduction as a check on our knowledge of the procedures involved.

**Figure 95.18** Block diagram with minor loop.



**Figure 95.19** A not-so-simple signal flow graph.



$$\frac{X(s)}{U(s)} = \frac{ABCD(1-KM)+AJKLD}{1-(AE+BF+CG+DH+KM+JKLFG)+AE(CG+DH+KM)+BF(DH+KM)+CG(KM)+DH(KM)-KM[AE(CG+DH)+BFDH]}$$

It turns out that we have obtained what was initially called the *general gain expression* for the gain, or transfer function, of a **signal flow graph**. The initial representation used for the signal flow graph is that illustrated in Fig. 95.19. Addition is accomplished at every node in a signal flow graph. To obtain subtraction, all we need to do is use a negative sign for the transfer function associated with appropriate path gains. Thus, we see that we have also developed an approach that may be used to evaluate the generalized gain, or transfer function, of a signal flow graph.

## 95.3 Summary

---

In this chapter of Section XVI, which has been devoted to a study of linear systems and models, we have studied block diagrams and block diagram reduction procedures. These are very useful for analysis and design of linear control systems and are among the many tools for analysis and design of linear systems.

### Defining Terms

**Block diagram:** A block diagram is generally a functional or input-output representation of a system. A block diagram may be a detailed structural representation of the architecture in a system, although this is generally not the case.

**Linear system:** A linear system is formally one that is described by a linear differential equation. Often, but not necessarily, the differential equation is time-invariant.

**Major loop feedback:** When there is a single feedback loop associated with a system, that loop is called a *major feedback loop*. When there are many feedback loops associated with a system, the outermost loop is generally called a *major loop*, and the feedback loops on the interior of the major loop are called *minor loops*.

**Signal flow graph:** A graph theoretic equivalent of a linear block diagram. It is composed of nodes and elements between nodes that represent gain and transfer function elements.

### References

Kuo, B. C. 1991. *Automatic Control Systems*, 6th ed. Prentice Hall, Englewood Cliffs, NJ.

Mason, S. J. 1953. Feedback theory—Some properties of signal flow graphs. *Proc. IRE*. 41(9):1144–1156.

Mason, S. J. 1956. Feedback theory—Further properties of signal flow graphs. *Proc. IRE*. 44(7): 920–926.

Sage, A. P. 1978. *Linear Systems Control*. Matrix, Champaign IL.

### Further Information

Because the material reviewed here is quite basic to any beginning study of linear control systems, there are many references appropriate for this chapter. Among the favorites of the author are the classic papers by Mason that describe the signal flow graph procedures, a work by the author, and a very classic work by Kuo that has seen many new editions. There are a plethora of good control systems texts that discuss block diagrams. You may consult the one most available to you or any of those cited in other chapters of this handbook that relate to linear systems and control for ancillary approaches to the study of block diagrams.

Kraus, A. D. "Signal Flow Analysis"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



- 96.1 The Signal Flow Graph
- 96.2 Transmission Gain
- 96.3 Signal Flow Graph Algebra
- 96.4 The Mason Gain Rule

**Allan D. Kraus**

*Naval Postgraduate School*

The signal flow graph is a viable alternative to the block diagram, and its use has many advocates. Because of this, signal flow analysis has, to some extent, replaced block diagram analysis as a means for reducing a complex system to a single transfer block or transfer function. Perhaps the outstanding feature of the signal flow graph is the use of node points or **nodes** to represent signals and the use of directed line segments (called **branches** or transmission **paths**) between the nodes.

## 96.1 The Signal Flow Graph

---

The signal flow graph applies only to linear systems, and an introduction to its use can derive from a consideration of the set of simultaneous, linear algebraic equations

$$\begin{aligned}
 8x_1 - 2x_2 - 4x_3 &= u_1 \\
 -2x_1 + 4x_2 - x_3 &= 0 \\
 -4x_1 - x_2 + 10x_3 &= u_2
 \end{aligned} \tag{96.1}$$

Here, the  $u$ 's are *excitations* or *forcing functions*. It should be noted that these equations represent an oversimplified case, and that it is common to see the coefficients as functions of the Laplace transform variable,  $s$ , to provide an  $s$ -domain representation of a system of linear ordinary differential equations. Equations (96.1) may be written in matrix form as

$$\begin{bmatrix} 8 & -2 & -4 \\ -2 & 4 & -1 \\ -4 & -1 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} u_1 \\ 0 \\ u_2 \end{bmatrix}$$

If  $u_1 = 24$  and  $u_2 = -12$ , the solution vector is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 15/4 \\ 2 \\ 1/2 \end{bmatrix}$$

which is easily verified.

Equations (96.1) may be rearranged to a form that isolates each of the three dependent variables:

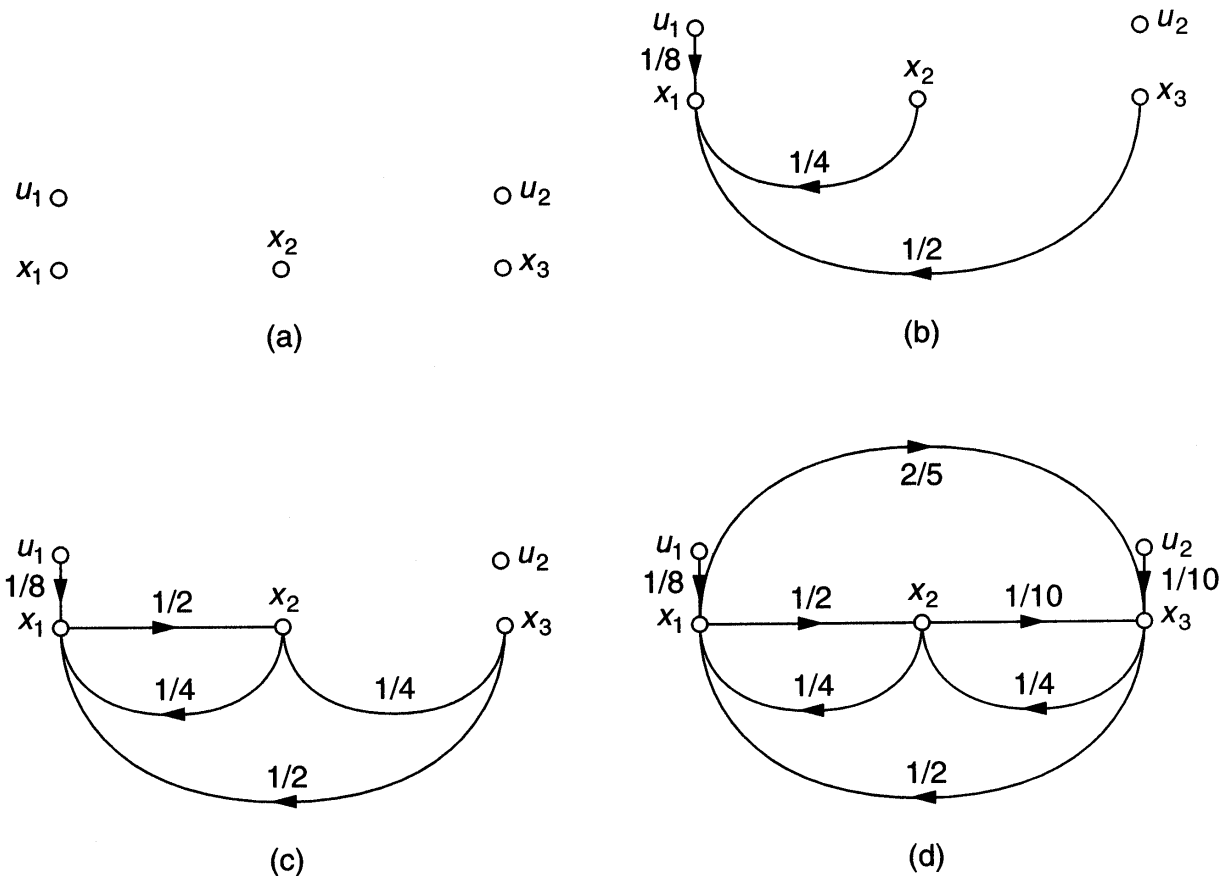
$$x_1 = \frac{1}{4}x_2 + \frac{1}{2}x_3 + \frac{1}{8}u_1 \quad (96.2a)$$

$$x_2 = \frac{1}{2}x_1 + \frac{1}{4}x_3 \quad (96.2b)$$

$$x_3 = \frac{2}{5}x_1 + \frac{1}{10}x_2 + \frac{1}{10}u_2 \quad (96.2c)$$

Observe that the objective here is to select a different dependent variable from each equation and to rewrite the equation in a manner such that the selected dependent variable is equal to a sum of terms involving the remaining variables. The signal flow graph representing these equations can then be constructed as indicated in Fig. 96.1. In this figure,  $x_1$ ,  $x_2$ , and  $x_3$  are considered as signals, and  $u_1$  and  $u_2$  are treated as inputs.

**Figure 96.1** Steps in the construction of the signal flow graph representing Eqs. (96.2). The final graph is indicated in (d).



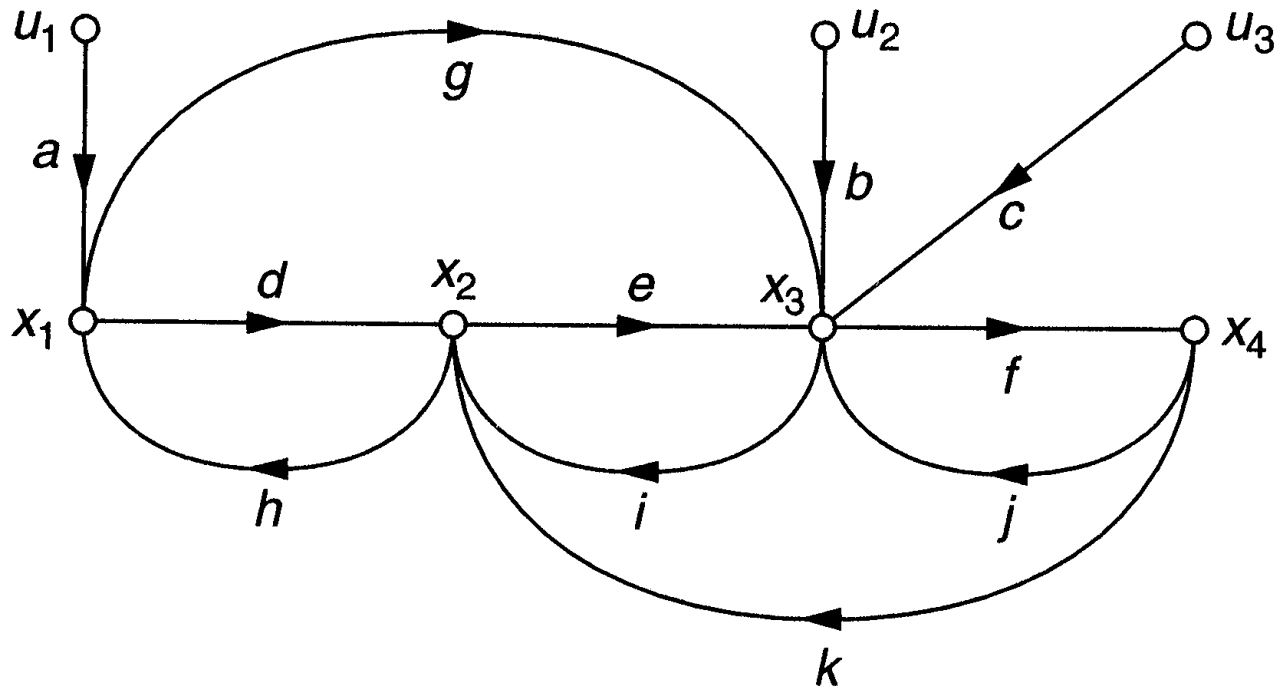
The graph is constructed by first defining nodes to represent all signals and inputs, with no connections at arbitrary points. Here the preference is to place all inputs at the top and all dependent variables in sequence, from left to right, below the inputs. This is indicated in Fig. 96.1(a). Then each equation in the system is added; Figs. 96.1(b) through 96.1(d) show how Eqs. (96.2a), (96.2b), and (96.2c) are handled. The net result is Fig. 96.1(d).

In Fig. 96.2, the inputs are designated  $u_1$ ,  $u_2$ , and  $u_3$ , and these inputs are located at the nodes so marked. The output nodes—

The nodes marked  $x_1$ ,  $x_2$ ,  $x_3$ , and  $x_4$  are output nodes because each of them contains at least one incoming branch.

are  $x_1$ ,  $x_2$ ,  $x_3$ , and  $x_4$ , and there are two paths from  $u_1$  to  $x_4$ :  $ade f$  and  $ag f$ . There are five loops, three of which are  $dh$ ,  $gfkh$ , and  $fke$ .

**Figure 96.2** A signal flow graph.



## 96.2 Transmission Gain

The form of the signal flow graph as well as the principle of linearity suggests that any of the outputs can be represented as a superposition of the inputs. In general, if there are  $n$  outputs, each represented by  $x_i$ , and  $m$  inputs, each represented by  $u_j$ , then each output  $x_i$  is related to each input  $u_k$  by a transmission gain  $T_{ik}$ :

$$x_i = \sum_{k=1}^{i=k} T_{ik} u_k \quad i = 1, 2, 3, \dots, n \quad (96.3)$$

Thus, in Fig. 96.1(d), there will be two transmission gains for each output. For example, for  $x_2$ ,

$$x_2 = T_{21} u_1 + T_{22} u_2$$

In Fig. 96.2, there are three transmission gains for each output  $w$ ,  $x$ ,  $y$ , and  $z$ , one each for  $u_1$ ,  $u_2$ , and  $u_3$ . Transmission gains are determined in an effective manner through an application of signal flow graph algebra, which is considered in the next section, or through the use of the *Mason gain rule*.

## 96.3 Signal Flow Graph Algebra

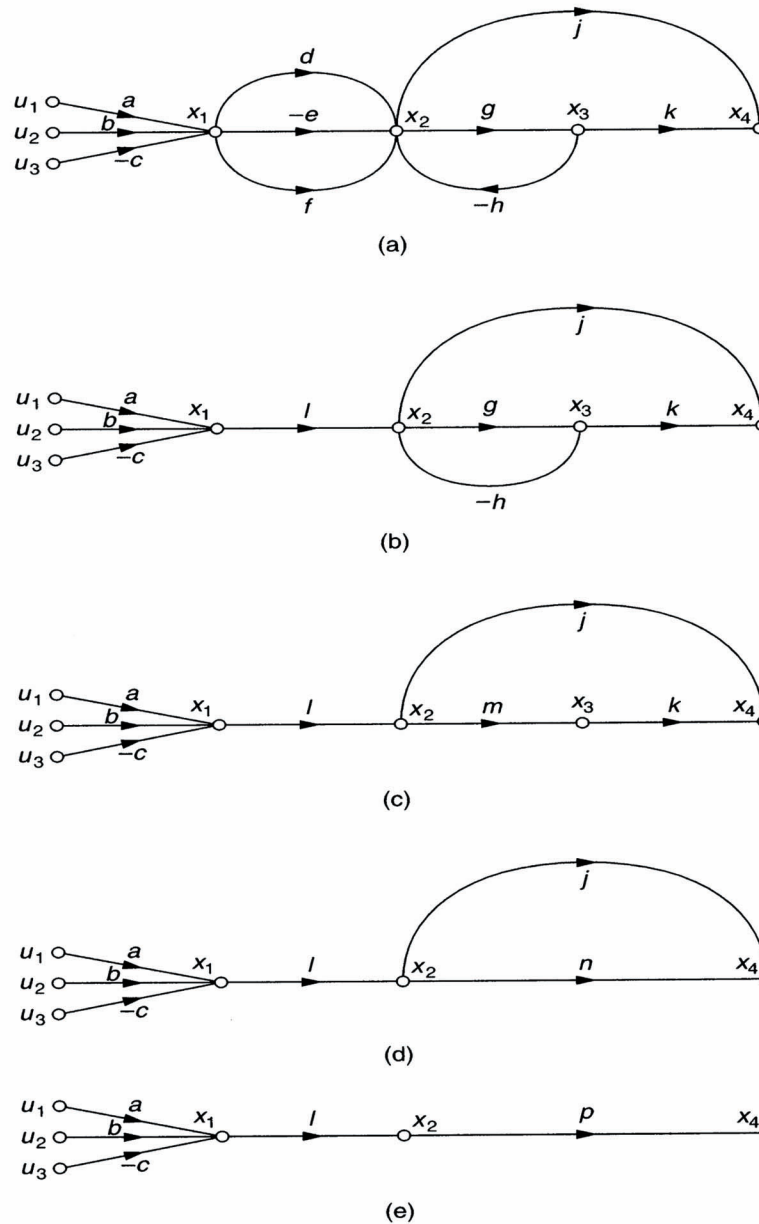
1. Incoming signals to a node add. That is, the value of a variable at any node will be equal to the sum of all of the signals that enter the node. For example, in Fig. 96.3 (a), the value of variable  $x_1$  is

$$x_1 = au_1 + bu_2 - cu_3$$

and the value at node  $x_4$  is

$$x_4 = jx_2 + kx_3$$

**Figure 96.3** Steps in the simplification of a particular signal flow graph. (a) The graph. (b) Simplification with the parallel branches from  $x_1$  to  $x_2$  replaced by a single branch,  $l$ . (c) Simplification with the feedback loop from  $x_2$  to  $x_3$  replaced by a single branch,  $m$ . (d) Simplification with the parallel branches from  $x_2$  to  $x_4$  replaced by a single branch,  $n$ . (e) Simplification with the parallel branches between  $x_2$  and  $x_4$  replaced by a single branch,  $p$ .



2. The value of the variable at any node is transmitted by all branches that leave the node. Notice in Fig. 96.3(a) that the value of the signal at the "upstream" ends of the branches with gains  $d$ ,  $-e$ , and  $f$  is  $x_1$ .
3. Parallel branches may be combined by addition, provided that they connect the same two nodes and are flowing in the same direction. In Fig. 96.3(a), the branches with gains  $d$ ,  $-e$ , and  $f$  are in parallel and flow from  $x_1$  to  $x_2$ . They may be replaced by a single branch with gain

$$l = d - e + f$$

as indicated in Fig. 96.3(b).

4. A feedback loop with forward transfer  $G(s)$  and feedback transfer  $H(s)$  may be replaced by a single branch whose gain,  $T(s)$ , is given by

$$T(s) = \frac{G(s)}{1 \mp G(s)H(s)} \quad (96.4)$$

where the plus (+) sign is used for negative feedback. In Fig. 96.3(b), the feedback loop with  $G(s) = g$  and  $H(s) = -h$  reduces to the single branch with gain

$$m = \frac{g}{1 + gh}$$

as shown in Fig. 96.3(c).

5. Single branches in series may be replaced by a single branch whose gain is the product of the gains of the individual branches. This is similar to the procedure for the cascade of transfer blocks in the block diagram, and it is observed that the two branches in series in Fig. 96.3(c) can be replaced by a single branch with gain

$$n = mk = \frac{gk}{1 + gh}$$

as shown in Fig. 96.3(d).

6. Additional combinations are possible. Observe that the parallel combination in Fig. 96.3(d) can be converted to a single branch

$$p = j + n = j + \frac{gk}{1 + gh}$$

or

$$p = \frac{j(1 + gh) + gk}{1 + gh}$$

as indicated in Fig. 96.3(e).

The overall transfer between the three inputs  $u_1$ ,  $u_2$ , and  $u_3$  to the single output  $x_4$  may now be obtained. Because this is a linear system, superposition must be applicable. Thus, with the series or cascade combination of the branches with gains  $l$  and  $p$  to form a single branch  $q$ ,

$$q = lp = (d - e + f) \left[ \frac{j(1 + gh) + gk}{1 + gh} \right]$$

one obtains  $x_4$  in terms of the inputs,  $u_1$ ,  $u_2$ , and  $u_3$ :

$$x_4 = \left[ \frac{(d - e + f)[j(1 + gh) + gk]}{1 + gh} \right] [au_1 + bu_2 - cu_3]$$

While simplification of a signal flow graph is always in order, a technique that does not rely on such simplification is available. This will be discussed in the next section.

## 96.4 The Mason Gain Rule

The transmission gains in a signal flow graph may be evaluated through the use of the Mason gain rule:

$$T_{ik} = \frac{x_i}{u_k} = \sum_j \frac{P_j \Delta_j}{\Delta} \quad (96.5)$$

where

$P_j$  is the  $j$ th path gain between the input  $u_k$  and the output  $x_i$

$\Delta$  is the graph determinant

$\Delta_j$  is the cofactor for the  $j$ th path

The graph determinant,  $\Delta$ , is determined from products involving the  $l$  loops in the graph:

$$\Delta = 1 - \sum L_i L_j + \sum L_i L_j L_k - \sum L_i L_j L_k L_m \cdots$$

where  $L_i L_j$  represents the products of all *nontouching* loops taken two at a time (doublets),  $L_i L_j L_k$  represents the products of all nontouching loops taken three at a time (triplets), and  $L_i L_j L_k L_m$  represents the products of all nontouching loops taken four at a time (quadruplets). Be aware that in signal flow graphs of even moderate complexity, it is possible to have quintuplets involving five loops, sextuplets involving six loops, and even more.

After the gain of the  $j$ th path has been determined, the cofactor,  $\Delta_j$ , of the  $j$ th path is formed by taking the graph determinant,  $\Delta$ , and striking out all terms associated with the loops that are touched by the  $j$ th path.

**Example.** Consider the system of Eqs. (96.1) and verify that  $x_3 = 1/2$  using a signal flow graph.

**Solution.** The signal flow graph of Fig. 96.1(d) is reproduced in Fig. 96.4 along with its five loops. Observe that all loops touch (loops that touch a single node are considered touching), so that the graph determinant is

$$\Delta = 1 - (L_1 + L_2 + L_3 + L_4 + L_5)$$

With the loop gains

$$L_1 = (1/2)(1/4) = 1/8 = 0.125$$

$$L_2 = (1/10)(1/4) = 1/40 = 0.025$$

$$L_3 = (2/5)(1/4)(1/4) = 1/40 = 0.025$$

$$L_4 = (2/5)(1/2) = 1/5 = 0.200$$

$$L_5 = (1/2)(1/10)(1/2) = 1/40 = 0.025$$

the graph determinant is computed as

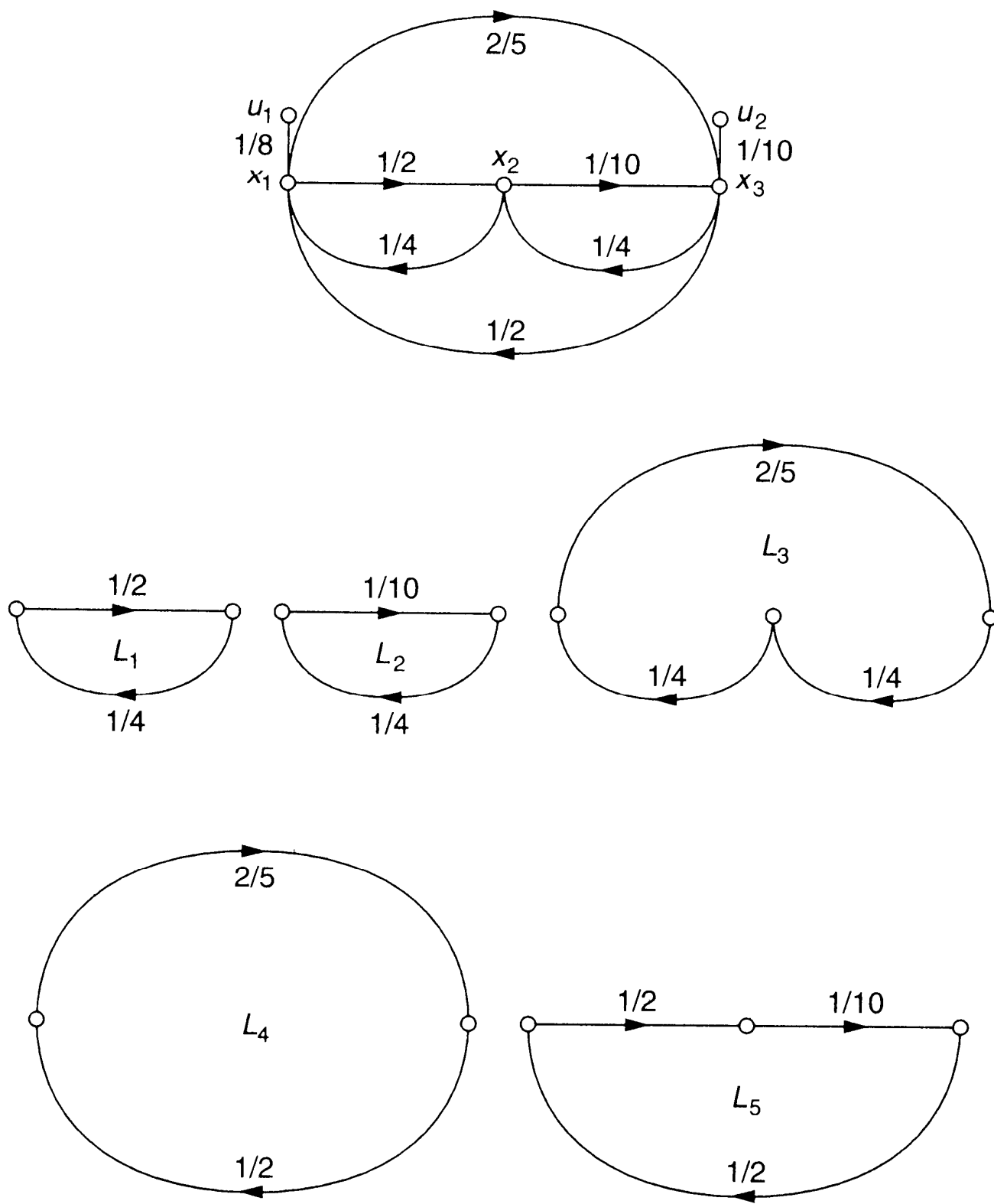
$$\Delta = 1 - (0.125 + 0.025 + 0.025 + 0.200 + 0.025) = 1 - 0.400 = 0.600$$

or

$$\Delta = 3/5$$

There are two paths between  $u_1$  and  $x_3$ . All loops touch both paths. Thus,  $\Delta_1 = \Delta_2 = 1$ , and

**Figure 96.4** The signal flow graph of Fig. 96.1(d) along with its five loops.





with

$$P_1 = (1/8)(1/2)(1/10) = 1/160$$

and

$$P_2 = (1/8)(2/5) = 1/20$$

the transmission gain will be in accordance with the Mason gain rule,

$$T_{31} = \frac{P_1 \Delta_1 + P_2 \Delta_2}{\Delta}$$

or

$$T_{31} = \frac{1/160 + 1/20}{3/5} = \frac{3}{32}$$

There is one path between  $u_2$  and  $x_3$ ,

$$P_1 = 1/10$$

but  $L_1$  does not touch this path. The cofactor of this path is therefore

$$\Delta_1 = 1 - L_1 = 1 - 1/8 = 7/8$$

and by the Mason gain rule,

$$T_{32} = \frac{P_1 \Delta_1}{\Delta}$$

or

$$T_{32} = \frac{(1/10)(7/8)}{3/5} = \frac{7}{48}$$

The value of  $x_3$  is determined from an application of Eq. (96.5). With  $u_1 = 24$  and  $u_2 = -12$ ,

$$\begin{aligned}
 x_3 &= T_{31} u_1 + T_{32} u_2 \\
 &= \frac{3}{32} 24 + \frac{7}{48} (-12) \\
 &= \frac{72}{32} - \frac{84}{48} \\
 &= \frac{1}{2}
 \end{aligned}$$

## Defining Terms

The following terminology pertains generally to all signal flow graphs and specifically to the graph in Fig. 96.1(d).

**Branch:** A directed line segment, having an associated gain, that connects two nodes in a graph. A signal  $x_1$  traveling along a branch between nodes  $x_1$  and  $x_2$  is multiplied by the gain of the branch. If the gain of the branch is unmarked, it is customary to assume that the gain is unity. Note well that signals travel along the branch only in the direction of the arrow; a branch that directs a signal from  $x_1$  to  $x_2$  indicates the dependency of  $x_2$  on  $x_1$ , and not the reverse. Just as in block diagram analysis, *unidirectionality* must hold.

**Loop:** A continuous sequence of branches, traversed in the indicated branch directions, from one node around a closed path back to the same node, in which no other node is encountered more than once.

**Nodes:** Points on the graph where the signals appear. At any node, quantities associated with incoming branches are added (summed), while outgoing branches have no effect on the signal that the node represents. In Fig. 96.1(d),  $x_1$ ,  $x_2$ ,  $x_3$ ,  $u_1$ , and  $u_2$  are all signals.

- *Input node:* A node at which there are only outgoing branches. In Fig. 96.1(d),  $u_1$  and  $u_2$  are input nodes.
- *Output node:* A node for which there is at least one incoming branch. This means that when all of the input nodes are identified, the rest of the nodes in the graph may be considered output nodes.

**Path:** A continuous sequence of branches, traversed as indicated by the branch directions, along which no node is encountered more than once.

## References

- Mason, S. J. 1953. Feedback theory: Some properties of signal flow graphs. *Proc. IRE*. 41(9): 1144–1156.
- Mason, S. J. 1956. Feedback theory: Further properties of signal flow graphs. *Proc. IRE*. 44(7): 920–926.
- Truxal, J. G. 1972. *Introductory Systems Engineering*. McGraw-Hill, New York.

Schimmel, B. D., Grantham, W. J. "Linear State-Space Models"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Linear State-Space Models

---

## 97.1 State-Space Models

### 97.2 Linearization

### 97.3 Linear System Representations

State-Space Systems • Input-Output Systems

### 97.4 Transforming System Representations

Controllability and Observability • State-Space Transformations • Input-Output Systems • State Equations from the Transfer Matrix

**Boyd D. Schimel**

*Washington State University*

**Walter J. Grantham**

*Washington State University*

## 97.1 State-Space Models

---

The general **state-space** model that we will consider for a continuous-time dynamical system consists of a system of  $n_x$  first-order differential equations of the form

$$\begin{aligned} \frac{dx_1}{dt} &= f_1(x_1, \dots, x_{n_x}, u_1, \dots, u_{n_u}) \\ &\vdots \\ \frac{dx_{n_x}}{dt} &= f_{n_x}(x_1, \dots, x_{n_x}, u_1, \dots, u_{n_u}) \end{aligned} \quad (97.1)$$

where  $t$  denotes time;  $\mathbf{x}(t) = [x_1 \dots x_{n_x}]^T$  is an  $n_x$ -dimensional **state vector**;  $\mathbf{u} = [u_1 \dots u_{n_u}]^T$  is an  $n_u$ -dimensional control or **input vector**; and  $[\cdot]^T$  denotes the transpose. Loosely speaking, the dimension of the state vector equals the number of initial conditions required to determine the motion, given the input—say  $\mathbf{u}(t)$  or  $\mathbf{u}(\mathbf{x}, t)$ —and the model of Eq. (97.1). As an example, the motion of a point mass  $m$  satisfying Newton's equations may be described by

$$m \frac{d^2 \mathbf{y}}{dt^2} = \mathbf{F}$$

where  $\mathbf{y} = [y_1 \ y_2 \ y_3]^T$  is the position vector and  $\mathbf{F} = [F_1 \ F_2 \ F_3]^T$  is the applied force. This second-order system can be converted to first-order state-space form by, for example, taking the state vector as  $\mathbf{x} = [y_1 \ y_2 \ y_3 \ dy_1/dt \ dy_2/dt \ dy_3/dt]^T$  and the input vector as  $\mathbf{u} = (1/m)\mathbf{F}$ , yielding

$$\dot{x}_1 = x_4$$

$$\dot{x}_2 = x_5$$

$$\dot{x}_3 = x_6$$

$$\dot{x}_4 = u_1$$

$$\dot{x}_5 = u_2$$

$$\dot{x}_6 = u_3$$

where  $(\dot{\phantom{x}}) = d(\phantom{x})/dt$ . This state-space formulation makes it clear that the required initial conditions include not only the initial position  $\mathbf{y}(t)$  but also the initial velocity  $\dot{\mathbf{y}}(0)$ . Writing dynamic system models as first-order differential equations also allows numerical analysts, for example, to focus on one type of numerical integration procedure, rather than having one scheme for first-order systems, another for second-order systems, and various other schemes for  $n$ th-order systems, since all systems of higher-order differential equations can be converted to equivalent systems of first-order differential equations, as illustrated in the preceding equations.

It should be noted that Eq. (97.1) does not explicitly cover systems governed by partial differential equations, such as the vibration of a drumhead, in which the motion of the object is distributed over space as well as time. Nor does Eq. (97.1) cover time-delay systems, in which the motion depends not only on the current state  $\mathbf{x}(t)$  and current input  $\mathbf{u}(t)$ , but also on a history of past conditions, such as the state  $\mathbf{x}(t - \tau)$  at some time  $\tau$  seconds in the past. Both of these situations could be (and typically are) converted to state-space models of the form in Eq. (97.1), by discretizing the spatial region for partial differential equations or discretizing the time-delay interval for time delay systems. In the limit as the discretization intervals become small the dimension of the state vector would become infinite.

In addition to the state and input vectors, there are typically various output or measurement quantities of interest, which are related to the state and input vectors by algebraic **output equations** of the general form

$$\begin{aligned} y_1 &= g_1(x_1, \dots, x_{n_x}, u_1, \dots, u_{n_u}) \\ &\vdots \\ y_{n_y} &= g_{n_y}(x_1, \dots, x_{n_x}, u_1, \dots, u_{n_u}) \end{aligned} \quad (97.2)$$

where  $\mathbf{y} = [y_1 \dots y_{n_y}]^T$  is an  $n_y$ -dimensional **output vector**. For our point-mass example, taking the position vector as the output would correspond to the output equations

$$y_1 = x_1$$

$$y_2 = x_2$$

$$y_3 = x_3$$

In vector form the state equations (97.1) and the output equations (97.2) can be written as

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}) \quad (97.3)$$

$$\mathbf{y} = \mathbf{g}(\mathbf{x}, \mathbf{u}) \quad (97.4)$$

where  $(\dot{\phantom{x}}) = d(\phantom{x})/dt$ ,  $\mathbf{f} = [f_1 \dots f_{n_x}]^T$ , and  $\mathbf{g} = [g_1 \dots g_{n_y}]^T$ . Systems in which time  $t$  appears explicitly in the right-hand sides of Eq. (97.3) or (97.4) can be handled, for example, either by including such terms in the input vector  $\mathbf{u}$  or by treating  $t$  as a state variable, say  $x_{n_x} = t$  with the differential equation  $\dot{t} = 1$ .

## 97.2 Linearization

---

For a system of the form

$$\dot{\mathbf{X}} = \mathbf{f}(\mathbf{X}, \mathbf{U}) \quad (97.5)$$

with output equations

$$\mathbf{Y} = \mathbf{g}(\mathbf{X}, \mathbf{U}) \quad (97.6)$$

let  $\mathbf{U}(t) = \bar{\mathbf{U}}(t)$  be a reference input. For a given initial state  $\bar{\mathbf{X}}(0)$  let  $\bar{\mathbf{X}}(t)$  be the corresponding solution to Eq. (97.5) generated by  $\bar{\mathbf{U}}(t)$  and let  $\bar{\mathbf{Y}}(t)$  be the resulting output, given by Eq. (97.6). Let  $\mathbf{x}(t)$ ,  $\mathbf{y}(t)$ , and  $\mathbf{u}(t)$  denote small deviations from the reference state, output, and input, respectively. Substituting  $\mathbf{X}(t) = \bar{\mathbf{X}}(t) + \mathbf{x}(t)$ ,  $\mathbf{Y}(t) = \bar{\mathbf{Y}}(t) + \mathbf{y}(t)$ , and  $\mathbf{U}(t) = \bar{\mathbf{U}}(t) + \mathbf{u}(t)$  into Eqs. (97.5) and (97.6), expanding the results using Taylor's theorem, and retaining only the first-order (linear) terms yields the time-varying **linear state equations**,

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u} \quad (97.7)$$

and the **linear output equations**

$$\mathbf{y} = \mathbf{C}(t)\mathbf{x} + \mathbf{D}(t)\mathbf{u} \quad (97.8)$$

where  $\mathbf{A}(t) = [a_{ij}(t)] = [\partial f_i / \partial X_j]$ ,  $i = \text{row}$ ,  $j = \text{column}$  is an  $n_x \times n_x$  matrix;  
 $\mathbf{B}(t) = [\partial f_i / \partial U_j]$  is an  $n_x \times n_u$  matrix;  $\mathbf{C}(t) = [\partial g_i / \partial X_j]$  is an  $n_y \times n_x$  matrix;  
 $\mathbf{D}(t) = [\partial g_i / \partial U_j]$  is an  $n_y \times n_u$  matrix; and all matrices are evaluated along the reference

conditions  $\bar{X}(t), \bar{U}(t)$ .

The most common occurrence of linear systems arises when the reference input is a constant  $\bar{U}(t) = \bar{U}$  and the reference state is an equilibrium (i.e., constant) state  $\bar{X}(t) = \bar{X}$ . Then the reference output  $\bar{Y}(t) = \bar{Y}$  is also constant. In this case the matrices in the linear model are also constant. This yields a constant-coefficient **multiple-input, multiple-output (MIMO)** linear state-space model, with linear state equations

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u \quad (97.9)$$

and the linear output equations

$$y = \mathbf{C}\mathbf{x} + Du \quad (97.10)$$

Henceforth we will be concerned only with constant-coefficient linear state-space models of the form given in Eqs. (97.9) and (97.10).

For the special case of a single input  $u$  and a single output  $y$  ( $n_u = n_y = 1$ ) we have a **single-input, single-output (SISO)** state-space system,

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u \quad (97.11)$$

$$y = \mathbf{C}\mathbf{x} + Du \quad (97.12)$$

where  $u$  and  $y$  are scalar variables,  $D$  is a scalar constant,  $\mathbf{B} = [b_1 \dots b_{n_x}]^T$ , and  $\mathbf{C} = [c_1 \dots c_{n_x}]$ .

## 97.3 Linear System Representations

---

The representation of a linear state-space system is generally not unique. For example, a change in the coordinate system for  $\mathbf{x}$  will change the matrices in the state-space representation. In addition, there are certain state-space representations and other types of representations that may be more convenient for various types of analyses.

### State-Space Systems

There are several special representations for a state-space system that occur frequently in control systems analysis:

#### Decoupled Form

The simplest representation of a state-space system occurs when the state equations are decoupled, that is, when the  $\mathbf{A}$  matrix is in the diagonalized form,

$$\mathbf{A} = \text{diag} [\lambda_1, \dots, \lambda_{n_x}] = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_{n_x} \end{bmatrix} \quad (97.13)$$

so that the evolution of each state variable depends only on itself and the inputs, but not on the other state variables.

### Block Diagonal Form

For a particular state-space system it may not be possible to completely diagonalize the  $\mathbf{A}$  matrix so that each state variable is decoupled from the others. In this case a generalization of the diagonal structure occurs where the  $\mathbf{A}$  matrix is in block diagonal form,

$$\mathbf{A} = \text{diag} [\mathbf{A}_1, \dots, \mathbf{A}_n] \quad (97.14)$$

in which the  $\mathbf{A}_i$  are square matrices on the diagonal of  $\mathbf{A}$ .

### Companion Form

For a special class of SISO state-space systems, a representation exists that can lead to a description of the system in terms of a single higher-order differential equation. A single-input, single-output state-space system is said to be in companion form if the  $\mathbf{A}$  matrix is a companion matrix,

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \cdots & 1 \\ -a_1 & -a_2 & -a_3 & \cdots & -a_{n_x} \end{bmatrix} \quad (97.15)$$

and the scalar control input enters only in the  $\dot{x}_{n_x}$  equation, with the column matrix  $\mathbf{B}$  of the form  $\mathbf{B} = [0 \ 0 \ \dots \ 1]^T$ . As we will see, a companion form representation has a particularly simple relationship to a representation involving a single  $n_x$ - order differential equation.

## Input-Output Systems

Equations (97.11) and (97.12) are the general representation of a constant-coefficient SISO system. Another common representation for a special class of SISO systems is referred to as an **input-output (IO)** representation. Using the notation  $y^{(n)} = d^n y / dt^n$  we call an SISO system an IO system if the system can be represented by a single  $n_x$ - order differential equation of the form

$$y^{(n_x)} + p_{n_x-1} y^{(n_x-1)} + \cdots + p_1 \dot{y} + p_0 y = q_0 u + q_1 \dot{u} + \cdots + q_{n_x} u^{(n_x)} \quad (97.16)$$

Note that the left-hand side of Eq. (97.16) is expressed in terms of the output  $y(t)$  and the right-hand side is expressed in terms of the input  $u(t)$ . In order for the output to depend on the input, at least one of the right-hand coefficients must be nonzero.

There is a close relationship between the state-space representation of an SISO system in companion form and the IO representation of the same system, provided that certain conditions are



satisfied. For example, if the output is just the first state component—that is, the  $\mathbf{C}$  matrix is of the form  $\mathbf{C} = [1 \ 0 \dots 0]$  and  $D = 0$ , with  $\mathbf{B} = [0 \dots 0 \ 1]^T$  — then from Eqs. (97.11) and (97.15) the last state equation can be written as

$$\dot{x}_{n_x} + a_{n_x} x_{n_x} + \dots + a_2 x_2 + a_1 x_1 = u \quad (97.17)$$

Since  $x_1 = y$ , it follows from the other state equations of (97.11) and (97.15),  $\dot{x}_i = x_{i+1}$  for  $i = 1, \dots, n_x - 1$ , that

$$\begin{aligned} x_2 &= \dot{x}_1 = \dot{y} \\ x_3 &= \dot{x}_2 = \ddot{y} \\ &\vdots \\ x_{n_x} &= \dot{x}_{n_x-1} = y^{(n_x-1)} \end{aligned}$$

Thus Eq. (97.17) becomes

$$y^{(n_x)} + a_{n_x} y^{(n_x-1)} + \dots + a_2 \dot{y} + a_1 y = u$$

which is in the IO format.

If the output is not just the first state component, then it may not be possible to convert an SISO system in companion form to an equivalent IO representation. In the next section we will discuss a general requirement that, when satisfied, will guarantee that a state-space SISO system (whether in companion form or not) can be converted to an equivalent IO representation.

## 97.4 Transforming System Representations

---

For the remainder of this chapter we will be concerned with transforming between various types of representations of constant-coefficient linear systems. It turns out that two fundamental concepts, controllability and observability, govern whether or not certain equivalent representations can be achieved.

### Controllability and Observability

**Controllability** is concerned with whether a control input  $u(t)$  exists that will transfer the state  $\mathbf{x}(t)$  from a given initial point  $\mathbf{x}(0)$  to a specified final state  $\mathbf{x}(t_f)$  in some finite time interval  $0 \leq t \leq t_f$ . A linear constant-coefficient system of the form in Eq. (97.9) is completely controllable (from any initial state to any final state) if and only if the Kalman controllability criterion,  $\text{rank } [\mathbf{P}] = n_x$ , is satisfied, where the  $n_x \times n_x n_u$  matrix,

$$\mathbf{P} = [\mathbf{B}, \mathbf{AB}, \mathbf{A}^2\mathbf{B}, \dots, \mathbf{A}^{n_x-1}\mathbf{B}] \quad (97.18)$$

is called the *controllability matrix*. Here,  $\text{rank } [\mathbf{P}]$  is the maximum number of linearly independent rows (or columns) in  $\mathbf{P}$  and is equal to the size of the largest nonzero square determinant obtained by deleting various rows and/or columns of  $\mathbf{P}$ . For a single-input system,  $\mathbf{P}$  is square and the controllability criterion requires  $|\mathbf{P}| \neq 0$ , where  $|\mathbf{P}|$  denotes the determinant of  $\mathbf{P}$ .

**Observability** is concerned with the problem of determining the state based on the measured outputs. In particular, a linear state-space system is observable if it is possible to uniquely determine the initial state  $\mathbf{x}(0)$  given the output and input histories  $y(t)$  and  $u(t)$  over a finite time interval  $0 \leq t \leq t_f$ . A linear constant-coefficient system of the form in Eqs. (97.9) and (97.10) is completely observable (over any nonzero time interval) if and only if the Kalman observability criterion,  $\text{rank } [\mathbf{Q}] = n_x$ , is satisfied, where the  $n_x n_y \times n_x$  matrix,

$$\mathbf{Q} = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \mathbf{CA}^2 \\ \vdots \\ \mathbf{CA}^{n_x-1} \end{bmatrix} \quad (97.19)$$

is called the *observability matrix*. For a single-output system,  $\mathbf{Q}$  is square and the observability criterion requires  $|\mathbf{Q}| \neq 0$ .

## State-Space Transformations

As indicated previously, the state-space representation of a linear system is not unique. Let  $\mathbf{M}$  be any constant nonsingular  $n_x \times n_x$  matrix with inverse  $\mathbf{M}^{-1}$ . Then the coordinate transformation  $\mathbf{z} = \mathbf{M}^{-1}\mathbf{x}$  transforms the system given in Eqs. (97.9) and (97.10) into

$$\dot{\mathbf{z}} = \hat{\mathbf{A}}\mathbf{z} + \hat{\mathbf{B}}\mathbf{u} \quad (97.20)$$

$$\mathbf{y} = \hat{\mathbf{C}}\mathbf{z} + \mathbf{D}\mathbf{u} \quad (97.21)$$

where

$$\hat{\mathbf{A}} = \mathbf{M}^{-1}\mathbf{A}\mathbf{M}, \quad \hat{\mathbf{B}} = \mathbf{M}^{-1}\mathbf{B}, \quad \hat{\mathbf{C}} = \mathbf{C}\mathbf{M} \quad (97.22)$$

### Diagonalization

For  $i = 1, \dots, n_x$  let  $\lambda_i$  denote the scalar **eigenvalues** of  $\mathbf{A}$ , with corresponding **eigenvectors**  $\xi_i \neq 0$ . That is,

$$\mathbf{A}\xi_i = \lambda_i\xi_i \quad (97.23)$$

where the eigenvalues satisfy the  $n_x$ -order polynomial characteristic equation

$$0 = |\lambda \mathbf{I} - \mathbf{A}| = \mathcal{P}(\lambda) = \lambda^{n_x} + p_{n_x-1} \lambda^{n_x-1} + \cdots + p_1 \lambda + p_0 \quad (97.24)$$

Suppose that the  $n_x$  eigenvectors are linearly independent. A sufficient condition for this, from linear algebra, is that the eigenvalues be distinct ( $\lambda_i \neq \lambda_j$  for  $i \neq j$ ). Then the eigenvector matrix (also called the *modal matrix*),

$$\mathbf{M} = [\xi_1 \cdots \xi_{n_x}] \quad (97.25)$$

whose columns are the eigenvectors of  $\mathbf{A}$ , has an inverse. Using the eigenvector matrix and the coordinate transformation  $\mathbf{z} = \mathbf{M}^{-1}\mathbf{x}$  applied to the system given in Eqs. (97.9) and (97.10) yields the transformed system of Eqs. (97.20) and (97.21), in which  $\hat{\mathbf{A}} = \mathbf{M}^{-1}\mathbf{A}\mathbf{M}$  is a diagonal matrix, with eigenvalues on the main diagonal in the same order as the eigenvector columns of  $\mathbf{M}$ . That is, the state equations become decoupled in the new state variables.

In component form the decoupled state equations are

$$\dot{z}_i = \lambda_i z_i + \hat{\mathbf{b}}_i^T \mathbf{u}, \quad i = 1, \dots, n_x \quad (97.26)$$

and the output equations can be written as

$$\mathbf{y} = z_1 \hat{\mathbf{c}}_1 + \cdots + z_{n_x} \hat{\mathbf{c}}_{n_x} + \mathbf{D}\mathbf{u} \quad (97.27)$$

where  $\hat{\mathbf{b}}_i^T$  is the  $i$ th row of  $\hat{\mathbf{B}} = \mathbf{M}^{-1}\mathbf{B}$  and  $\hat{\mathbf{c}}_i$  is the  $i$ th column of  $\hat{\mathbf{C}} = \mathbf{C}\mathbf{M}$ .

In terms of the decoupled system equations, we have direct alternate tests for controllability and observability. The system of Eqs. (97.9) and (97.10) is controllable if and only if  $\hat{\mathbf{B}}$  contains no zero row, so that  $\mathbf{u}$  affects each eigenstate  $z_i$ . Similarly, the system of Eqs. (97.9) and (97.10) is observable if and only if  $\hat{\mathbf{C}}$  contains no zero column, so that  $\mathbf{y}$  reflects each eigenstate  $z_i$ .

### Block Diagonalization

If the  $n_x$  eigenvectors of  $\mathbf{A}$  are not linearly independent, then it is not possible to transform  $\mathbf{A}$  to a diagonal matrix. That is, we cannot find a coordinate system in which each state variable is decoupled from the other state variables. However, we can always find a coordinate transformation matrix  $\mathbf{M}$  so that  $\mathbf{z} = \mathbf{M}^{-1}\mathbf{x}$  transforms  $\mathbf{A}$  to an  $n_x \times n_x$  block diagonal *Jordan matrix*,

$$\mathbf{J} = \text{diag} [\mathbf{J}_1, \mathbf{J}_2, \dots, \mathbf{J}_n] \quad (97.28)$$

where  $n$  is the number of linearly independent eigenvectors and each Jordan block  $\mathbf{J}_i$ , associated with an eigenvalue  $\lambda_i$  of  $\mathbf{A}$ , is either a  $1 \times 1$  matrix  $[\lambda_i]$  for distinct eigenvalues or, for repeated eigenvalues, a square submatrix in upper triangular form with  $\lambda_i$  on the diagonal, ones above and adjacent to the diagonal, and zeroes elsewhere:

$$\mathbf{J}_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & 1 & 0 \\ & & \lambda_i & \ddots \\ 0 & & & \ddots & 1 \\ & & & & \lambda_i \end{bmatrix} \quad (97.29)$$

The resulting transformed system, of the form

$$\dot{\mathbf{z}}_i = \mathbf{J}_i \mathbf{z}_i + \mathbf{B}_i \mathbf{u} \quad (97.30)$$

will be composed of block-decoupled subsystems, whose state variables are decoupled from those of the other subsystems.

The development of the matrix  $\mathbf{M}$ , so that  $\mathbf{z} = \mathbf{M}^{-1} \mathbf{x}$  transforms  $\mathbf{A}$  to Jordan block diagonal form, is essentially the same as in the case of complete diagonalization. The difference is that in order for the matrix  $\mathbf{M}$  to have an inverse, any linearly dependent eigenvectors of  $\mathbf{A}$  are replaced by "generalized eigenvectors" to form an  $\mathbf{M}$  matrix with  $n_x$  linearly independent columns. For a repeated eigenvalue  $\lambda$  with eigenvector  $\xi$ , a *generalized eigenvector*  $\hat{\xi}$  is a nonzero solution to the equation

$$[\lambda_i \mathbf{I} - \mathbf{A}] \hat{\xi} = -\xi \quad (97.31)$$

This equation can be used repeatedly, if necessary, with the right-hand side being either an eigenvector or a previously generated generalized eigenvector.

Let  $\lambda_1, \dots, \lambda_{n_x}$  and  $\xi_1, \dots, \xi_{n_x}$  be the eigenvalues and the associated linearly independent eigenvectors or generalized eigenvectors. Order them such that each generalized eigenvector  $\xi_i$  is generated from  $\xi_{i-1}$  by

$$[\lambda_i \mathbf{I} - \mathbf{A}] \xi_i = -\xi_{i-1} \quad (97.32)$$

whereas each eigenvector  $\xi_i$  satisfies

$$[\lambda_i \mathbf{I} - \mathbf{A}] \xi_i = 0 \quad (97.33)$$

Then the transformation  $\mathbf{z} = \mathbf{M}^{-1} \mathbf{x}$ , with  $\mathbf{M} = [\xi_1 \dots \xi_{n_x}]$ , converts  $\mathbf{A}$  to a Jordan block diagonal form  $\mathbf{J} = \mathbf{M}^{-1} \mathbf{A} \mathbf{M}$ , since  $\mathbf{M} \mathbf{J} = \mathbf{A} \mathbf{M}$ . For more details on block diagonalization see [Brogan \[1982, p. 143\]](#).

Instead of block diagonalization one can employ a matrix perturbation technique and only consider the case where the eigenvalues of  $\mathbf{A}$  are distinct, so that the eigenvectors are linearly independent and complete diagonalization is possible. This technique is based on the fact that distinct eigenvalues can always be achieved by making a small perturbation  $\varepsilon$  in the elements of  $\mathbf{A}$  that cause repeated eigenvalues [[Luenberger, 1979, p. 149](#)]. After any ensuing analysis has been

performed—for example, for the solution  $\mathbf{x}(t, \varepsilon)$  — the results can be examined in the limit as  $\varepsilon \rightarrow 0$ .

### Companion Form Systems

An SISO state-space system such as in Eqs. (97.11) and (97.12) can be transformed to a unique companion form if, and only if, the system is controllable. By way of construction, we note from Eqs. (97.11) and (97.15) that in companion form the state variables  $z_i$  all follow from  $z_1$  in a cascade:  $z_{i+1} = \dot{z}_i, i = 1, \dots, n_x - 1$ . Thus we can construct the transformation matrix by finding an  $n_x$ -dimensional vector  $\rho$  such that, by choosing  $z_1 = \rho^T \mathbf{x}$ , repeated differentiation yields a system in companion form. In particular, we choose

$$\begin{aligned}
 z_1 &= \rho^T \mathbf{x} \\
 z_2 &= \dot{z}_1 = \rho^T \mathbf{A} \mathbf{x} && \text{with } \rho^T \mathbf{B} = 0 \\
 z_3 &= \dot{z}_2 = \rho^T \mathbf{A}^2 \mathbf{x} && \text{with } \rho^T \mathbf{A} \mathbf{B} = 0 \\
 &\vdots && \\
 z_{n_x} &= \dot{z}_{n_x-1} = \rho^T \mathbf{A}^{n_x-1} \mathbf{x} && \text{with } \rho^T \mathbf{A}^{n_x-2} \mathbf{B} = 0 \\
 \dot{z}_{n_x} &= \rho^T \mathbf{A}^{n_x} \mathbf{x} + u && \text{with } \rho^T \mathbf{A}^{n_x-1} \mathbf{B} = 1
 \end{aligned} \tag{97.34}$$

From the left-hand sides of Eq. (97.34) and  $\mathbf{z} = \mathbf{M}^{-1} \mathbf{x}$ , we have

$$\mathbf{M}^{-1} = \begin{bmatrix} \rho^T \\ \rho^T \mathbf{A} \\ \rho^T \mathbf{A}^2 \\ \vdots \\ \rho^T \mathbf{A}^{n_x-1} \end{bmatrix} \tag{97.35}$$

From the right-hand sides of Eq. (97.34),  $\rho$  is the solution to

$$\rho^T [\mathbf{B}, \mathbf{A} \mathbf{B}, \mathbf{A}^2 \mathbf{B}, \dots, \mathbf{A}^{n_x-1} \mathbf{B}] = [0 \ 0 \dots 1]$$

For SISO systems, the controllability matrix  $\mathbf{P}$  is square and the controllability condition ensures the existence of  $\mathbf{P}^{-1}$ . Thus

$$\rho^T = [0 \ 0 \dots 1] \mathbf{P}^{-1} \tag{97.36}$$

Hence  $\rho^T$  is the last row of  $\mathbf{P}^{-1}$  and we construct  $\mathbf{M}^{-1}$  as shown in Eq. (97.35). The matrix  $\mathbf{M}^{-1}$  has an inverse, since the right-hand sides of Eq. (97.34) imply that the rows in Eq. (97.35) are linearly independent [Luenberger, 1979, p. 292].

## Input-Output Systems

We can transform a state-space SISO system to a unique equivalent  $n_x$ -\*order IO form if, and only if, the system is observable. To perform the transformation we differentiate the output  $y(t)$   $n_x$  times. Using the state equation (97.11) to substitute for  $\dot{\mathbf{x}}$  at each step yields the following system of equations:

$$\begin{aligned}
 y &= \mathbf{C}\mathbf{x} + D u \\
 \dot{y} &= \mathbf{C}\mathbf{A}\mathbf{x} + D\dot{u} + \mathbf{C}\mathbf{B} u \\
 \ddot{y} &= \mathbf{C}\mathbf{A}^2\mathbf{x} + D\ddot{u} + \mathbf{C}\{\mathbf{B}\dot{u} + \mathbf{A}\mathbf{B} u\} \\
 y^{(3)} &= \mathbf{C}\mathbf{A}^3\mathbf{x} + D u^{(3)} + \mathbf{C}\{\mathbf{B}\ddot{u} + \mathbf{A}\mathbf{B} \dot{u} + \mathbf{A}^2\mathbf{B} u\} \\
 &\vdots \\
 y^{(n_x-1)} &= \mathbf{C}\mathbf{A}^{n_x-1}\mathbf{x} + D u^{(n_x-1)} + \mathbf{C}\{\mathbf{B} u^{(n_x-2)} + \mathbf{A}\mathbf{B} u^{(n_x-3)} + \dots + \mathbf{A}^{n_x-2}\mathbf{B} u\} \\
 y^{(n_x)} &= \mathbf{C}\mathbf{A}^{n_x}\mathbf{x} + D u^{(n_x)} + \mathbf{C}\{\mathbf{B} u^{(n_x-1)} + \mathbf{A}\mathbf{B} u^{(n_x-2)} + \dots + \mathbf{A}^{n_x-1}\mathbf{B} u\}
 \end{aligned} \tag{97.37}$$

The first  $n_x$  of these equations can be solved for  $\mathbf{x}$  in terms of  $y, u$ , and their derivatives. For SISO systems the observability matrix  $\mathbf{Q}$  is square and the observability condition ensures the existence of an inverse. Thus  $\mathbf{x}$  will be a unique function of  $y, u$ , and their derivatives. Substituting this result into the  $y^{(n_x)}$  equation in (97.37) and collecting terms yields an IO system of the form in Eq. (97.16).

## State Equations from the Transfer Matrix

For a multiple-input, multiple-output state-space system of the form in Eqs. (97.9) and (97.10), taking the Laplace transform of both equations and setting the initial condition terms to zero yields

$$\mathbf{Y}(s) = \mathbf{G}(s)\mathbf{U}(s) \tag{97.38}$$

where  $\mathbf{Y}(s) = \mathcal{L}[\mathbf{y}(t)]$  and  $\mathbf{U}(s) = \mathcal{L}[\mathbf{u}(t)]$  are the Laplace transforms of  $\mathbf{y}(t)$  and  $\mathbf{u}(t)$ , respectively, and

$$\mathbf{G}(s) = \mathbf{C}[s\mathbf{I} - \mathbf{A}]^{-1}\mathbf{B} + \mathbf{D} \tag{97.39}$$

is the  $n_y \times n_u$  transfer matrix. Suppose that the transfer matrix is known, perhaps from experimental results, and is given by

$$\mathbf{G}(s) = \frac{1}{\mathcal{P}(s)} \begin{bmatrix} \mathcal{Q}_{11}(s) & \cdots & \mathcal{Q}_{1n_u}(s) \\ \vdots & \ddots & \vdots \\ \mathcal{Q}_{n_y 1}(s) & \cdots & \mathcal{Q}_{n_y n_u}(s) \end{bmatrix} \quad (97.40)$$

where each  $\mathcal{Q}_{ij}(s)$  is a polynomial of order less than or equal to the order  $n_x$  of the characteristic polynomial  $\mathcal{P}(s) = |s\mathbf{I} - \mathbf{A}|$ . We wish to construct a state-space model from the transfer matrix model.

We will first obtain a decoupled model in which the  $\mathbf{A}$  matrix is diagonal. Then a suitable coordinate transformation is applied to yield a state-space representation having specified properties, such as a desired set of eigenvectors or an  $\mathbf{A}$  matrix that is in companion form.

We compute the eigenvalues from the  $n_x$ -order characteristic equation  $\mathcal{P}(\lambda) = 0$ . Assuming that these eigenvalues are distinct, we define the diagonalized state matrix as  $\hat{\mathbf{A}} = \text{diag} [\lambda_1, \dots, \lambda_{n_x}]$ . The next step is to determine an  $n_x \times n_u$  matrix  $\hat{\mathbf{B}}$ , an  $n_y \times n_x$  matrix  $\hat{\mathbf{C}}$ , and an  $n_y \times n_u$  matrix  $\mathbf{D}$  such that the following conditions hold:

1. The transfer matrix corresponds to a state-space system, that is,

$$\mathbf{G}(s) = \hat{\mathbf{C}} \left[ s\mathbf{I} - \hat{\mathbf{A}} \right]^{-1} \hat{\mathbf{B}} + \mathbf{D}.$$

2.  $\hat{\mathbf{B}}$  has no zero rows (controllability is satisfied).
3.  $\hat{\mathbf{C}}$  has no zero columns (observability is satisfied).

Usually, the system of equations that result from equating like powers of  $s$  in condition 1 involves fewer equations than unknowns in  $\hat{\mathbf{B}}$ ,  $\hat{\mathbf{C}}$ , and  $\mathbf{D}$ , so there is some degree of freedom in choosing the elements of  $\hat{\mathbf{B}}$ ,  $\hat{\mathbf{C}}$ , and  $\mathbf{D}$ . The resulting decoupled state-space system is

$$\begin{aligned} \dot{\mathbf{z}} &= \hat{\mathbf{A}}\mathbf{z} + \hat{\mathbf{B}}\mathbf{u} \\ \mathbf{y} &= \hat{\mathbf{C}}\mathbf{z} + \mathbf{D}\mathbf{u} \end{aligned} \quad (97.41)$$

From the diagonalized representation, which may have complex-valued matrices, we can change to a final set of state variables  $\mathbf{x} = \mathbf{M}\mathbf{z}$  by choosing a set of desired linearly independent eigenvectors as the columns of the transformation matrix  $\mathbf{M} = [\xi_1 \dots \xi_{n_x}]$ . This transformation yields

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \\ \mathbf{y} &= \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u} \end{aligned} \quad (97.42)$$

where  $\mathbf{A} = \mathbf{M}\hat{\mathbf{A}}\mathbf{M}^{-1}$ ,  $\mathbf{B} = \mathbf{M}\hat{\mathbf{B}}$ , and  $\mathbf{C} = \hat{\mathbf{C}}\mathbf{M}^{-1}$ . The matrix  $\mathbf{A}$  will have eigenvalues  $\lambda_i$  and the chosen eigenvectors. For real-valued final matrices, complex conjugate eigenvectors should be chosen for any corresponding conjugate eigenvalues.

As an example, consider the IO system,

$$\ddot{y} + 3\dot{y} + 2y = u_1 + 3u_2 + \dot{u}_2$$

that has the transfer matrix representation

$$Y(s) = \begin{bmatrix} \frac{1}{s^2 + 3s + 2} & \frac{s + 3}{s^2 + 3s + 2} \end{bmatrix} \begin{bmatrix} U_1(s) \\ U_2(s) \end{bmatrix} = \mathbf{G}(s)\mathbf{U}(s)$$

The characteristic equation  $\lambda^2 + 3\lambda + 2 = 0$  yields the eigenvalues  $\lambda_1 = -1, \lambda_2 = -2$ . Thus the diagonalized state equations will have the  $\hat{\mathbf{A}}$  matrix given by

$$\hat{\mathbf{A}} = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}$$

The condition  $\mathbf{G}(s) = \hat{\mathbf{C}} [s\mathbf{I} - \hat{\mathbf{A}}]^{-1} \hat{\mathbf{B}} + \mathbf{D}$  yields

$$\frac{\begin{bmatrix} 1 & s + 3 \end{bmatrix}}{s^2 + 3s + 2} = \begin{bmatrix} \hat{c}_1 & \hat{c}_2 \end{bmatrix} \begin{bmatrix} \frac{1}{s + 1} & 0 \\ 0 & \frac{1}{s + 2} \end{bmatrix} \begin{bmatrix} \hat{b}_{11} & \hat{b}_{12} \\ \hat{b}_{21} & \hat{b}_{22} \end{bmatrix} + \begin{bmatrix} d_1 & d_2 \end{bmatrix}$$

Thus the elements of  $\hat{\mathbf{B}}, \hat{\mathbf{C}},$  and  $\mathbf{D}$  must satisfy

$$\begin{aligned} 1 &= d_1 s^2 + (\hat{c}_1 \hat{b}_{11} + \hat{c}_2 \hat{b}_{21} + 3d_1) s + (2\hat{c}_1 \hat{b}_{11} + \hat{c}_2 \hat{b}_{21} + 2d_1) \\ s + 3 &= d_2 s^2 + (\hat{c}_1 \hat{b}_{12} + \hat{c}_2 \hat{b}_{22} + 3d_2) s + (2\hat{c}_1 \hat{b}_{12} + \hat{c}_2 \hat{b}_{22} + 2d_2) \end{aligned}$$

Equating like powers of  $s$  and solving the resulting equations yields

$$d_1 = d_2 = 0, \quad \hat{c}_1 \hat{b}_{11} = 1, \quad \hat{c}_2 \hat{b}_{21} = -1, \quad \hat{c}_1 \hat{b}_{12} = 2, \quad \hat{c}_2 \hat{b}_{22} = -1$$

The observability condition in 3 requires that  $\hat{c}_1 \neq 0$  and  $\hat{c}_2 \neq 0$ . Therefore, these results can be solved for the  $\hat{b}_{ij}$  in terms of  $\hat{c}_1$  and  $\hat{c}_2$ . The resulting  $\hat{b}_{ij}$  are all nonzero, so the controllability condition in 2 is also satisfied. For this example, we choose  $\hat{c}_1 = \hat{c}_2 = 1$ , yielding the decoupled system of Eq. (97.41) with

$$\hat{\mathbf{A}} = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}, \quad \hat{\mathbf{B}} = \begin{bmatrix} 1 & 2 \\ -1 & -1 \end{bmatrix}, \quad \hat{\mathbf{C}} = \begin{bmatrix} 1 & 1 \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} 0 & 0 \end{bmatrix}$$

Since the eigenvalues were all real, the diagonalized state-space model is real, so further transformation is not required. However, to complete the example, suppose we want to change from this diagonal form to a state-space system having a specified set of eigenvectors, such as  $\xi_1 = [1 \quad -1]^T$  and  $\xi_2 = [1 \quad -2]^T$ . As a final step, we employ a coordinate transformation  $\mathbf{x} = \mathbf{M}\mathbf{z}$ , where  $\mathbf{M} = [\xi_1 \quad \xi_2]$ . This yields a state-space system as in Eq. (97.42), with



$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{C} = [1 \quad 0], \quad \mathbf{D} = [0 \quad 0]$$

Note that our choice of eigenvectors happened to produce an  $\mathbf{A}$  matrix in companion form, with the same eigenvalues as  $\hat{\mathbf{A}}$ , but with different eigenvectors. If our objective had been to directly produce an  $\mathbf{A}$  matrix in companion form, we could have chosen  $\xi_1 = [1 \quad \alpha]^T$  and  $\xi_2 = [1 \quad \beta]^T$  as variables, computed  $\mathbf{A} = \mathbf{M}\hat{\mathbf{A}}\mathbf{M}^{-1} = [a_{ij}]$ , and then solved for the parameters  $\alpha$  and  $\beta$  to satisfy the two companion matrix conditions  $a_{11} = 0$  and  $a_{12} = 1$ .

## Defining Terms

**Cause and effect:** In this context for the IO formulation, a condition that requires that the order of any derivative in the input be less than or equal to the highest order of derivative in the output.

**Controllability:** The ability to drive a system from an arbitrary initial state to an arbitrary final state in finite time.

**Eigenvalue:** Any scalar  $\lambda$  satisfying the equation  $\det[\lambda\mathbf{I} - \mathbf{A}] = 0$ , where  $\mathbf{A}$  is an  $n_x \times n_x$  matrix and  $\mathbf{I}$  is the identity matrix; alternately, a scalar  $\lambda$  satisfying the equation  $\mathbf{A}\xi = \lambda\xi$ , where  $\xi$  is an  $n_x$ -dimensional eigenvector.

**Eigenvector:** Any nonzero  $n_x$ -dimensional vector  $\xi$  satisfying the equation  $\mathbf{A}\xi = \lambda\xi$ , where  $\mathbf{A}$  is an  $n_x \times n_x$  dimensional matrix and  $\lambda$  is an eigenvalue.

**Input-output (IO) model:** A representation of a dynamic system in terms of a single  $n_x$ -order differential equation relating the output  $y(t)$  to the input  $u(t)$ .

**Input vector:** An  $n_u$ -dimensional column vector consisting of the variable quantities, other than state variables, that affect the evolution of the state of a dynamic system.

**Linear output equations:** A set of algebraic equations defining the output variables in terms of linear combinations of the state and input variables.

**Linear state equations:** A set of first-order differential equations that model the behavior of a physical system in terms of a linear combination of the state and input variables.

**Multiple-input, multiple-output (MIMO) model:** A state-space model with an input vector,  $\mathbf{u}$ , of dimension  $n_u$  greater than one and output vector,  $\mathbf{y}$ , of dimension  $n_y$  greater than one.

**Observability:** The ability to determine a system's initial state given the output and input histories over a finite time interval.

**Output equations:** A set of algebraic equations defining the output variables as functions of the state variables and the input variables.

**Output vector:** An  $n_y$ -dimensional column vector whose elements model the measurements of a physical system.

**Single-input, single-output (SISO) model:** A state-space model with a single scalar input,  $u$ , and single scalar output,  $y$ .

**State equations:** A set of first-order differential equations that model the behavior of a physical system.

**State space:** A geometric space with dimension  $n_x$  equal to the number of state variables. Any possible state of a dynamic model can be represented as a point in state space.

**State variables:** The smallest set of time-differentiated variables whose initial conditions, along with the inputs and the model, allow complete prediction of the behavior of a dynamic system. It is possible to define more than one set of state variables for any particular model. However, the number of state variables is a unique quantity for a system.

**State vector:** An  $n_x$ -dimensional column vector consisting of the state variables of a model.

## References

Brogan, W. L. 1982. *Modern Control Theory*. Prentice Hall, Englewood Cliffs, NJ.

Luenberger, D. G. 1979. *Introduction to Dynamic Systems*. John Wiley & Sons, New York.

## Further Information

The following two texts present examples and in-depth information on state-space models:

Friedland, B. 1986. *Control System Design: An Introduction to State-Space Methods*.

McGraw-Hill, New York.

Grantham, W. J. and Vincent, T. L. 1993. *Modern Control Systems Analysis and Design*. John Wiley & Sons, New York.

Neudorfer, P., Gehlen, P. "Frequency Response"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

### 98.1 Frequency Response Plotting

Linear Plots • Bode Diagrams

### 98.2 A Comparison of Methods

**Paul Neudorfer**

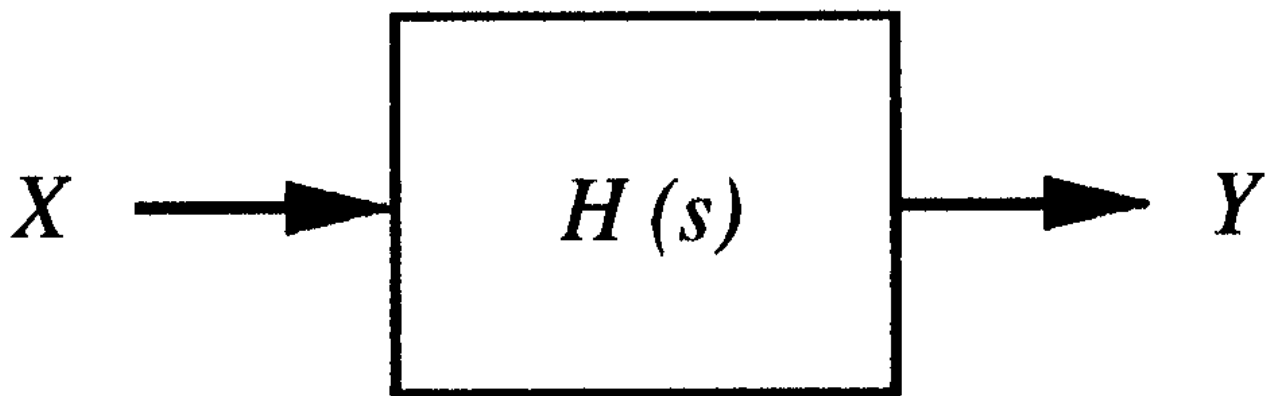
*Seattle University*

**Pierre Gehlen**

*Seattle University*

**Frequency response** in stable, linear systems is defined as "the frequency-dependent relation in both gain and phase difference between steady-state sinusoidal inputs and the resultant steady-state sinusoidal outputs" [IEEE, 1988]. The frequency response characteristics of a system can be found analytically from its transfer function. They are also commonly generated from laboratory or field tests. A single-input/single-output linear time-invariant system is shown in Fig. 98.1.

**Figure 98.1** A single-input/single-output linear system.



For systems with no time delay, the transfer function  $H(s)$  is in the form of a ratio of polynomials in the complex frequency  $s$ ,

$$H(s) = K \frac{N(s)}{D(s)}$$

where  $K$  is a frequency-independent constant. For a system in the sinusoidal steady state,  $s$  is replaced by the sinusoidal frequency  $j\omega$  ( $j = \sqrt{-1}$ ) and the system function becomes

$$H(j\omega) = K \frac{N(j\omega)}{D(j\omega)} = |H(j\omega)| e^{j \arg H(j\omega)}$$

$H(j\omega)$  is a complex quantity. Its magnitude,  $|H(j\omega)|$ , and its argument or angle,  $\arg H(j\omega)$ , relate, respectively, the amplitudes and phase angles of sinusoidal steady state input and output signals. Referring to [Fig. 98.1](#), if the input and output signals are

$$x(t) = X \cos(\omega t + \theta_x)$$

$$y(t) = Y \cos(\omega t + \theta_y)$$

then the output's amplitude  $Y$  and phase angle  $\theta_y$  are related to those of the input by the two equations

$$Y = |H(j\omega)| X$$

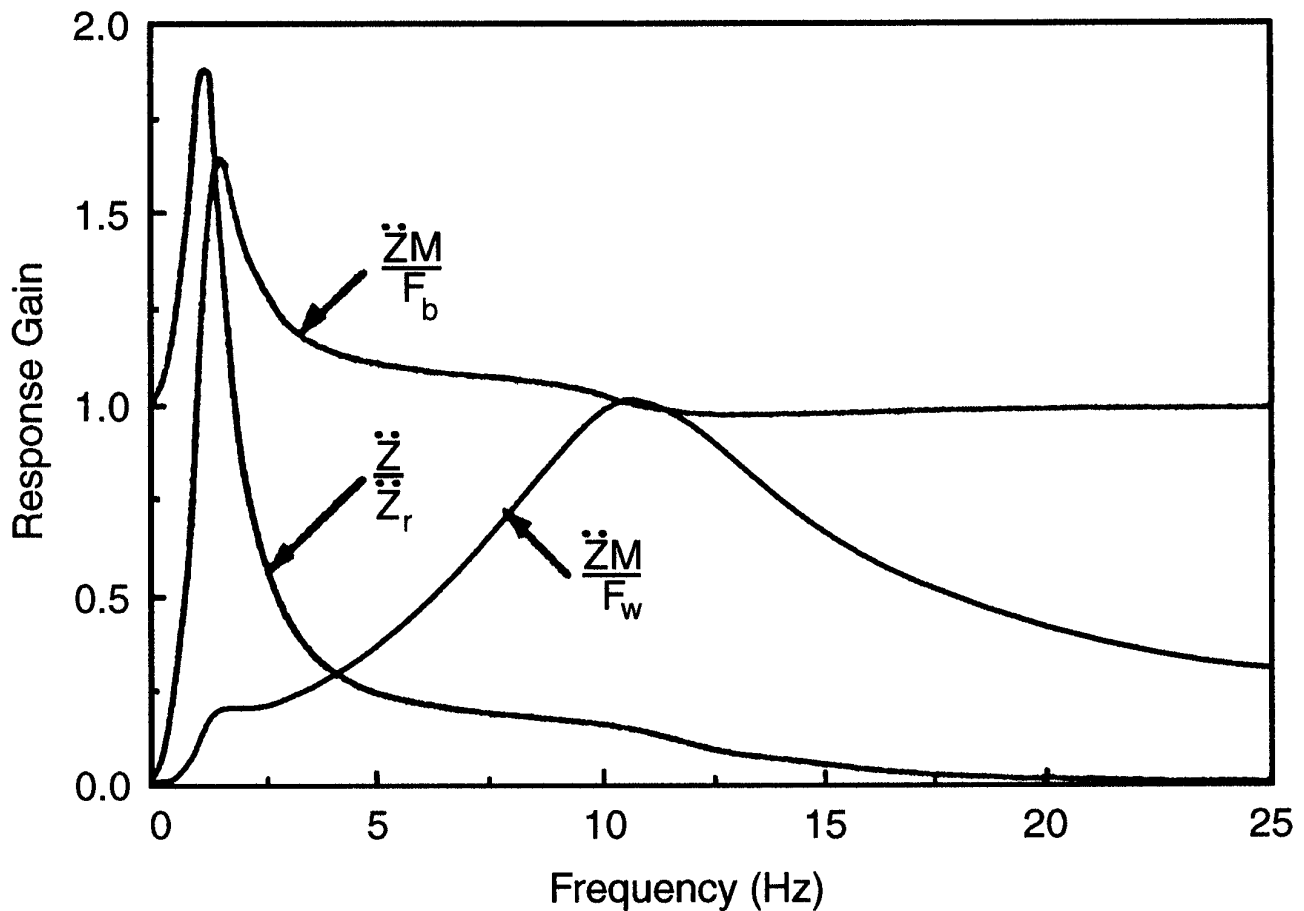
$$\theta_y = \arg H(j\omega) + \theta_x$$

The term *frequency response characteristic* usually implies a complete description of a system's sinusoidal steady state behavior as a function of frequency. Because  $H(j\omega)$  is complex, frequency response characteristics cannot be graphically displayed as a single curve plotted with respect to frequency. Instead, the magnitude and argument of  $H(j\omega)$  can be separately plotted as functions of frequency. It is often advantageous to plot frequency response curves on other than linearly scaled Cartesian coordinates. **Bode diagrams** (developed in the 1930s by H. W. Bode of Bell Labs) use a logarithmic scale for frequency and a decibel measure for magnitude. In **Nyquist plots** (from Harry Nyquist, also of Bell Labs),  $H(j\omega)$  is displayed in Argand diagram form on the complex plane— $\text{Re}[H(j\omega)]$  being on the horizontal axis and  $\text{Im}[H(j\omega)]$  on the vertical. Frequency is a parameter of such curves. It is sometimes numerically identified at selected points of the curve and sometimes omitted. The **Nichols chart** (developed by N. B. Nichols) graphs magnitude versus phase for the system function—frequency again being a parameter of the curve.

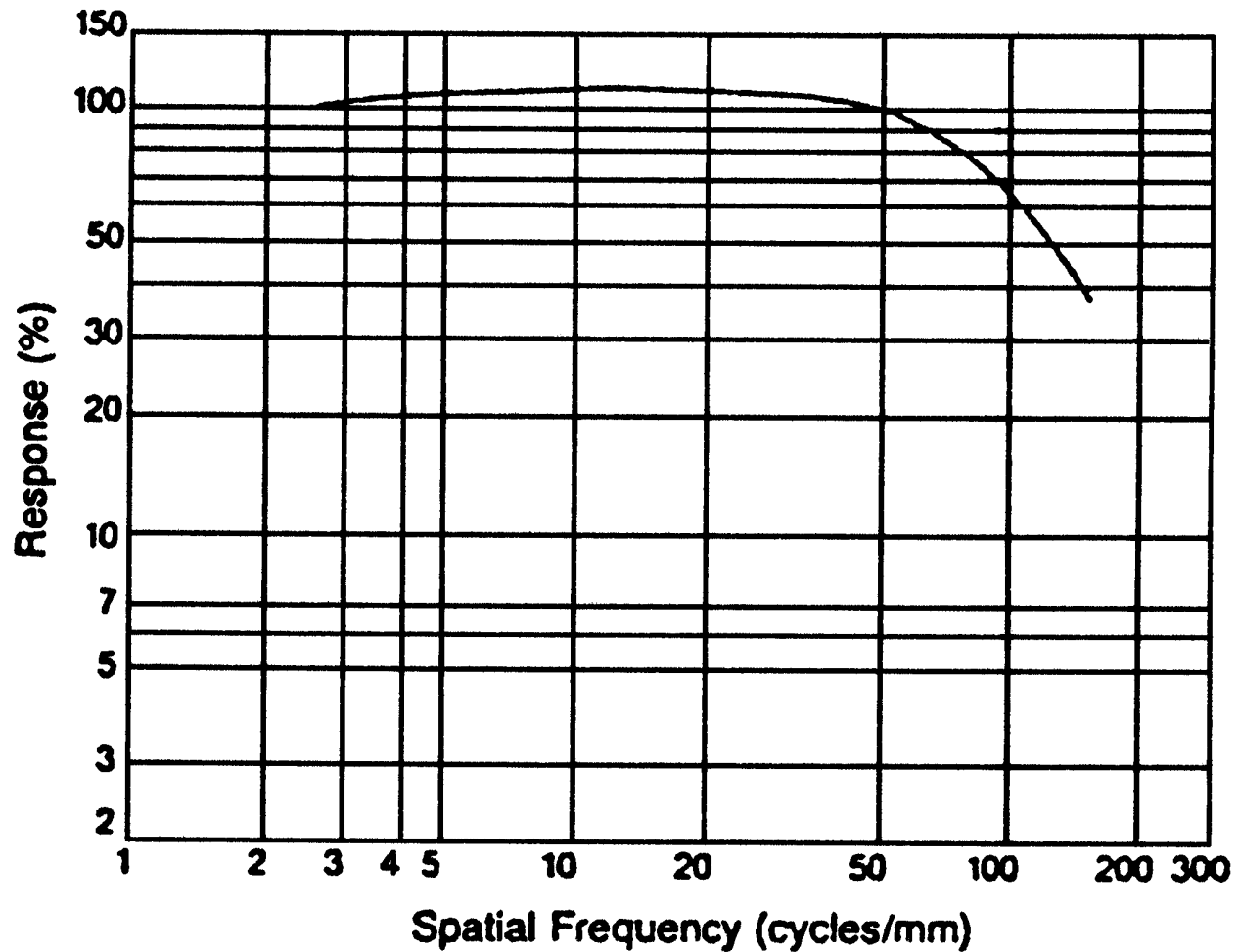
Frequency response techniques are most obviously applicable to topics such as communications and electrical filters, in which the frequency response behaviors of systems are central to an understanding of their operations. It is, however, in the area of control systems that frequency response techniques are most fully developed as analytical and design tools. The Nichols chart, for instance, is used exclusively in the analysis and design of classical feedback control systems. Although frequency response concepts are most often associated with electrical engineering, they are also commonly employed in other branches of engineering and the natural sciences. Two

nonelectrical examples are given in Figs. 98.2 and 98.3, which show, respectively, a frequency response test of a scale model of the suspension system of a truck and the modulation transfer function (MTF) of a photographic film. MTFs provide an accurate description of the quality of photographic systems and have largely replaced more subjective parameters, such as resolving power, sharpness, and acutance. Note that frequency response concepts are equally applicable to problems involving temporal (Fig. 98.2) and spatial (Fig. 98.3) frequencies.

**Figure 98.2** Frequency response test of a truck cab and suspension. (Source: Gillespie, T. B. 1992. *Fundamentals of Vehicle Dynamics*. Society of Automotive Engineers, Warrendale, PA. Courtesy of the Society of Automotive Engineers, Inc.)



**Figure 98.3** Modulation transfer function of photographic film. (Source: Williams, J. B. 1991. *Image Clarity: High Resolution Photography*. Focal, Boston, MA. Courtesy of Eastman Kodak.)



## 98.1 Frequency Response Plotting

Frequency response plots are prepared by computing the magnitude and argument of  $H(j\omega)$ .

### Linear Plots

In linear plots  $|H(j\omega)|$  and  $\arg H(j\omega)$  are shown in separate diagrams as functions of frequency (either  $f$  or  $\omega$ ). Cartesian coordinates are used and all scales are linear.

**Example 98.1.** Consider the transfer function

$$H(s) = \frac{160\,000}{s^2 + 220s + 160\,000}$$

The complex frequency variable  $s$  is replaced by the sinusoidal frequency  $j\omega$  and the magnitude

and argument found.

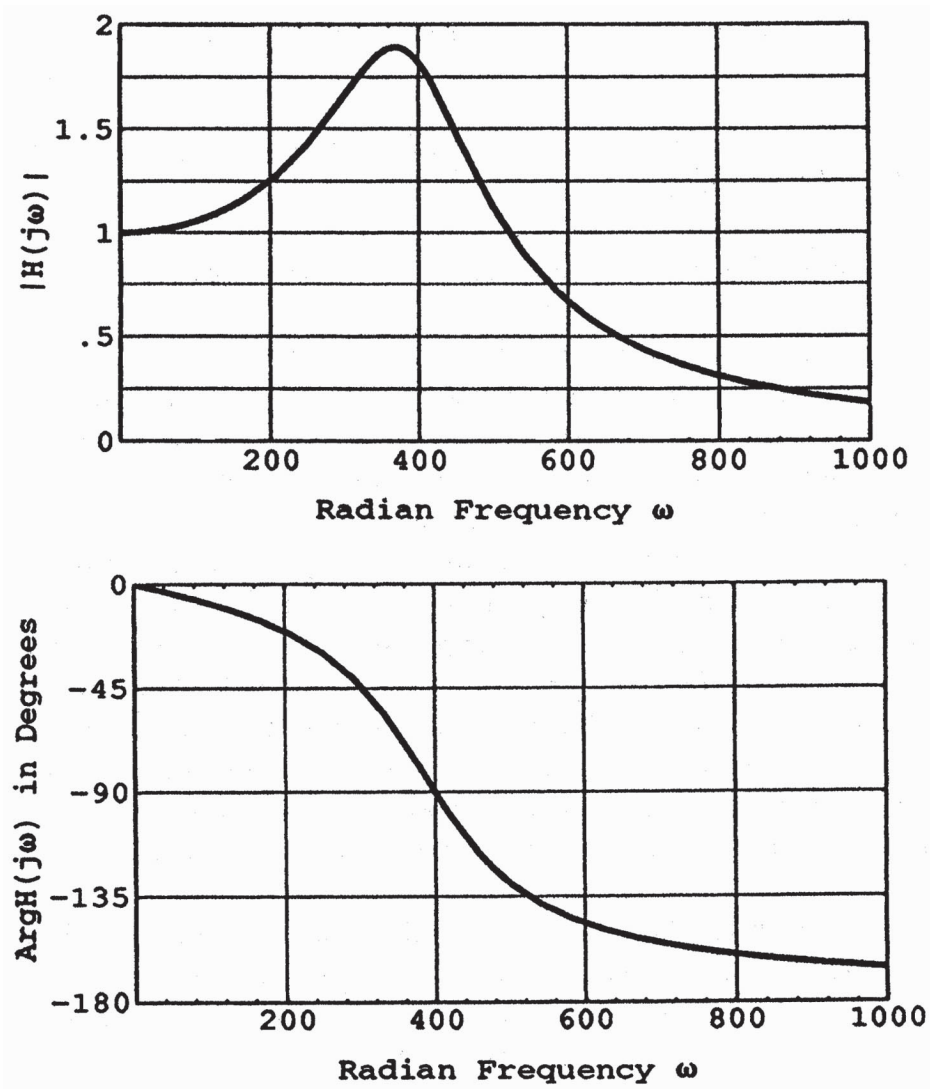
$$H(j\omega) = \frac{160\,000}{(j\omega)^2 + 220(j\omega) + 160\,000}$$

$$|H(j\omega)| = \frac{160\,000}{\sqrt{(160\,000 - \omega^2)^2 + (220\omega)^2}}$$

$$\arg H(j\omega) = -\tan^{-1} \frac{220\omega}{160\,000 - \omega^2}$$

The plots of magnitude and argument are shown in Fig. 98.4. Linear plots are most useful when the frequency range of interest is small, as is commonly the case for mechanical systems. Such plots give a straightforward representation of system response.

**Figure 98.4** Rectangular frequency response curves of  $H(j\omega)$ . (Source: Dorf, R. C. (Ed.) 1993. *The Electrical Engineering Handbook*, Chap. 11. CRC, Boca Raton, FL.)





The peaking of the magnitude function near  $\omega = 400$  rad/s in Fig. 98.4 is a reflection of the phenomenon of **resonance**. Resonance may exist in lightly damped second or higher order systems. Examples of resonance are common in the natural and human-made worlds. Organ pipes, for example, are designed to resonate at desired pitches. Resonances can have undesirable effects such as when the steering mechanism of a vehicle with misaligned front wheels shimmies at certain operating speeds. Two well-known examples of the destructive effects of resonance are the Tacoma Narrows suspension bridge, which underwent catastrophic failure during a moderate windstorm, and the collapse of an elevated walkway in the lobby of a Kansas City hotel under the load of a crowd of dancers. All mechanical systems resonate at certain critical frequencies. The careful design engineer makes sure that vibrations that might affect the system's structural integrity are well outside the service range.

## Bode Diagrams

A Bode diagram consists of plots of the gain and phase of a transfer function, each with respect to logarithmically scaled frequency axes. In addition, the gain of the transfer function is scaled in **decibels** according to the definition

$$H_{\text{dB}} = 20 \log_{10} |H(j\omega)|$$

Bode diagrams have the advantage of clearly identifying system features even if they occur over wide ranges of frequency or dynamic response. Before constructing a Bode diagram, the transfer function is normalized so that each pole or zero term (except those at  $s = 0$ ) has a DC gain of one. For instance,

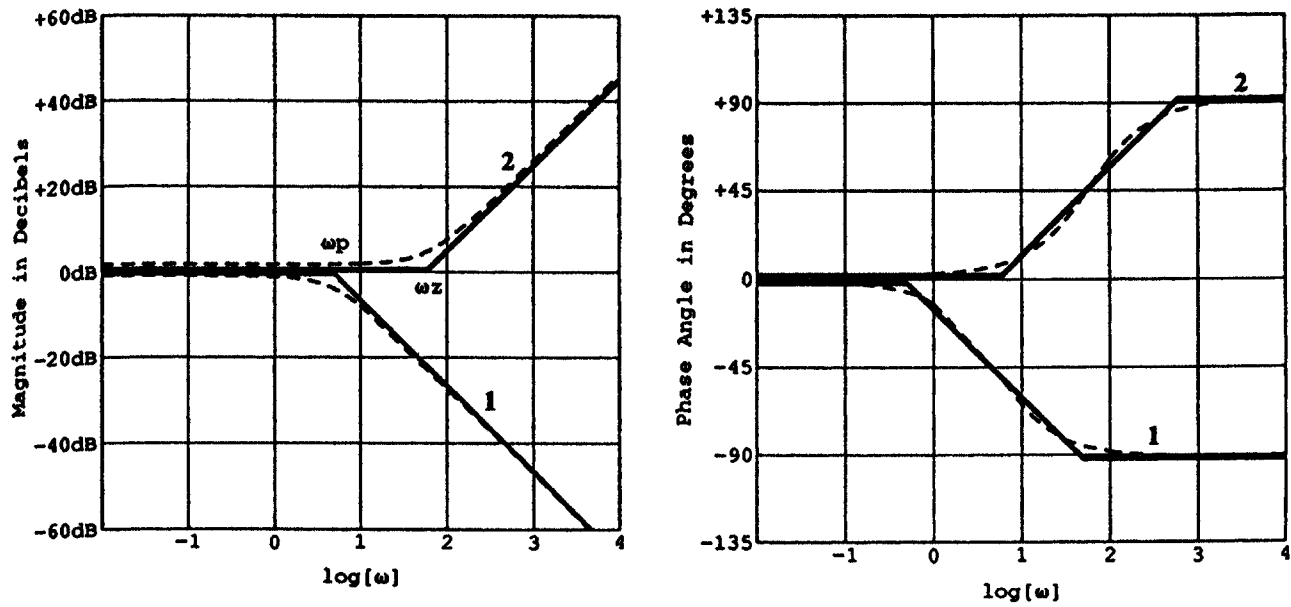
$$H(s) = K \frac{s + \omega_z}{s(s + \omega_p)} = \frac{K\omega_z}{\omega_p} \frac{s/\omega_z + 1}{s(s/\omega_p + 1)} = K' \frac{s\tau_z + 1}{s(s\tau_p + 1)}$$

It is common to draw Bode diagrams directly from  $H(s)$  without making the formal substitution  $s = j\omega$ .

When drawn by hand, Bode magnitude and phase curves are developed by adding the individual contributions of the factored terms of the transfer function's numerator and denominator. In general, these factored terms may include (1) a constant  $K$ , (2) a simple  $s$  term corresponding to either a zero or a pole at the origin, (3) a term such as  $(s\tau + 1)$  corresponding to a real valued (nonzero) pole or zero, and (4) a quadratic term with a possible standard form of  $[(s/\omega_n)^2 + 2\zeta(s/\omega_n) + 1]$  corresponding to a pair of complex conjugate poles or zeros and for which  $0 < \zeta < 1$ . With the exception of quadratic terms having small  $\zeta$  (**damping ratio**), the Bode magnitude and phase curves for all such expressions can be reasonably approximated by a series of straight-line segments. Detailed procedures for drawing Bode diagrams from these basic forms are described in many references. Examples are given in Fig. 98.5, which shows straight-line approximations for both a numerator-factored term  $(s/\omega_z + 1)$  and a denominator-factored term,  $1/(s/\omega_p + 1)$ . The approximations are shown as heavy straight lines. The exact curves are shown

as dashed lines and have actually been displaced somewhat so as not to be obscured by the approximations.

**Figure 98.5** Bode curves for (1) a pole at  $s = -\omega_p$  and (2) a zero at  $s = -\omega_z$ . (Source: Dorf, R. C. (Ed.). 1993. *The Electrical Engineering Handbook*, Chap. 11. CRC, Boca Raton, FL.)



Note in Fig. 98.5 that both decibel magnitude and phase are plotted semilogarithmically. The frequency axis is logarithmically scaled so that every tenfold, or **decade**, change in frequency occurs over an equal distance. The magnitude axis is given in decibels. Positive decibel magnitudes correspond to amplifications between input and output that are greater than one (output amplitude larger than input). Negative decibel gains correspond to attenuation between input and output (output amplitude smaller than input). The straight-line approximations differ most greatly from the exact curves at points where the approximation changes slope. In magnitude curves these are called **breakpoints**. At these points the approximate and exact curves differ by 3 dB for each real pole or zero.

Bode diagrams are easily constructed because, with the exception of lightly damped quadratic terms, each contribution can be reasonably approximated with straight lines. Also, the overall frequency response curve is found by adding the individual contributions. Two examples follow.

### Example 98.2

$$A(s) = \frac{10^4 s}{s^2 + 1100s + 10^5} = \frac{10^4 s}{(s + 100)(s + 1000)} = 10^{-1} \frac{s}{(s/100 + 1)(s/1000 + 1)}$$

In Fig. 98.6 the individual contributions of the four factored terms of  $H(s)$  are shown as long-dashed lines. The straight-line approximations for gain and phase are shown with solid lines.

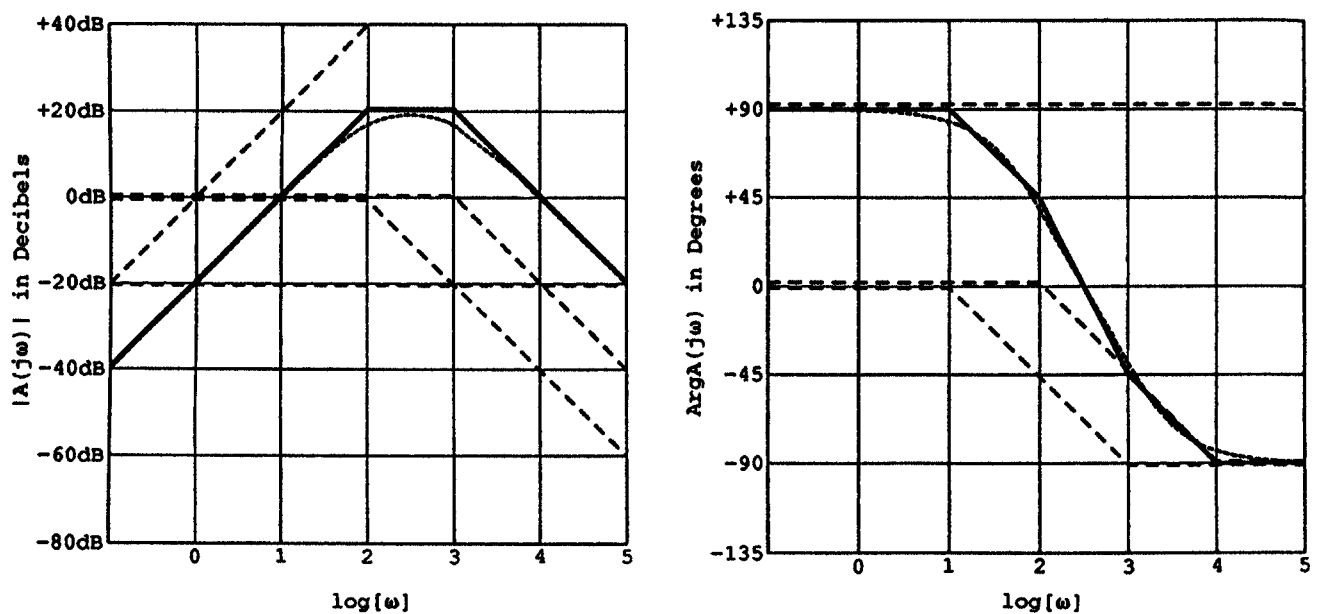
The exact curves are presented with short-dashed lines.

### Example 98.3

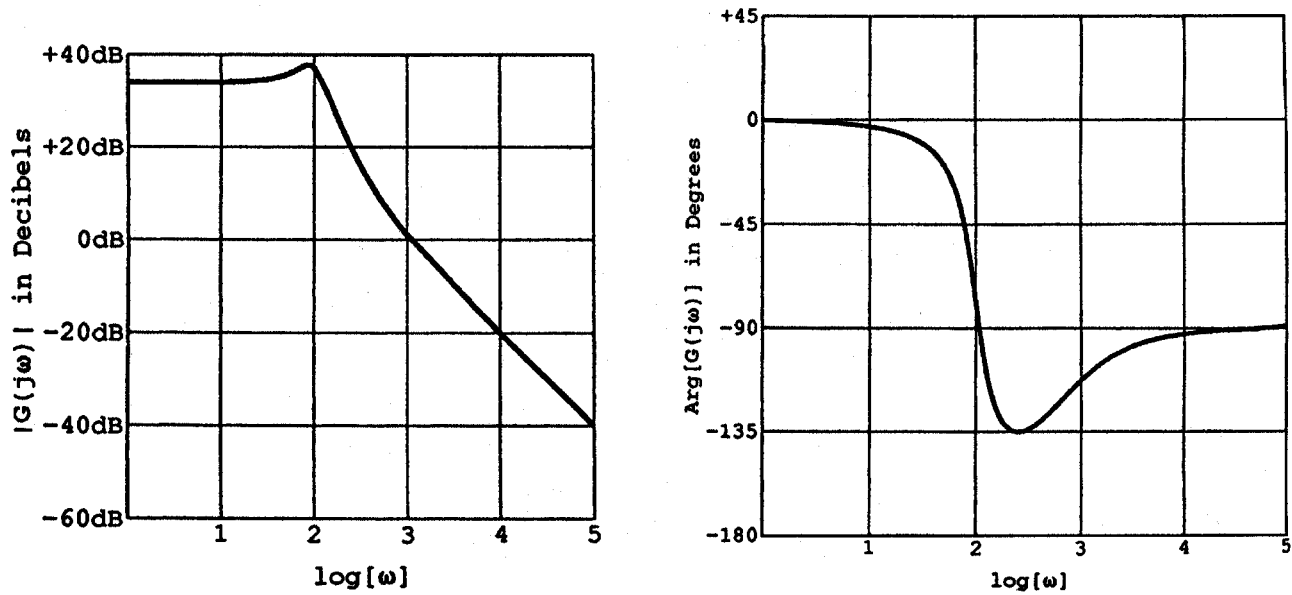
$$G(s) = \frac{1000(s + 500)}{s^2 + 70s + 10\,000} = \frac{50(s/500 + 1)}{(s/100)^2 + 2(.35)(s/100) + 1}$$

Note that, for the quadratic term in the denominator, the damping ratio is .35, an indication of resonance. For small damping ratios the straight-line approximations of Bode magnitude and phase plots vary significantly from the exact curves. For improved accuracy the approximations would have to be adjusted near the frequency of  $\omega = 100$  rad/s. This is not a consideration when a computer is used to generate a Bode diagram. Figure 98.7 shows the exact gain and phase frequency response curves for  $G(s)$ .

**Figure 98.6** Bode diagram of  $A(s)$ . (Source: Dorf, R. C. (Ed.). 1993. *The Electrical Engineering Handbook*, Chap. 11. CRC, Boca Raton, FL.)



**Figure 98.7** Bode diagram of  $G(s)$ . (Source: Dorf, R. C. (Ed.). 1993. *The Electrical Engineering Handbook*, Chap. 11. CRC, Boca Raton, FL.)



## 98.2 A Comparison of Methods

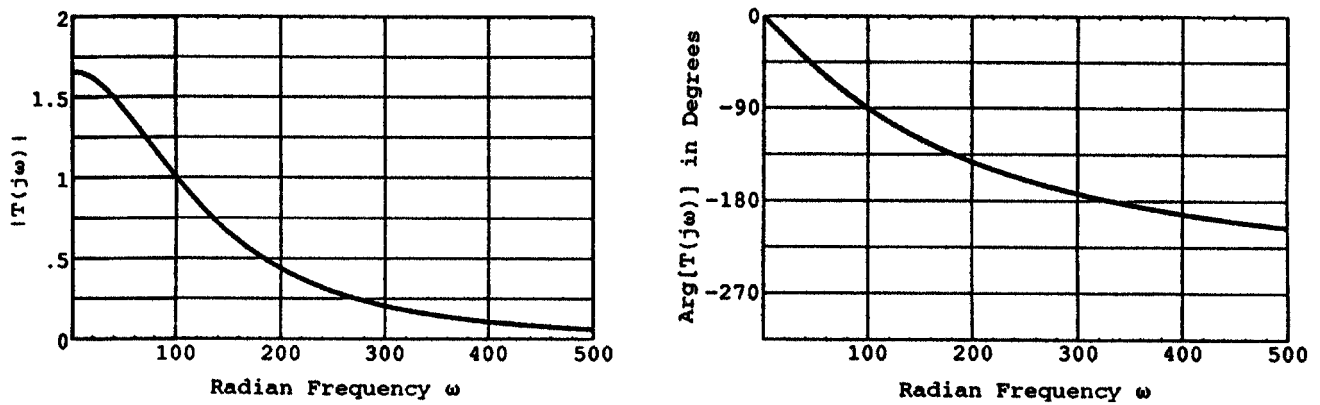
This chapter concludes with the frequency response of a simple system function plotted in three different ways.

### Example 98.4

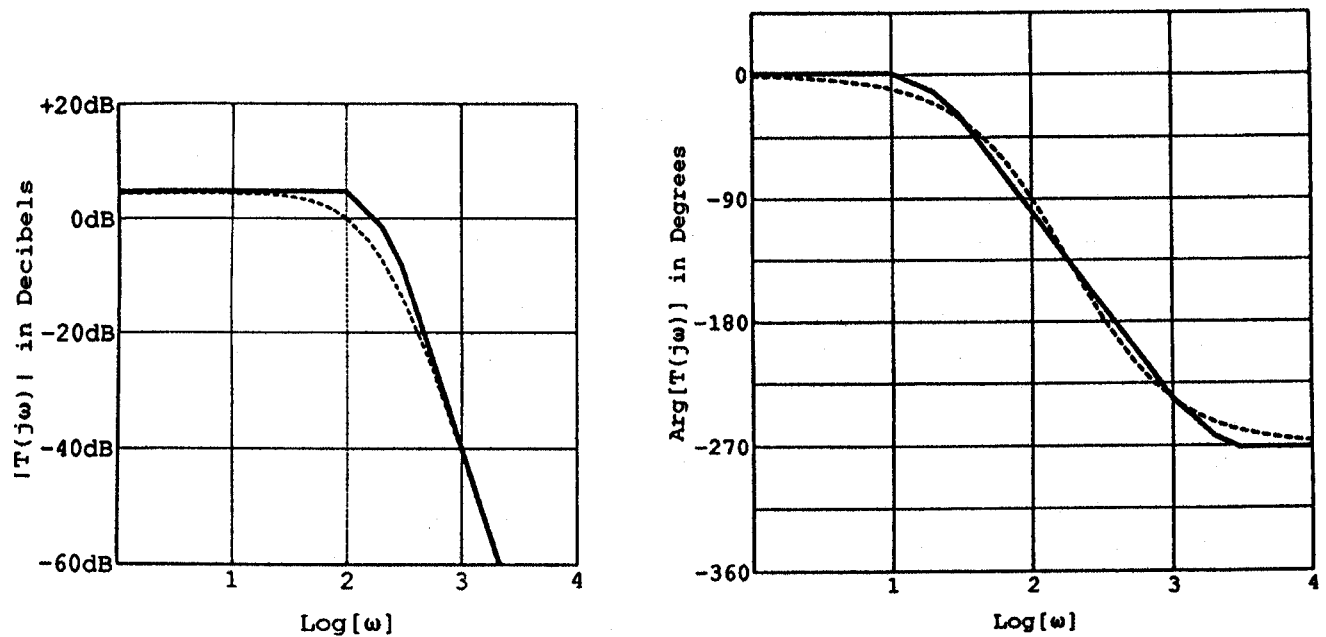
$$T(s) = \frac{10^7}{(s + 100)(s + 200)(s + 300)}$$

Figure 98.8 shows linear frequency response curves for  $H(s)$ . Corresponding Bode and Nyquist diagrams are shown, respectively, in Figs. 98.9 and 98.10.

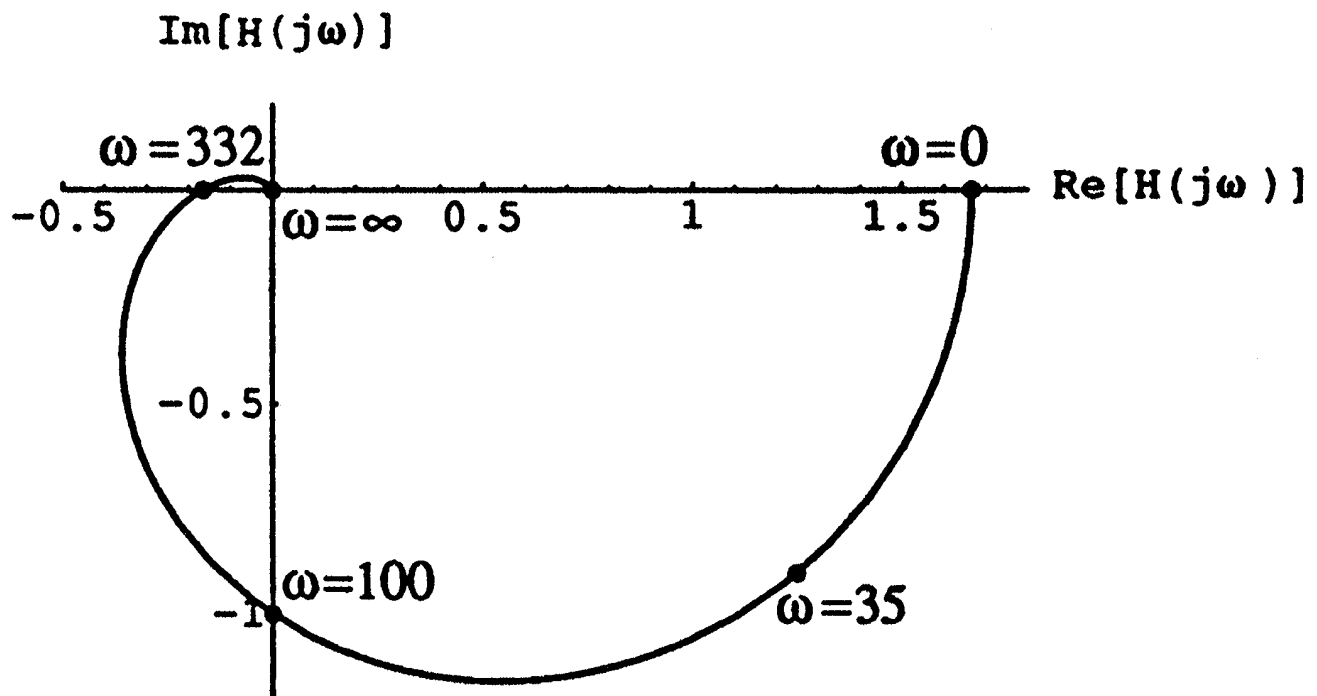
**Figure 98.8** Linear frequency response plot of  $T(s)$ . (Source: Dorf, R. C. (Ed.). 1993. *The Electrical Engineering Handbook*, Chap. 11. CRC, Boca Raton, FL.)



**Figure 98.9** Bode diagram of  $T(s)$ . (Source: Dorf, R. C. (Ed.). 1993. *The Electrical Engineering Handbook*, Chap. 11. CRC, Boca Raton, FL.)



**Figure 98.10** Nyquist plot of  $T(s)$ . (Source: Dorf, R. C. (Ed.). 1993. *The Electrical Engineering Handbook*, Chap. 11. CRC, Boca Raton, FL.)



## Defining Terms

**Bode diagram:** A frequency response plot of 20-log gain and phase angle on a log-frequency base.

**Breakpoint:** A point of abrupt change in slope in the straight line approximation of a Bode magnitude curve.

**Damping ratio:** The ratio between a system's damping factor (measure of rate of decay of response) and the damping factor when the system is critically damped.

**Decade:** Power of ten. In context, a tenfold change in frequency.

**Decibel:** A measure of relative size. The decibel gain between voltages  $V_1$  and  $V_2$  is  $20 \log_{10}(V_1/V_2)$ .

**Frequency response:** The frequency-dependent relation in both gain and phase difference between steady state sinusoidal inputs and the resultant steady state sinusoidal outputs.

**Nichols chart:** A plot showing magnitude contours and phase contours of the closed-loop transfer function referred to ordinates of logarithmic loop gain and abscissas of loop phase angle.

**Nyquist plot:** A parametric frequency response plot with the real part of the transfer function on the abscissa and the imaginary part of the transfer function on the ordinate.

**Resonance:** The enhancement of the response of a physical system to a steady state sinusoidal input when the excitation frequency is near a natural frequency of the system.

## References

Dorf, R. C. 1986. *Modern Control Systems*, 4th ed. Addison-Wesley, Reading, MA.

- Dorf, R. C. (Ed.) 1993. *The Electrical Engineering Handbook*. CRC, Boca Raton, FL.
- Franklin, G. F., Powell, J. D., and Emani-Naeini, A. 1994. *Feedback Control of Dynamic Systems*, 3rd ed. Addison-Wesley, Reading, MA.
- Gillespie, T. B. 1992. *Fundamentals of Vehicle Dynamics*. Society of Automotive Engineers, Warrendale, PA.
- Golten, J. and Verwer, A. 1991. *Control System Design and Simulation*. McGraw-Hill, New York.
- IEEE. *IEEE Standard Dictionary of Electrical and Electronics Terms*, 4th ed. 1988. Institute of Electrical and Electronics Engineers, Piscataway, NJ.
- Neudorfer, P. O. and Hassul, M. 1990. *Introduction to Circuit Analysis*. Prentice Hall, Englewood Cliffs, NJ.
- Palm, W. J., III. 1986. *Control Systems Engineering*. John Wiley & Sons, New York.
- Williams, J. B. 1991. *Image Clarity: High Resolution Photography*. Focal, Boston, MA.

Ziemer, R. E. "Convolution Integral"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



# Convolution Integral

---

## 99.1 Fundamentals

Continuous-Time Convolution • Discrete-Time Convolution

## 99.2 Properties of the Convolution Operation

## 99.3 Applications of the Convolution Integral

Filtering • Spectral Analysis • Correlation or Matched Filtering

## 99.4 Two-Dimensional Convolution

## 99.5 Time-Varying System Analysis

### Rodger E. Ziemer

*University of Colorado, Colorado Springs*

The mathematical operation of **convolution** is defined in the first sections of this chapter, both for **continuous-time** and **discrete-time signals**. Properties of convolution are given and illustrated. Several applications are then enumerated and illustrated. Further discussions and illustrations of convolution are given in Ziemer *et al.* [1993] and Close and Frederick [1993]. A good reference on engineering mathematics—including many of the system analysis tools arising in linear systems where the convolution integral arises—is Kreyszig [1988].

## 99.1 Fundamentals

---

### Continuous-Time Convolution

Let  $x(t)$  and  $h(t)$  be two continuous-time signals defined for  $-\infty < t < \infty$ . Their **continuous-time convolution** is defined as the integral operation

$$y(t) = \int_{-\infty}^{\infty} x(\tau)h(t - \tau) d\tau \quad (99.1)$$

With the change of variables  $\lambda = t - \tau$ , Eq. (99.1) can be written in the equivalent form

$$y(t) = \int_{-\infty}^{\infty} x(t - \lambda)h(\lambda) d\lambda \quad (99.2)$$

Inspection of Eq. (99.1) reveals that the convolution operation is composed of four steps with respect to the variable of integration:

1. **Shifting** of  $h(\lambda)$ , represented by replacing  $\lambda$  by  $\lambda - t$
2. Reversal with respect to the independent variable, also known as **folding**, which gives the signal  $h(t - \lambda)$
3. Multiplication by the second signal to produce the integrand  $x(t)h(t - \lambda)$
4. Integration of the product with respect to  $\lambda$  for all values of the delay variable  $t$ , which is the independent variable of the new signal  $y(t)$

**Example 99.1.** Consider the convolution of the following two signals:

$$x(t) = \begin{cases} e^{-\alpha t}, & t \geq 0, \\ 0, & t < 0 \end{cases} \quad \alpha > 0 \quad (99.3)$$

$$h(t) = \begin{cases} e^{-\beta t}, & t \geq 0, \\ 0, & t < 0 \end{cases} \quad \alpha \neq \beta > 0 \quad (99.4)$$

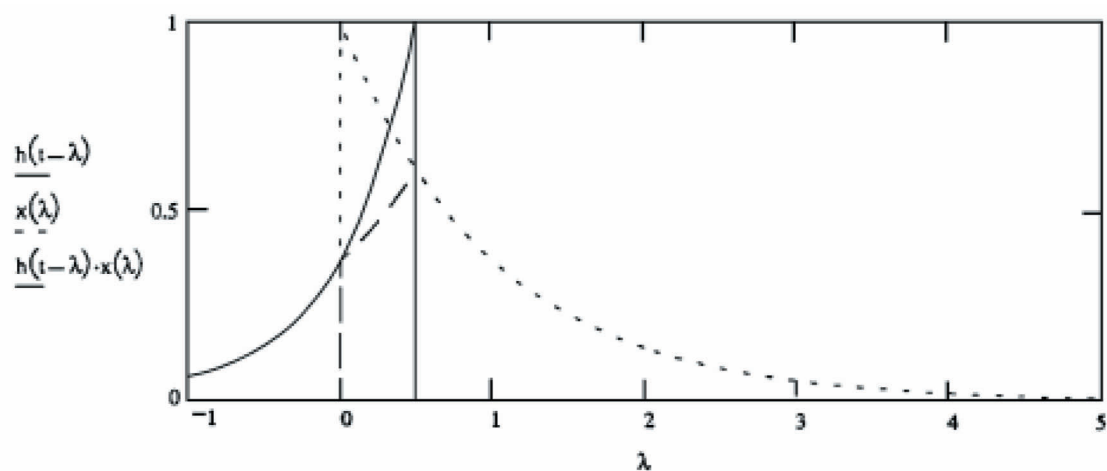
The convolution of these two signals, using Eq. (99.1), is given by

$$y(t) = \int_0^t e^{-\alpha \lambda} e^{-\beta(t-\lambda)} d\lambda = e^{-\beta t} \int_0^t e^{-(\alpha-\beta)\lambda} d\lambda = \frac{e^{-\beta t} - e^{-\alpha t}}{\alpha - \beta}, \quad (99.5)$$

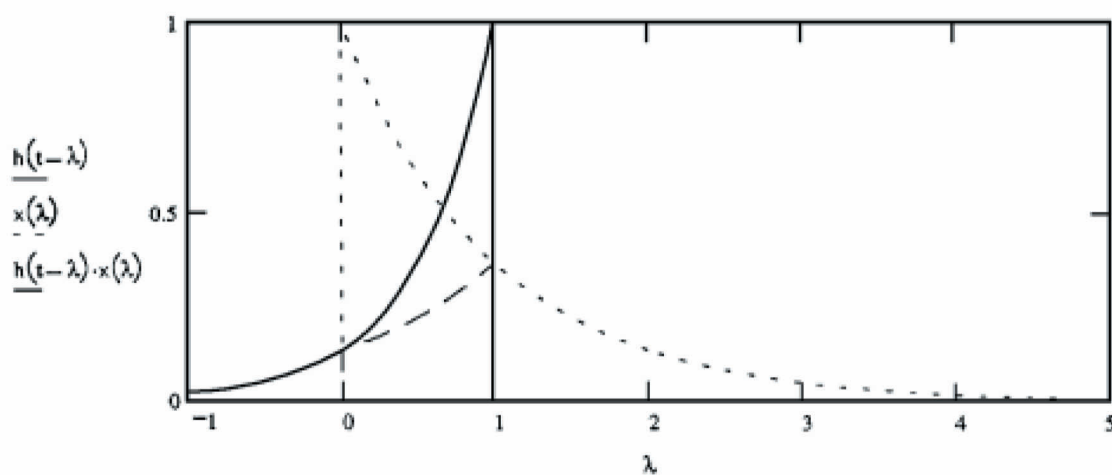
$t \geq 0$

where the lower limit on the integral is 0, by virtue of  $x(\tau) = 0$  for  $\tau < 0$ , and the upper limit is  $t$ , by virtue of  $h(t - \tau) = 0$  for  $t - \tau < 0$  or  $\tau > t$ . [Figures 99.1\(a–c\)](#) illustrate the factors in the integrand for various values of  $t$ , and [Fig. 99.1\(d\)](#) shows the result of the convolution for  $0 \leq t \leq 6$ .

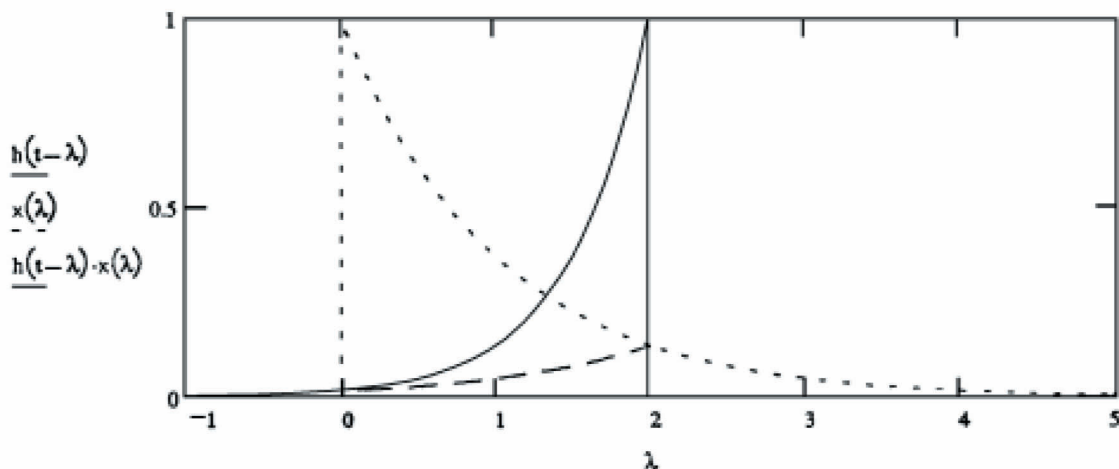
**Figure 99.1** Steps in the convolution of two exponential signals and the final result ( $\alpha = 2$  and  $\beta = 1$ ). See Example 99.1 for mathematical details.



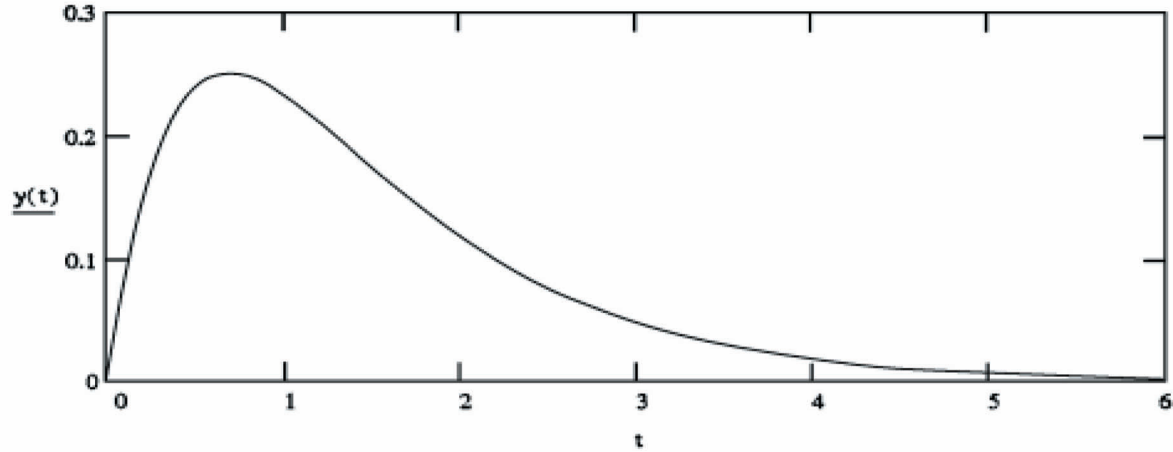
(a) Integrand factors for  $t = 0.5$ .



(b) Integrand factors for  $t = 1$ .



(c) Integrand factors for  $t = 2$ .



(d) Final result for the convolution.

Convolution is often used to find the output of a **linear time-invariant system** to an input signal. (The linearity property means that superposition holds; that is, an arbitrary linear combination of two inputs applied to a linear system results in the same linear combination of outputs due to the inputs applied separately. *Time invariant* means that if a given input results in a certain output, the output due to the input delayed is the original output delayed by the same amount as the input.) When used in this context, Eq. (99.1) or (99.2) involves the following functions or **signals**: the input signal  $x(t)$ , the output signal  $y(t)$ , and the **impulse response**,  $h(t)$ , which is the response of the system to a unit impulse function applied at time zero. The latter completely characterizes the system. A system is **causal** if its output does not anticipate the application of an input, which is manifested in the impulse response being zero for  $t < 0$ . The system is **stable** if every bounded input results in a bounded output, which is manifested by the impulse response being **absolutely integrable**.

## Discrete-Time Convolution

Given two discrete-time signals  $x(n)$  and  $h(n)$  defined for  $-\infty < n < \infty$ , their **discrete-time convolution** (sum) is

$$y(n) = \sum_{k=-\infty}^{\infty} x(k)h(n-k) \quad (99.6)$$

With the change of variables  $j = n - k$ , this can be written as

$$y(n) = \sum_{k=-\infty}^{\infty} x(n-j)h(j) \quad (99.7)$$

Discrete-time convolution involves the same operations as continuous-time convolution except that integration is replaced by summation.

**Example 99.2.** Consider the convolution of the following two discrete-time signals:

$$x(n) = \begin{cases} 1, & n = 0, 1, 2 \\ 0, & n < 0 \text{ and } n > 2 \end{cases} \quad (99.8)$$

$$h(n) = \begin{cases} 0, & n \leq 0 \text{ and } n > 3 \\ 1, & n = 1 \\ 2, & n = 2 \\ 3, & n = 3 \end{cases} \quad (99.9)$$

The discrete-time convolution can be carried out as shown in [Table 99.1](#). Note that the duration of  $x(n)$  is 3, that of  $h(n)$  is 3, and that of  $y(n)$  is  $3 + 3 - 1 = 5$ .

**Table 99.1** Discrete-Time Convolution

$x(n)$							
$h(-n)$	1	1	1	0	0	$n =$	$y(n)$
0 3 2 1	0					-1	0
0 3 2	1	0				0	1
0 3	2	1	0			1	3
0	3	2	1	0		2	6
	0	3	2	1	0	3	5
	0	0	3	2	1	4	3
	0	0	0	3	2	5	0
	0	0	0	0	3	6	0

## 99.2 Properties of the Convolution Operation

Several convenient **properties of convolution** can be proved. These are listed in [Table 99.2](#). Hints at proving these properties are given in the right-hand column.

**Table 99.2** Properties of the Convolution Integral

Number	Property	Comments on Proof
1	$x(t) * h(t) = h(t) * x(t)$ $\text{or } x(n) * h(n) = h(n) * x(n)$ <p style="text-align: center;">(commutative)</p>	Compare Eqs. (99.1) and (99.2) or Eqs. (99.6) and (99.7).
2	$x(t) * [\alpha h(t)] = \alpha [x(t) * h(t)]$ $\text{or } x(n) * [\alpha h(n)] = \alpha [x(n) * h(n)]$ <p style="text-align: center;"><math>\alpha</math> constant</p>	Convolution is an integral from which the constant $\alpha$ can be taken outside.
3	$x(t) * [y(t) * z(t)] = [x(t) * y(t)] * z(t)$ $\text{or } x(n) * [y(n) * z(n)] = [x(n) * y(n)] * z(n)$ <p style="text-align: center;">(associative)</p>	Write the two convolution operations as integrals and reverse the orders of integration.
4	$x(t) * [y(t) + z(t)] = x(t) * y(t) + x(t) * z(t)$ $\text{or } x(n) * [y(n) + z(n)] = x(n) * y(n) + x(n) * z(n)$ <p style="text-align: center;">(distributive)</p>	Separate the integral involving the convolution of $x(t)$ with the sum of the other two signals into the sum of two convolution integrals.
5	If the duration of $x(t)$ is $T_1$ and the duration of $y(t)$ is $T_2$ , then the duration of their convolution is $T_1 + T_2$ . For discrete-time sequences, the length of the convolution is $N_1 + N_2 - 1$ , where $x(n)$ has length $N_1$ and $y(n)$ has length $N_2$ .	Evident by sketching time-limited versions of the integrand or summand factors and considering the overlap.
6	$x(t) * \delta(t - \tau) = x(t - \tau),$ <p style="text-align: center;"><math>\tau</math> constant</p> $x(n) * \delta(n - n_0) = x(n - n_0),$ <p style="text-align: center;"><math>n_0</math> constant</p>	Use the sifting property of the delta function inside convolution integral, or the sifting property of the unit pulse function inside the convolution sum.
7	$\mathcal{F}[x(t) * h(t)] = \mathcal{F}[x(t)]\mathcal{F}[h(t)]$ $\mathcal{L}[x(t) * h(t)] = \mathcal{L}[x(t)]\mathcal{L}[h(t)]$ $\mathcal{Z}[x(n) * h(n)] = \mathcal{Z}[x(n)]\mathcal{Z}[h(n)]$ <p style="text-align: center;"><math>\mathcal{F}(\cdot)</math> = Fourier transform</p> <p style="text-align: center;"><math>\mathcal{L}(\cdot)</math> = Laplace transform</p> <p style="text-align: center;"><math>\mathcal{Z}(\cdot)</math> = <math>z</math> transform</p>	Write the Fourier or Laplace transform integrals and interchange the order of transforming and convolution. For the $z$ transform, interchange the order of the $z$ transform and convolution sums.
8	$x(t) * y(-t) = \int_{-\infty}^{\infty} x(\lambda)y(t + \lambda) d\lambda$ <p style="text-align: center;">= correlation of</p> <p style="text-align: center;"><math>x(t)</math> and <math>y(t)</math></p>	Write out the convolution integral with the argument of the second signal negated. A similar expression holds for discrete-time sequences.
9	$x(t) * e^{j\omega t} = H(j\omega)e^{j\omega t}$ <p style="text-align: center;">where <math>H(j\omega) = \int_{-\infty}^{\infty} h(t)e^{j\omega t} dt</math></p> $x(n) * e^{jn\omega T} = H(e^{j\omega T})e^{jn\omega T}$ <p style="text-align: center;">where <math>H(e^{j\omega T}) = \sum_{k=-\infty}^{\infty} h(k)e^{jk\omega T}</math></p>	Do a direct substitution into the convolution integral, factor out the term $\exp(j\omega t)$ , and use the given definition of $H(j\omega)$ . A similar derivation holds for discrete-time sequences.

## 99.3 Applications of the Convolution Integral

---

### Filtering

From property 7 of Table 99.2, it follows that

$$\begin{aligned}|Y(f)| &= |H(f)||X(f)| \\ \angle Y(f) &= \angle H(f) + \angle X(f)\end{aligned}\quad (99.10)$$

where  $f = \omega/2\pi$  is frequency in hertz. The first equation shows that the magnitude of the spectral content of  $x(t)$  is enhanced or attenuated by the magnitude of  $H(f)$ , which is known as the **amplitude response** of the system. Similarly, the phase components of the spectrum of  $x(t)$  are shifted by the argument of  $H(f)$ , which is known as the **phase response** of the system. Standard design techniques can be used to design filters to modify the amplitude and phase components of  $x(t)$  as desired. Similar statements can be applied to a discrete-time system where  $H(e^{j2\pi fT})$  plays a role similar to  $H(f)$ .

### Spectral Analysis

The approach for determining the presence of spectral components in a signal is similar to filtering, where now the amplitude response function of the filter is designed to pass the desired spectral components of the signal. In the case of random signals, spectral analysis is sometimes better carried out by estimating the autocorrelation function of the signal, Fourier transforming it, and looking for suspected spectral content [Bendat and Piersol, 1980].

### Correlation or Matched Filtering

An important problem in signal detection involves maximizing the peak signal-to-rms-noise ratio at the output of a filter at a certain time. The filter that does this is called the **matched filter** for the particular signal under consideration. Its impulse response is the time reverse of the signal to which it is matched, or, for a signal  $s(t)$ , the matched filter impulse response is

$$h_m(t) = s(t_0 - t) \quad (99.11)$$

where  $t_0$  is the time at which the output peak signal-to-rms-noise ratio is a maximum. The constant  $t_0$  is often chosen to make the impulse response of the matched filter causal. In this case **causality** can be taken to mean that the impulse response is zero for  $t < 0$ . If  $x(t) = 0$  outside of the interval  $[0, T]$ , Eq. (99.1) can be used to write the output of a matched filter to signal alone at its input as

$$y(t) = \int_0^T s(\tau)s(t_0 - t + \tau) d\tau \quad (99.12)$$

If  $t_0 = T$  and the output is taken at  $t = T$ , Eq. (99.12) becomes the energy in the

signal.

$$y(T) = \int_0^T s(\tau)s(\tau) d\tau = \int_0^T s^2(\tau) d\tau = E \quad (99.13)$$

The peak output signal squared-to-mean-square noise ratio can be shown to be

$$\text{SNR}_{\text{out}} = \frac{2E}{N_0} \quad (99.14)$$

where  $N_0$  is the single-sided power spectral density of the white input noise.

## 99.4 Two-Dimensional Convolution

---

Convolution of two-dimensional signals, such as images, is important for several reasons, including pattern recognition and image compression. This topic will be considered briefly here in terms of discrete-spatial-variable signals. Consider two-dimensional signals defined as  $x(m, n)$  and  $h(m, n)$ . Their **two-dimensional convolution** is defined as

$$y(m, n) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} x(j, k)h(m - j, n - k) \quad (99.15)$$

A double change of summation variables permits this to also be written as

$$y(m, n) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} x(m - j, n - k)h(j, k) \quad (99.16)$$

Properties similar to those shown for the one-dimensional case can also be given.

**Example 99.3.** The shift property given in [Table 99.2](#) (generalized to two dimensions) and superposition will be used to convolve the two-dimensional signals given in the table that follows (assumed zero outside of its definition array), with

$$h(m, n) = \delta(m - 1, n - 1) + \delta(m - 2, n - 1) \quad (99.17)$$

For  $x(m, n)$ ,



2	2	1	0	0
2	1	0	0	0
1	0	0	0	0
0	0	0	0	0
0	0	0	0	0

The two-dimensional convolution of these two signals is given by

$$y(m, n) = x(m - 1, n - 1) + x(m - 2, n - 1) \quad (99.18)$$

The result of the two-dimensional convolution is given in the table that follows, for  $y(m, n)$  :

0	0	0	0	0
0	2	4	3	1
0	2	3	1	0
0	1	1	0	0
0	0	0	0	0

## 99.5 Time-Varying System Analysis

---

The form of the convolution integral given in Eq. (99.1), when applied to finding the output of a system to a given input, assumes that the system is time-invariant—that is, its properties do not change with time. In this context  $x(t)$  is viewed as the input to the system,  $y(t)$  its output, and  $h(t)$  the response of the system to a delta function applied at time zero. The generalization of Eq. (99.1) to time-varying (i.e., *not* time-invariant) systems is

$$y(t) = \int_{-\infty}^{\infty} x(\tau)h(t, \tau) d\tau \quad (99.19)$$

where now  $h(t, \tau)$  is the response of the system at time  $t$  to a delta function applied at time  $\tau$ .

### Defining Terms

**Amplitude response:** The amount of attenuation or amplification given to a steady state sinusoidal input by a filter or time-invariant linear system. This is the magnitude of the Fourier transform of the impulse response of the system at the frequency of the sinusoidal input.

**Causal:** A property of a system stating that the system does not respond to a given input before that input is applied. Also, for a linear time-invariant system, **causality** implies that its impulse response is zero for  $t < 0$ .

**Continuous-time signal (system):** A signal for which the independent variable—quite often time—takes on a continuum of values. When applied to a system, continuous time refers to

the fact that the system processes continuous-time signals.

**Convolution:** The process of taking two signals, reversing one in time and shifting it, then multiplying the signals point by point and integrating the product (**continuous-time convolution**). The result of the convolution is then a signal whose independent variable is the shift used in the operation. When done for all shifts, the resulting convolution then produces the output signal for all values of its independent variable, quite often time. When convolving discrete-time signals, the integration of continuous-time convolution is replaced by summation and referred to as the **convolution sum** or **discrete-time convolution**.

**Discrete-time signal (system):** A signal for which the independent variable takes on a discrete set of values. When applied to a system, discrete time refers to the fact that the system processes discrete-time signals.

**Folding:** The process of reversing one of the signals in the convolution integral or sum.

**Impulse response:** The response of an LTI system to a unit impulse function applied at time zero. For a system that is linear but not time invariant, the impulse response,  $h(t, \tau)$ , is the response of the system at time  $t$  to a unit impulse function applied at time  $\tau$ .

**Linear time-invariant (LTI) system:** A system for which superposition holds and for which the output is invariant to time shifts of the input.

**Matched filter:** An LTI system whose impulse response is the time reverse of the signal to which it is matched.

**Phase response:** The amount of phase shift given to a steady state sinusoidal input by a filter or linear time-invariant system. This is the argument of the Fourier transform of the impulse response of the system at the frequency of the sinusoidal input.

**Properties of convolution:** Properties of the convolution operation given in [Table 99.2](#) result from the form of the defining integral (continuous-time case) or sum (discrete-time case). They can be used to simplify application of the convolution integral or sum.

**Signal:** Usually a function of time, but also may be a function of spatial and time variables. Signals can be classified in many different ways. Two classifications used here are continuous time and discrete time.

**Shifting:** An operation used in evaluating the convolution integral or sum. It consists of shifting the folded signal by the other one, multiplying point-by-point, and integrating.

**Stable:** A system is stable if every bounded input results in a bounded output. For an LTI system, stability means that the impulse response is **absolutely integrable**.

**Superposition:** A term that can refer to the superposition integral or sum, which is identical to the convolution integral or sum except that it refers to the response of an LTI system, or can apply to the property of superposition that defines a linear system.

**Time-invariant system:** A system whose properties do not change with time. When expressed in terms of the impulse response of a time-invariant linear system, this means that the impulse response is a function of a single variable, because the unit impulse function that is applied to give the impulse response can be applied at time zero.

**Two-dimensional convolution:** The process of convolution applied to signals dependent on two variables, usually spatial.

## References

- Bendat, J. S. and Piersol, A. G. 1980. *Engineering Applications of Correlation and Spectral Analysis*, John Wiley & Sons, New York.
- Close, C. M. and Frederick, D. K. 1993. *Modeling and Analysis of Dynamic Systems*, 2nd ed. Houghton Mifflin, Boston, MA.
- Kreyszig, E. 1988. *Advanced Engineering Mathematics*, 6th ed. John Wiley & Sons, New York.
- Rioul, O. and Vetterli, M. 1991. Wavelets in signal processing. *IEEE Signal Processing Magazine*. 8:14–38, October.
- Ziemer, R. E., Tranter, W. H., and Fannin, D. R. 1993. *Signals and Systems: Continuous and Discrete*, 3rd ed. Macmillan, New York.

## Further Information

The convolution integral finds application in linear system analysis and signal processing. In the book by Ziemer, Tranter, and Fannin [1993] chapter 2 is devoted to the convolution integral and its application to continuous-time linear system analysis, whereas chapter 8 includes material on the discrete-time convolution sum.

The *IEEE Signal Processing Magazine* has tutorial articles on various aspects of signal processing, whereas the *IEEE Transactions on Signal Processing* include research papers on all aspects of signal processing. One notable and timely application of convolution is to wavelet transform theory, which has many potential applications including signal and image compression. The April 1992 issue of the *IEEE Signal Processing Magazine* contains a tutorial article on wavelet transforms.

Stefani, R. T. "Stability Analysis"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

100.1 Response Components

100.2 Internal (Asymptotic) and External (BIBO) Stability

100.3 Unstable and Marginally Stable Responses

100.4 Structural Integrity

**Raymond T. Stefani**

*California State University, Long Beach*

The output response of a linear system has two components: the zero-state response caused by the input and the zero-input response caused by the initial conditions. It is desirable that each of those responses be well behaved. The speed of an automobile, for example, must provide a comfortable ride for the passengers. The term *stability* refers to how those two response components behave. The following sections define types of stability, relationships between stability types, and methods of determining the type of stability for a given system. Several response examples are provided.

## 100.1 Response Components

---

Figure 100.1 shows the two kinds of response components for a linear system. The **zero-state** component occurs when the initial condition vector  $x_0$  is zero and the input  $r$  is nonzero.

Conversely, the **zero-input** response occurs when the input  $r$  is zero but the initial condition vector  $x_0$  is nonzero. Then, when both inputs are nonzero, a typical system has a response containing both zero-input and zero-state responses.

**Figure 100.1** Zero-state and zero-input components.

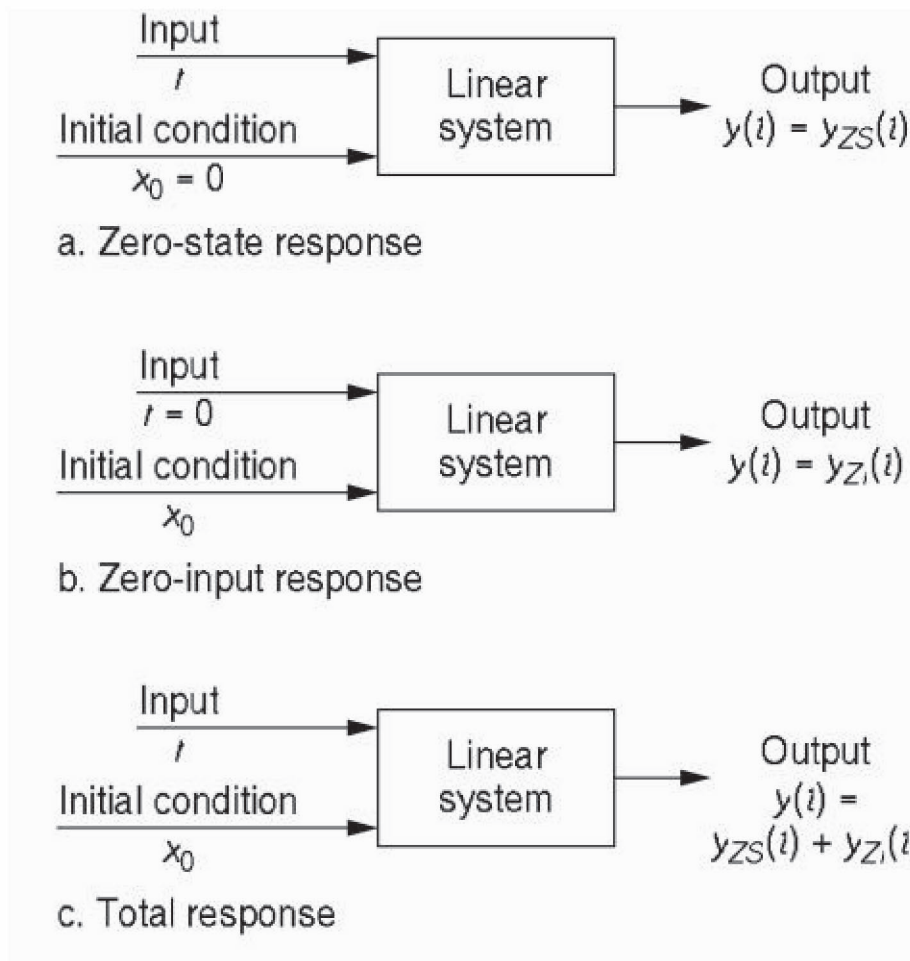


Table 100.1 shows the Laplace transform and resulting time response where the zero-state and zero-input responses are defined using classical mathematics (based on transfer functions) and state variable mathematics (based on system matrices). In the time domain the state variable–based description is

$$dx/dt = Ax + Br \quad (100.1)$$

$$y = Cx \quad (100.2)$$

The classical transfer function can be related to the  $(A, B, C)$  system by

$$H(s) = Y(s)/R(s) = C[sI - A]^{-1}B \quad (100.3)$$

**Table 100.1** Classical and State Variable Forms for Zero-State and Zero-Input Responses

	Laplace	Time
<b>Zero State</b>		
Classical(general)	$H(s)R(s) = \frac{N(s)}{D(s)}R(s)$	$\int_0^t h(\tau)r(t-\tau) d\tau$
Classical (impulse)	$H(s)$	$h(t)$
State variable (general)	$C\phi(s)BR(s)$	$\int_0^t C\phi(\tau)Br(t-\tau) d\tau$
State variable (impulse)	$C\phi(s)B$	$C\phi(t)B$
<b>Zero Input</b>		
Classical	$H_1(s) = \frac{N_1(s, x_0)}{D(s)}$	$h_1(t)$
State variable	$C\phi(s)x_0$	$C\phi(t)x_0$
Transfer function = $H(s) = C\phi(s)B$		
Resolvent matrix = $\phi(s) = [sI - A]^{-1}$		
Characteristic polynomial = $D(s) =  sI - A $		
Initial state vector = $x_0$		

An important type of zero-state response is the **impulse response** in which the input  $r$  is a unit impulse function  $\delta(t)$  whose Laplace transform  $R(s)$  equals 1. Notice in Table 100.1 the close relationship between the form of the zero-state impulse response  $C\phi(t)B$  and the zero-input response  $C\phi(t)x_0$ . Thus stability definitions based on the impulse response bear a close relationship to stability definitions based on the zero-input response.

## 100.2 Internal (Asymptotic) and External (BIBO) Stability

Stability definitions relate either to the zero-input response (**internal stability**) or to the zero-state response (**external stability**). The latter is called *external* because performance depends on the external input as it influences the external output (internal performance is ignored). Table 100.2 contains two commonly used stability definitions: asymptotic stability relates to zero-input (internal) stability, whereas the bounded-input, bounded-output (BIBO) stability relates to zero-state (external) stability.

**Table 100.2** Stability Definitions for Linear Systems

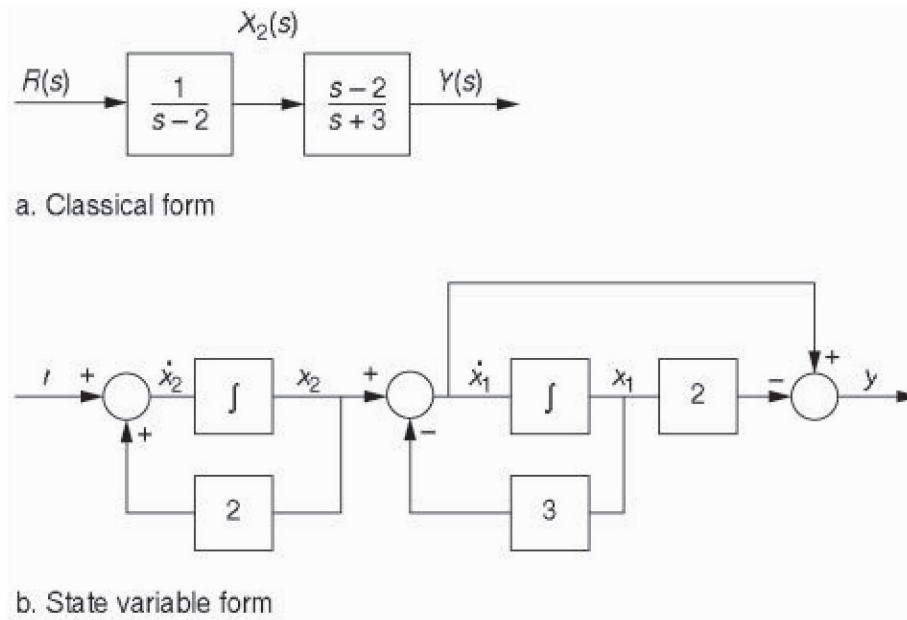
Type of Stability	Requirement	Necessary and Sufficient Condition
Asymptotic (internal) (zero input)	$C\phi(t)x_0 \rightarrow 0$ as $t \rightarrow \infty$ for all $x_0$	All eigenvalues of $A$ (closed-loop poles) are in the LHP
BIBO (external) (zero state)	If $ r(t)  \leq M_1 < \infty$ then $ y(t)  \leq M_2 < \infty$	$\int_0^\infty  h(t) dt \leq M_2 < \infty$ (all poles of $H(s)$ are in the LHP)

Asymptotic stability implies that the zero-input response tends to zero for all initial conditions. The necessary and sufficient condition for asymptotic stability is that all the eigenvalues  $A$  are in the left half plane (LHP). BIBO stability implies that all bounded inputs will cause the output to be bounded. The necessary and sufficient condition for BIBO stability is that the integral of the magnitude of the impulse response is bounded, which occurs if all the poles of  $H(s)$  are in the

LHP. These two stability definitions will result in different conclusions when there are right half plane (RHP) poles that cancel out of  $H(s)$ , obscuring internal responses that may not be bounded.

Figure 100.2 shows a system in both classical form and state variable form. From the classical form, notice that  $H(s) = 1/(s + 3)$  (an RHP pole cancels out), so the impulse response of  $y(t)$  is  $e^{-3t}$  (whose integral is bounded), causing the system to be BIBO (externally) stable. However, for an impulse input,  $X_2(s)$  is  $1/(s - 2)$  and  $x_2(t)$  is  $e^{2t}$ , which is obviously not bounded. That second (undesirable) effect is predictable from the state variable description

**Figure 100.2** Condition 1 (BIBO stable but asymptotically unstable).



$$A = \begin{bmatrix} -3 & 1 \\ 0 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C = [-5 \quad 1] \quad (100.4)$$

Although  $H(s) = C[sI - A]^{-1}B = 1/(s + 3)$ , passing the test for BIBO stability, the eigenvalues of  $A$  are  $-3$  and  $2$ , failing the test for asymptotic stability.

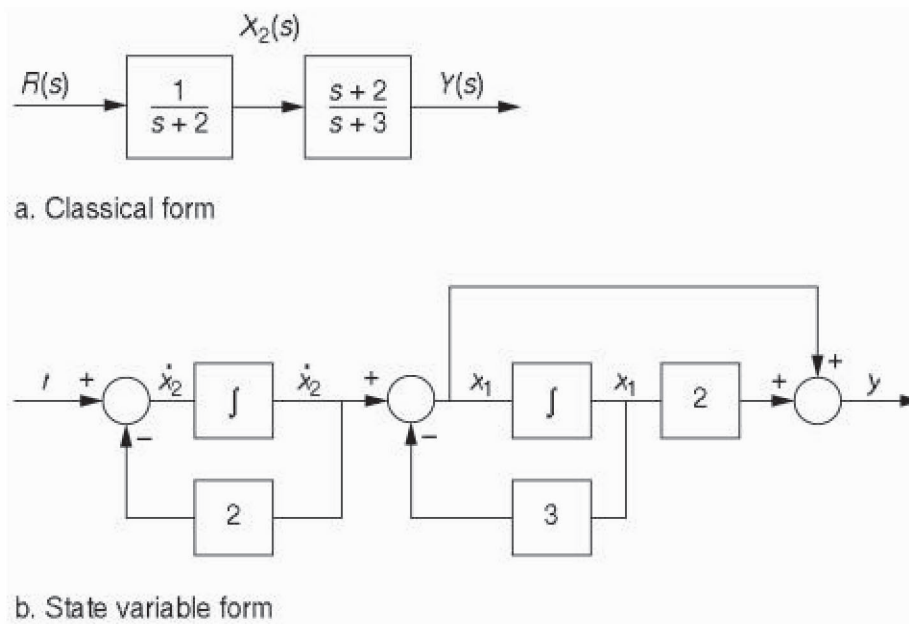
Table 100.3 shows that asymptotic (internal) stability always implies BIBO (external) stability but that BIBO stability implies asymptotic stability only if no RHP poles cancel from the transfer function. For example, Fig. 100.3 shows another system in both classical form and state variable form. From the classical form, notice that  $H(s) = 1/(s + 3)$  (no RHP poles cancel this time, just one LHP pole), so the impulse response of  $y(t)$  is  $e^{-3t}$  (whose integral is bounded), causing the system to be BIBO (externally) stable. Now, for an impulse input,  $X_2(s)$  is  $1/(s + 2)$  and  $x_2(t)$  is  $e^{-2t}$ , which is also bounded. From the state variable description,



**Table 100.3** Equivalence of Stability Types

Equivalence	Requirement
Asymptotic $\rightarrow$ BIBO	Always
BIBO $\rightarrow$ asymptotic	No cancellation of RHP poles in $H(s)$

**Figure 100.3** Condition 2 (BIBO stable and asymptotically stable).



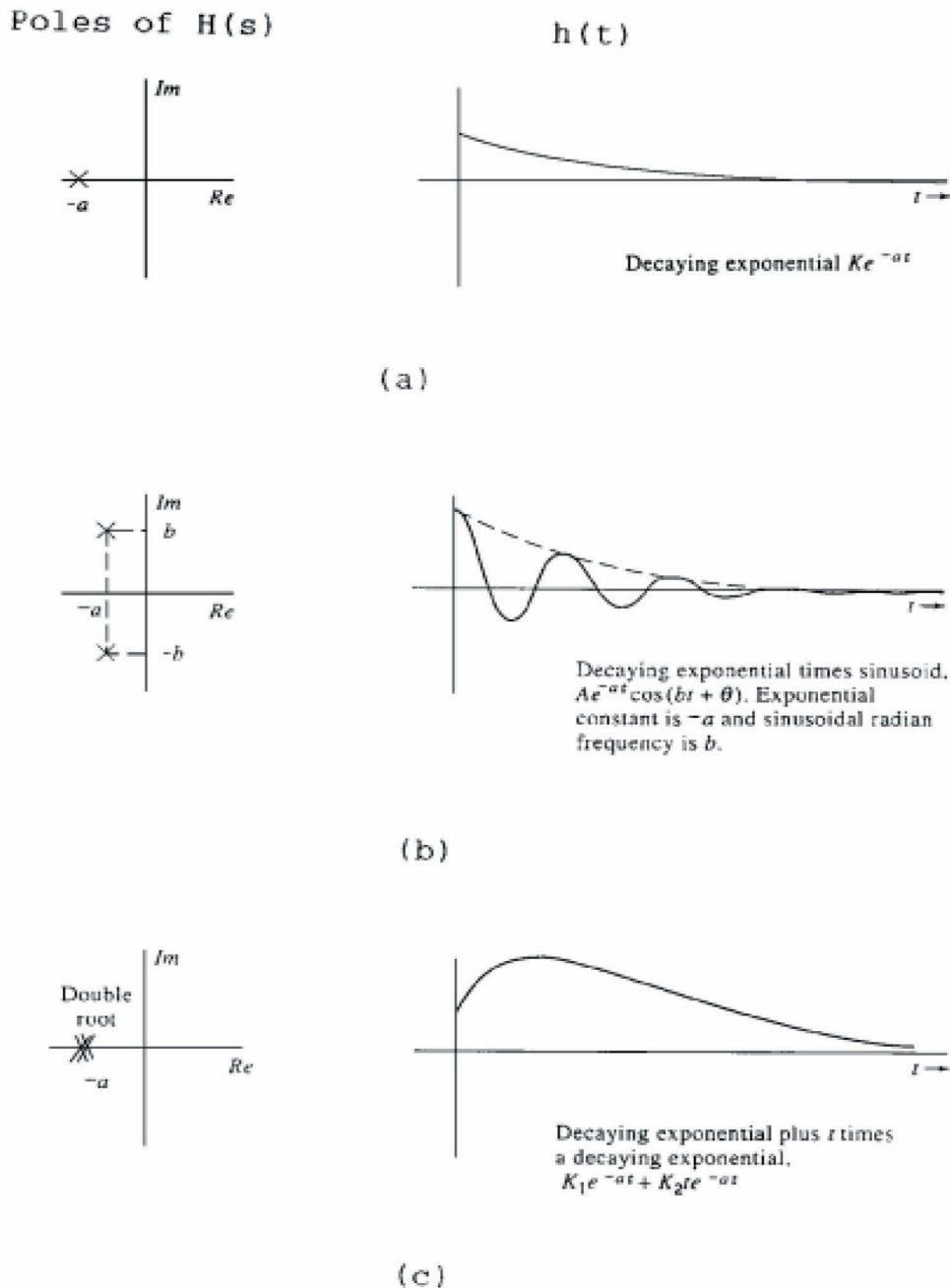
$$A = \begin{bmatrix} -3 & 1 \\ 0 & -2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C = [-1 \quad 1] \quad (100.5)$$

As before,  $H(s) = C[sI - A]^{-1}B = 1/(s + 3)$ , passing the test for BIBO stability; but now the eigenvalues of  $A$  are  $-3$  and  $-2$ , also passing the test for asymptotic stability.

Figure 100.4 shows a number of stable  $H(s)$  functions and the corresponding impulse responses  $h(t)$ . The step responses would differ by a constant.

**Figure 100.4** Impulse response of stable BIBO systems. (Adapted from Stefani, R. T., Savant, C. J., Shahian, B., and Hostetter, G. H. 1994. Design of Feedback Control Systems, 3rd ed, Saunders College, Boston, MA.)

**Figure 100.4**



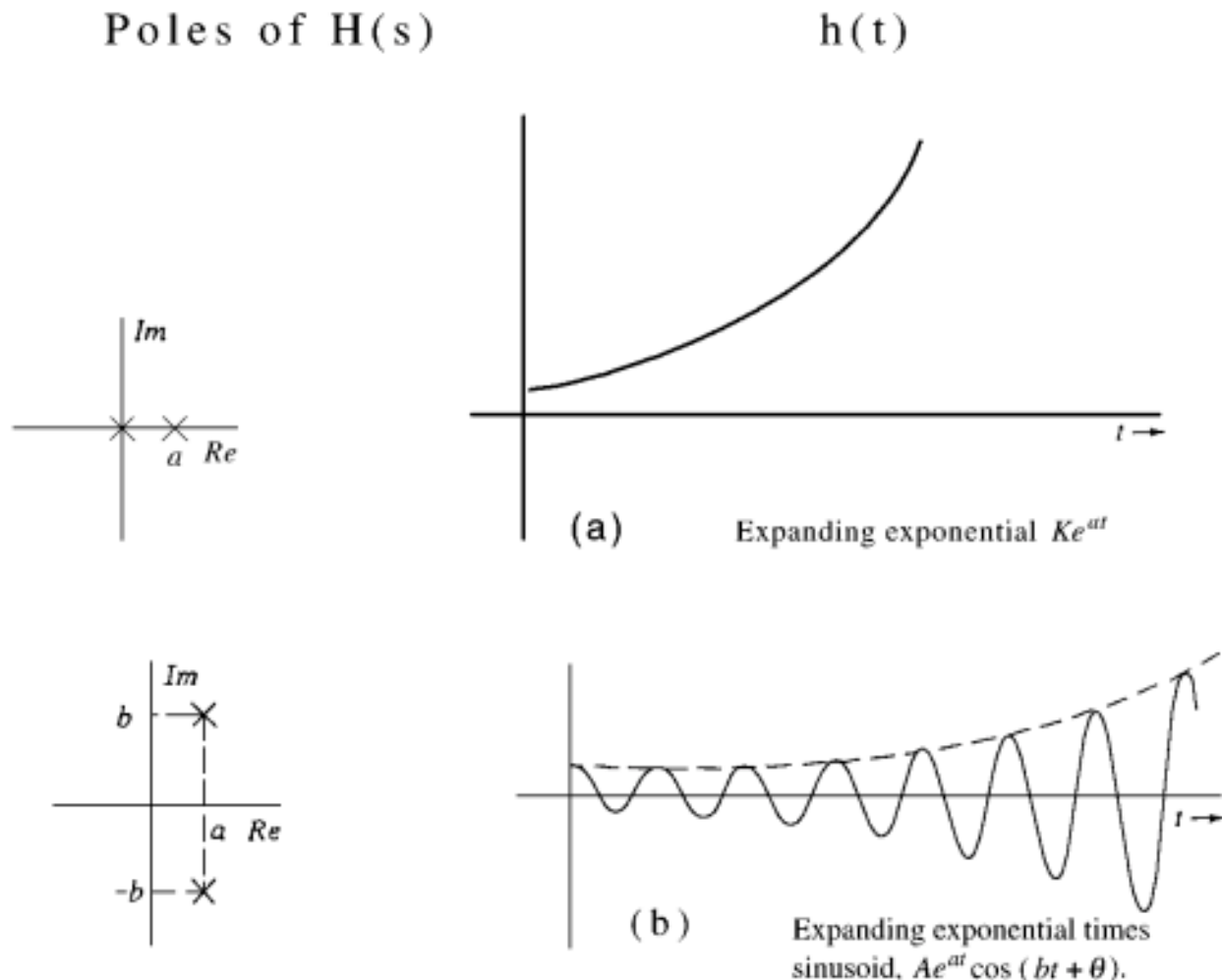
## 100.3 Unstable and Marginally Stable Responses

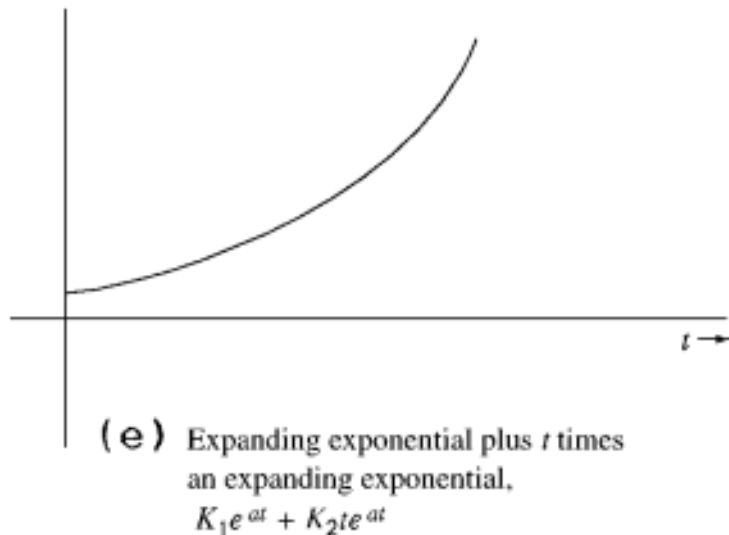
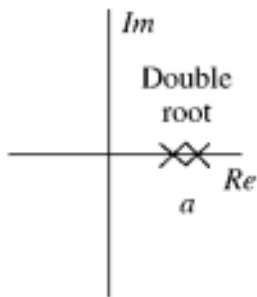
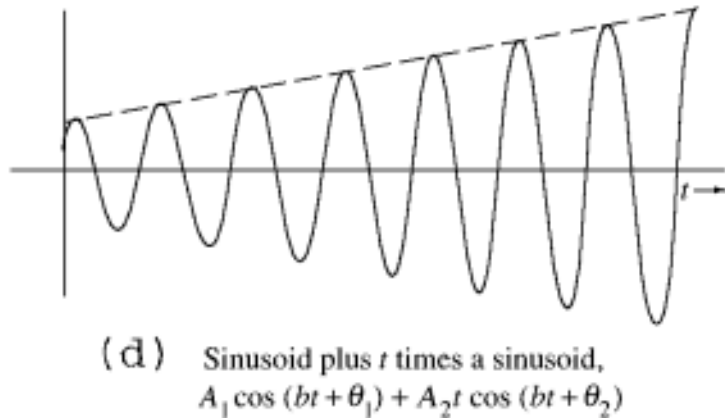
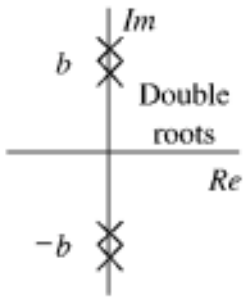
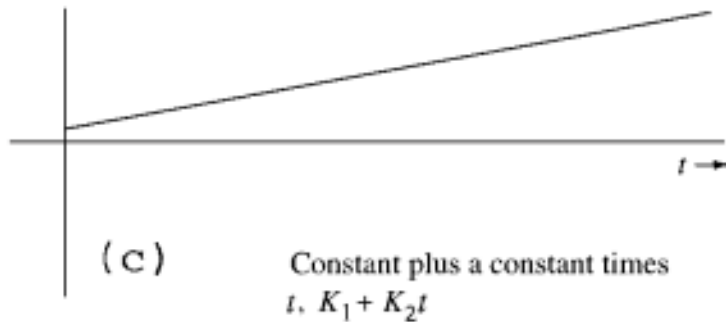
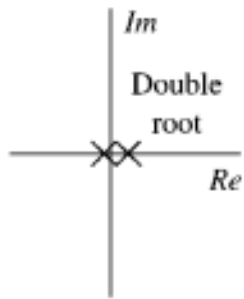
Systems that do not pass the tests for stability may be classified as **unstable** or **marginally stable**, as in Table 100.4. As before, BIBO and asymptotic stability may differ if certain poles of  $H(s)$  are canceled by zeros at the same place. An unstable system has a BIBO response that is unbounded for all bounded inputs and a zero-input response that diverges for at least one initial condition. A system is unstable for closed loop poles (BIBO) and eigenvalues (asymptotic) that are RHP and/or repeated along the imaginary axis (IA). Figure 100.5 shows a number of unstable (BIBO)  $H(s)$  functions and the corresponding impulse responses  $h(t)$ .

**Table 100.4** Unstable and Marginally Stable Definitions for Linear Systems

Type of Stability	Requirement	Necessary and Sufficient Condition
Unstable (asymptotic)	$C\phi(t)x_0 \rightarrow \infty$ as $t \rightarrow \infty$ for at least one $x_0$	At least one eigenvalue of $A$ in RHP and/or repeated IA
Unstable (BIBO)	If $ r(t)  \leq M_1 < \infty$ then $ y(t)  \rightarrow \infty$	At least one pole of $H(s)$ in RHP and/or repeated IA
Marginally stable (asymptotic)	$C\phi(t)x_0 \rightarrow C_0$ $0 <  C_0  < \infty$ as $t \rightarrow \infty$ for all $x_0$	One eigenvalue of $A$ at 0 and/or nonrepeated complex conjugate IA; the rest are LHP
Marginally stable (BIBO)	For some but not all $ r(t)  \leq M_1 < \infty$ then $ y(t)  \leq M_2 < \infty$	One pole of $H(s)$ at 0 and/or nonrepeated complex conjugate IA; the rest are LHP

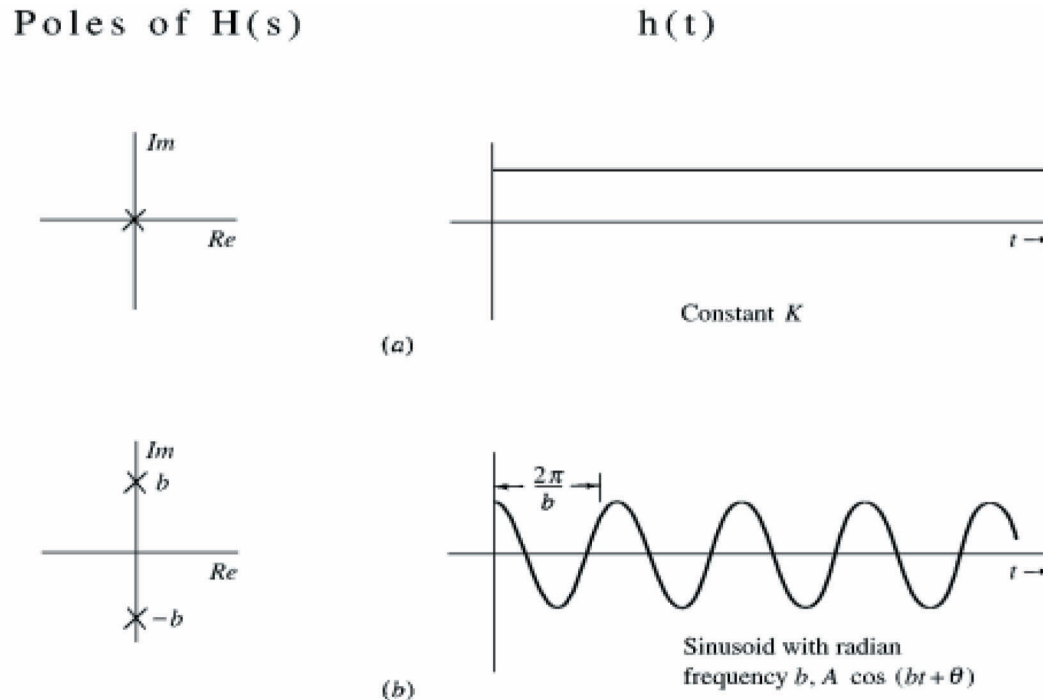
**Figure 100.5** Impulse response of unstable BIBO systems. (Adapted from Stefani, R. T., Savant, C. J., Shahian, B., and Hostetter, G. H. 1994. *Design of Feedback Control Systems*, 2nd ed. Saunders College, Boston, MA.)





A marginally stable system has a BIBO response that is bounded for some inputs but unbounded for others and a zero-input response that is bounded (going neither to zero nor infinity) for all initial conditions. A system is marginally stable for closed loop poles (BIBO) and eigenvalues (asymptotic) that are IA and nonrepeated. Figure 100.6 shows two marginally stable (BIBO)  $H(s)$  functions and the corresponding impulse responses  $h(t)$ . For Fig. 100.6(a),  $H(s) = 1/s$ , so the zero-state impulse response is a constant (which is bounded). However, if the input is a step, the zero-state response becomes  $1/s^2$  and the time response for this case is the same as the impulse response in Fig. 100.5(c), which is obviously unbounded. Similarly, the example of Fig. 100.6(b) has a bounded sinusoidal impulse response for IA poles of  $H(s)$ , but if the bounded  $r(t)$  is also sinusoidal and the IA poles of  $R(s)$  are at the same place as for that  $H(s)$ , then the multiple IA poles of  $Y(s)$  create a time response like that of Fig. 100.5(d), which is unbounded.

**Figure 100.6** Impulse response of marginally stable BIBO systems. (Adapted from Stefani, R. T., Savant, C. J., Shahian, B., and Hostetter, G. H. 1994. *Design of Feedback Control Systems*, 2nd ed. Saunders College, Boston, MA.)



## 100.4 Structural Integrity

A stable system always has a bounded BIBO response; however, that response can have very large values. For example, multiple (LHP) poles often create momentarily large outputs, even though the system is considered stable [see Fig. 100.4(c)]. These large outputs can exceed structural limits. For example, a famous bridge over the Tacoma Narrows collapsed in 1939 because of winds that created lifting forces whose  $R(s)$  had poles near the bridge's  $H(s)$  poles. When a building or bridge collapses during an earthquake, the cause may be similar. Reinforcing a building or bridge moves the poles of  $H(s)$  away from those of  $R(s)$ , thus maintaining the response within structural limits. An engineer must examine more than just stability when ensuring that a design results in a satisfactory time response.

### Defining Terms

**External stability:** Stability based on the bounded-input, bounded-output relationship, ignoring the performance of internal states. This stability is based on the poles of the transfer function, which may contain RHP pole cancellations.

**Impulse response:** The zero-state response of a system to a unit impulse input whose Laplace transform is 1.

**Internal stability:** Stability based on the performance of all the states of the system, and therefore on all the eigenvalues of the system. This type of stability will differ from external stability if

RHP poles cancel from the transfer function.

**Marginally stable:** For external stability, the response is bounded for some but not all bounded inputs.

**Unstable:** For external stability, the output is unbounded for all bounded inputs.

**Zero input:** Response component calculated by setting the input to zero.

**Zero state:** Response component calculated by setting the initial state to zero.

## References

Kailath, T. *Linear Systems*. 1980. Prentice Hall, Englewood Cliffs, NJ.

Kuo, B. C. 1987. *Automatic Control Systems*, 5th. ed. Prentice Hall, Englewood Cliffs, NJ.

Stefani, R. T., Savant, C. J., Shahian, B., and Hostetter, G. H. 1994. *Design of Feedback Control Systems*, 3rd ed. Saunders College, Boston, MA.

## Further Information

*IEEE Transactions on Automatic Control* is a rather theoretically-oriented journal including aspects of stability related to a wide spectrum of linear, nonlinear and adaptive control systems.

*Control Systems* is a magazine published by IEEE with a more basic level of presentation than the *IEEE Transactions on Automatic Control*. Applications are made to the stability of various practical systems.

*IEEE Transactions on Systems, Man, and Cybernetics* is a journal providing coverage of stability as it relates to a variety of systems with human interface or models that emulate the response of humans.

*Automatica* and the *International Journal of Control* cover stability concepts for an international audience.

Johansson, R. "z Transform and Digital Systems"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# **z Transform and Digital Systems**

---

## 101.1 The $z$ Transform

## 101.2 Digital Systems and Discretized Data

The Discrete Fourier Transform

## 101.3 The Transfer Function

State-Space Systems

## 101.4 Digital Systems Described by Difference Equations(ARMAX Models)

## 101.5 Prediction and Reconstruction

## 101.6 The Kalman Filter

### **Rolf Johansson**

*Department of Automatic ControlLund Institute of Technology*

A digital system (or discrete-time system or sampled-data system) is a device such as a digital controller or a digital filter or, more generally, a system intended for digital computer implementation and usually with some periodic interaction with the environment and with a supporting methodology for analysis and design. Of particular importance for modeling and analysis are recurrent algorithms—for example, difference equations in input-output data—and the  $z$  transform is important for the solution of such problems.

The  **$z$  transform** is being used in the analysis of linear time-invariant systems and discrete-time signals—for example, for digital control or filtering—and may be compared to the Laplace transform as used in the analysis of continuous-time signals and systems, a useful property being that the convolution of two time-domain signals is equivalent to multiplication of their corresponding  $z$  transforms. The  $z$  transform is important as a means to characterize a linear time-invariant system in terms of its pole-zero locations, its transfer function and Bode diagram, and its response to a large variety of signals. In addition, it provides important relationships between temporal and spectral properties of signals. The  $z$  transform generally appears in the analysis of difference equations as used in many branches of engineering and applied mathematics.

## **101.1 The $z$ Transform**

---

The  $z$  transform of the sequence  $\{x_k\}_{-\infty}^{+\infty}$  is defined as the generating function

$$X(z) = \mathcal{L}\{x\} = \sum_{k=-\infty}^{\infty} x_k z^{-k} \quad (101.1)$$



where the variable  $z$  has the essential interpretation of a forward shift operator so that

$$\mathcal{L}\{x_{k+1}\} = z\mathcal{L}\{x_k\} = zX(z) \quad (101.2)$$

The  $z$  transform is an infinite power series in the complex variable  $z^{-1}$  where  $\{x_k\}$  constitutes a sequence of coefficients. As the  $z$  transform is an infinite power series, it exists only for those values of  $z$  for which this series converges and the *region of convergence* of  $X(z)$  is the set of  $z$  for which  $X(z)$  takes on a finite value. A sufficient condition for existence of the  $z$  transform is convergence of the power series

$$\sum_{k=-\infty}^{\infty} |x_k| \cdot |z^{-k}| < \infty \quad (101.3)$$

The region of convergence for a finite-duration signal is the entire  $z$  plane except  $z = 0$  and  $z = \infty$ . For a one-sided infinite-duration sequence  $\{x_k\}_{k=0}^{\infty}$ , a number  $r$  can usually be found so that the power series converges for  $|z| > r$ . Then, the *inverse  $z$  transform* can be derived as

$$x_k = \frac{1}{2\pi i} \oint X(z) z^{k-1} dz \quad (101.4)$$

where the contour of integration encloses all singularities of  $X(z)$ . In practice it is standard procedure to use tabulated results; some standard  $z$  transform pairs are to be found in [Table 101.1](#).

**Table 101.1** Properties of the  $z$ 

$z$ transform	$\mathcal{Z}(\{f_k\}) = F(z)$	
Convolution	$\mathcal{Z}(\{f_k * g_k\}) = \mathcal{Z}(\{f_k\}) \cdot \mathcal{Z}(\{g_k\})$	
	$\mathcal{Z}(\{f_k \cdot g_k\}) = \mathcal{Z}(\{f_k\}) * \mathcal{Z}(\{g_k\})$	
Forward shift	$\mathcal{Z}(\{f_{k+1}\}) = z\mathcal{Z}(\{f_k\}) = zF(z)$	
Backward shift	$\mathcal{Z}(\{f_{k-1}\}) = z^{-1}\mathcal{Z}(\{f_k\}) = z^{-1}F(z)$	
Linearity	$\mathcal{Z}(\{af_k + bg_k\}) = a\mathcal{Z}(\{f_k\}) + b\mathcal{Z}(\{g_k\})$	
Multiplication	$\mathcal{Z}(\{a^k f_k\}) = F(a^{-1}z)$	
Final value	$\lim_{k \rightarrow \infty} f_k = \lim_{z \rightarrow 1} (1 - z^{-1})F(z)$	
Initial value	$f_0 = \lim_{z \rightarrow \infty} F(z)$	

	Time Domain		$z$ Transform
Impulse	$\delta_k = \begin{cases} 1, & k = 0 \\ 0, & k \neq 0 \end{cases}$	$\Leftrightarrow$	$\mathcal{Z}\{\delta_k\} = 1, \quad z \in \mathbb{C}$
Step function	$\sigma_k = \begin{cases} 0, & k < 0 \\ 1, & k \geq 0 \end{cases}$	$\Leftrightarrow$	$\mathcal{Z}\{\sigma_k\} = \frac{z}{z-1}, \quad  z  > 1$
Ramp function	$x_k = k \cdot \sigma_k$	$\Leftrightarrow$	$X(z) = \frac{z}{(z-1)^2}, \quad  z  > 1$
Exponential	$x_k = a^k \cdot \sigma_k$	$\Leftrightarrow$	$X(z) = \frac{z}{z-a}, \quad  z  >  a $
Sinusoid	$x_k = \sin \omega k \cdot \sigma_k$	$\Leftrightarrow$	$X(z) = \frac{z \sin \omega}{z^2 - 2z \cos \omega + 1}, \quad  z  > 1$

## 101.2 Digital Systems and Discretized Data

Periodic sampling of signals and subsequent computation or storing of the results requires the computer to schedule sampling and to handle the resulting sequences of numbers. A measured variable  $x(t)$  may be available only as periodic observations of  $x(t)$  as sampled with a time interval

$T$  (the sampling period). The sample sequence can be represented as

$$\{x_k\}_{-\infty}^{\infty}; \quad x_k = x(kT) \quad \text{for} \quad k = \dots, -1, 0, 1, 2, \dots \quad (101.5)$$

and it is important to ascertain that the sample sequence adequately represents the original variable  $x(t)$ ; see [Fig. 101.1](#). For ideal sampling it is required that the duration of each sampling be very short and the sampled function may be represented by a sequence of infinitely short impulses  $\delta(t)$  (the Dirac impulse). Let the sampled function of time be expressed thus:

$$x_{\Delta}(t) = x(t) \cdot T \sum_{k=-\infty}^{\infty} \delta(t - kT) = x(t) \cdot \sqcup_T(t) \quad (101.6)$$

where

$$\sqcup\sqcup_T(t) \triangleq T \sum_{k=-\infty}^{\infty} \delta(t - kT) \quad (101.7)$$

and where the sampling period  $T$  is multiplied to ensure that the averages over a sampling period of the original variable  $x$  and the sampled signal  $x_{\Delta}$ , respectively, are of the same magnitude. A direct application of the discretized variable  $x_{\Delta}(t)$  in Eq. (101.6) verifies that the spectrum of  $x_{\Delta}$  is related to the  $z$  transform  $X(z)$  as

$$X_{\Delta}(i\omega) = \mathcal{F}\{x(t) \cdot \sqcup\sqcup_T(t)\} = T \sum_{k=-\infty}^{\infty} x_k \exp(-i\omega kT) = TX(e^{i\omega T}) \quad (101.8)$$

Obviously, the original variable  $x(t)$  and the sampled data are not identical, and thus it is necessary to consider the distortive effects of discretization. Consider the spectrum of the sampled signal  $x_{\Delta}(t)$  obtained as the Fourier transform

$$X_{\Delta}(i\omega) = \mathcal{F}\{x_{\Delta}(t)\} = \mathcal{F}\{x(t)\} * \mathcal{F}\{\sqcup\sqcup_T(t)\} \quad (101.9)$$

where

$$\mathcal{F}\{\sqcup\sqcup_T(t)\} = \sum_{k=-\infty}^{\infty} \delta\left(\omega - \frac{2\pi}{T}k\right) = \frac{T}{2\pi} \sqcup\sqcup_{2\pi/T}(\omega) \quad (101.10)$$

so that

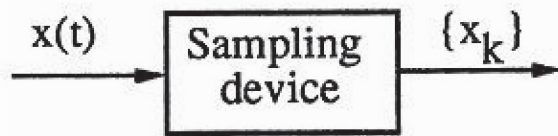
$$X_{\Delta}(i\omega) = \mathcal{F}\{x(t)\} * \mathcal{F}\{\sqcup\sqcup_T(t)\} = \sum_{k=-\infty}^{\infty} X\left[i\left(\omega - \frac{2\pi}{T}k\right)\right] \quad (101.11)$$

Thus, the Fourier transform  $X_{\Delta}$  of the sampled variable has a periodic extension of the original spectrum  $X(i\omega)$  along the frequency axis with a period equal to the sampling frequency  $\omega_s = 2\pi/T$ . There is an important result based on this observation known as *the Shannon sampling theorem*, which states that the continuous-time variable  $x(t)$  may be reconstructed from the samples  $\{x_k\}_{-\infty}^{+\infty}$  if and only if the sampling frequency is at least twice that of the highest frequency for which  $X(i\omega)$  is nonzero. The original variable  $x(t)$  may thus be recovered as

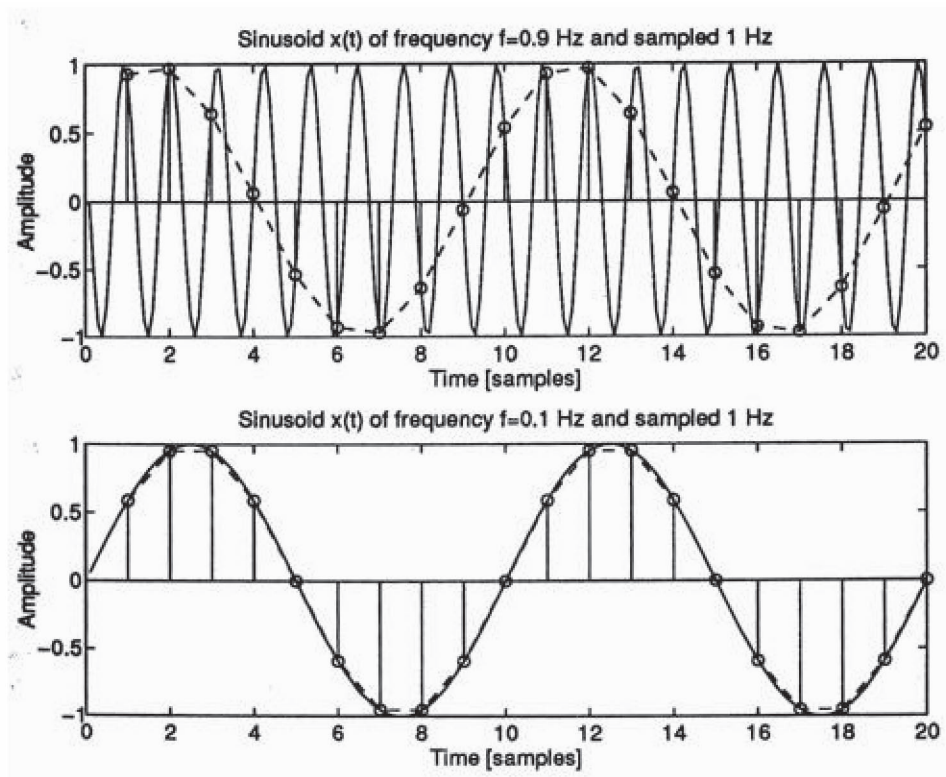
$$x(t) = \sum_{k=-\infty}^{\infty} x_k \frac{\sin \frac{\pi}{T}(t - kT)}{\frac{\pi}{T}(t - kT)} \quad (101.12)$$

The formula given in Eq. (101.12) is called *Shannon interpolation*, which is often quoted though it is valid only for infinitely long data sequences and though it would require a noncausal filter to reconstruct the continuous-time signal  $x(t)$  in real-time operation. The frequency  $\omega_n = \omega_s/2 = \pi/T$  is called the *Nyquist frequency* and indicates the upper limit of distortion-free sampling. A nonzero spectrum beyond this limit leads to interference between the sampling frequency and the sampled signal (*aliasing*); see Fig. 101.2.

**Figure 101.1** A continuous-time signal  $x(t)$  and a sampling device that produces a sample sequence  $\{x_k\}$ .



**Figure 101.2** Illustration of aliasing appearing during sampling of a sinusoid  $x(t) = \sin 2\pi \cdot 0.9t$  at the insufficient sampling frequency 1 Hz (sampling period  $T = 1$ ) (*upper graph*). The sampled signal exhibits aliasing with its major component similar to a signal  $x(t) = \sin 2\pi \cdot 0.1t$  sampled with the same rate (*lower graph*).



## The Discrete Fourier Transform

Consider a finite length sequence  $\{x_k\}_{k=0}^{N-1}$  that is zero outside the interval  $0 \leq k \leq N-1$ .

Evaluation of the  $z$  transform  $X(z)$  at  $N$  equally spaced points on the unit circle

$z = \exp(i\omega_k T) = \exp[i(2\pi/NT)kT]$  for  $k = 0, 1, \dots, N-1$  defines the *discrete Fourier transform* (DFT) of a signal  $x$  with a sampling period  $h$  and  $N$  measurements

$$X_k = \text{DFT} \{x(kT)\} = \sum_{l=0}^{N-1} x_l \exp(-i\omega_k lT) = X(e^{i\omega_k T}) \quad (101.13)$$

Notice that the discrete Fourier transform  $\{X_k\}_{k=0}^{N-1}$  is only defined at the discrete frequency points

$$\omega_k = \frac{2\pi}{NT}k, \quad \text{for } k = 0, 1, \dots, N-1 \quad (101.14)$$

In fact, the discrete Fourier transform adapts the Fourier transform and the  $z$  transform to the practical requirements of finite measurements. Similar properties hold for the discrete Laplace transform with  $z = \exp(sT)$ , where  $s$  is the Laplace transform variable.

### 101.3 The Transfer Function

Consider the following discrete-time linear system with input sequence  $\{u_k\}$  (stimulus) and output sequence  $\{y_k\}$  (response). The dependency of the output of a linear system is characterized by the convolution-type equation and its  $z$  transform,

$$y_k = \sum_{m=0}^{\infty} h_m u_{k-m} + v_k = \sum_{m=-\infty}^k h_{k-m} u_m + v_k, \quad k = \dots, -1, 0, 1, 2, \dots$$

$$Y(z) = H(z)U(z) + V(z) \quad (101.15)$$

where the sequence  $\{v_k\}$  represents some external input of errors and disturbances and with  $Y(z) = \mathcal{L}\{y\}$ ,  $U(z) = \mathcal{L}\{u\}$ ,  $V(z) = \mathcal{L}\{v\}$  as output and inputs. The *weighting function*  $h(kT) = \{h_k\}_{k=0}^{\infty}$ , which is zero for negative  $k$  and for reasons of causality is sometimes called *pulse response* of the digital system (compare *impulse response* of continuous-time systems). The pulse response and its  $z$  transform, the *pulse transfer function*,

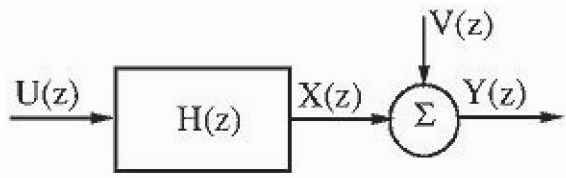
$$H(z) = \mathcal{L}\{h(kT)\} = \sum_{k=0}^{\infty} h_k z^{-k} \quad (101.16)$$

determine the system's response to an input  $U(z)$ ; see Fig. 101.3. The pulse transfer function  $H(z)$  is obtained as the ratio

$$H(z) = \frac{X(z)}{U(z)} \quad (101.17)$$

and provides the frequency domain input-output relation of the system. In particular, the Bode diagram is evaluated as  $|H(z)|$  and  $\arg H(z)$  for  $z = \exp(i\omega_k T)$  and for  $|\omega_k| < \omega_n = \pi/T$  — that is, when  $H(z)$  is evaluated for frequency points up to the Nyquist frequency  $\omega_n$  along the unit circle.

**Figure 101.3** Block diagram with an assumed transfer function relationship  $H(z)$  between input  $U(z)$ , disturbance  $V(z)$ , intermediate  $X(z)$ , and output  $Y(z)$ .



## State-Space Systems

Alternatives to the input-output representations by means of transfer functions are the state-space representations. Consider the following finite dimensional discrete state-space equation with a state vector  $x_k \in \mathcal{R}^n$ , input  $u_k \in \mathcal{R}^p$ , and observations  $y_n \in \mathcal{R}^m$ .

$$\begin{cases} x_{k+1} = \Phi x_k + \Gamma u_k \\ y_k = C x_k + D u_k \end{cases}, \quad k = 0, 1, \dots \quad (101.18)$$

with the pulse transfer function

$$H(z) = C(zI - \Phi)^{-1}\Gamma + D \quad (101.19)$$

and the output variable

$$Y(z) = C \sum_{k=0}^{\infty} \Phi^k z^{-k} x_0 + H(z)U(z) \quad (101.20)$$

where possible effects of initial conditions  $x_0$  appear as the first term. Notice that the initial conditions  $x_0$  can be viewed as the net effects of the input in the time interval  $(-\infty, 0)$ .

## 101.4 Digital Systems Described by Difference Equations(ARMAX Models)

---

An important class of nonstationary stochastic processes is one in which some deterministic response to an external input and a stationary stochastic process are superimposed. This is relevant, for instance, when the external input cannot be effectively described by some probabilistic distribution. A discrete-time model can be formulated in the form of a difference equation with an external input  $\{u_k\}$  that is usually considered to be known:

$$\begin{aligned} y_k = & -a_1 y_{k-1} - \cdots - a_n y_{k-n} \\ & + b_1 u_{k-1} + \cdots + b_n u_{k-n} \\ & + w_k + c_1 w_{k-1} + \cdots + c_n w_{k-n} \end{aligned} \quad (101.21)$$

Application of the  $z$  transform permits formulation of Eq. (101.21) as

$$A(z^{-1})Y(z) = B(z^{-1})U(z) + C(z^{-1})W(z) \quad (101.22)$$

where

$$\begin{aligned} A(z^{-1}) &= 1 + a_1 z^{-1} + \cdots + a_n z^{-n} \\ B(z^{-1}) &= 1 + b_1 z^{-1} + \cdots + b_n z^{-n} \\ C(z^{-1}) &= 1 + c_1 z^{-1} + \cdots + c_n z^{-n} \end{aligned} \quad (101.23)$$

Stochastic models including the  $A$  polynomial, according to Eq. (101.22) and Eq. (101.23), are known as **autoregressive models (AR)** and models including the  $C$  polynomial are known as **moving-average models (MA)**, whereas the  $B$  polynomial determines the effects of the external input ( $X$ ). Notice that the term *moving average* is here somewhat misleading, as there is no restriction that the coefficients should add to 1 or that the coefficients are nonnegative. An alternative description is *finite impulse response* or *all-zero filter*.

Thus, the full model of Eq. (101.22) is an **autoregressive moving average model** with external input (ARMAX) and its pulse transfer function  $H(z) = B(z^{-1})/A(z^{-1})$  is stable if and only if the *poles*<sup>34</sup> that is, the complex numbers  $z_1, \dots, z_n$  solving the equation  $A(z^{-1}) = 0$  — are strictly inside the unit circle, that is,  $|z_i| < 1$ . The *zeros* of the system—that is, the complex numbers  $z_1, \dots, z_n$  solving the equation  $B(z^{-1}) = 0$  — may take on any value without any instability arising, although it is preferable to obtain zeros located strictly inside the unit circle, that is,  $|z_i| < 1$  (*minimum-phase zeros*). By linearity  $\{y_k\}$  can be separated into one purely deterministic process  $\{x_k\}$  and one purely stochastic process  $\{v_k\}$ :

$$\begin{cases} A(z^{-1})X(z) = B(z^{-1})U(z) \\ A(z^{-1})V(z) = C(z^{-1})W(z) \end{cases} \quad \text{and} \quad \begin{cases} y_k = x_k + v_k \\ Y(z) = X(z) + V(z) \end{cases} \quad (101.24)$$

The type of decomposition [Eq. (101.24)] that separates the deterministic and stochastic processes is known as the *Wold decomposition*.

## 101.5 Prediction and Reconstruction

---

Consider the problem of predicting the output  $d$  steps ahead when the output  $\{y_k\}$  is generated by the ARMA model,

$$A(z^{-1})Y(z) = C(z^{-1})W(z) \quad (101.25)$$

which is driven by a zero-mean white noise  $\{w_k\}$  with covariance  $\mathcal{E}\{w_i w_j\} = \sigma_w^2 \delta_{ij}$ . In other words, assuming that observations  $\{y_k\}$  are available up to the present time, how should the output  $d$  steps ahead be predicted optimally? Assume that the polynomials  $A(z^{-1})$  and  $C(z^{-1})$  are mutually prime with no zeros for  $|z| \geq 1$ . Let the  $C$  polynomial be expanded according to the *Diophantine equation*,

$$C(z^{-1}) = A(z^{-1})F(z^{-1}) + z^{-d}G(z^{-1}) \quad (101.26)$$

which is solved by the two polynomials

$$\begin{aligned} F(z^{-1}) &= 1 + f_1 z^{-1} + \cdots + f_{n_F} z^{-n_F}, & n_F &= d - 1 \\ G(z^{-1}) &= g_0 + g_1 z^{-1} + \cdots + g_{n_G} z^{-n_G}, & n_G &= \max(n_A - 1, n_C - d) \end{aligned} \quad (101.27)$$

Interpretation of  $z^{-1}$  as a *backward shift operator* and application of Eqs. (101.25) and (101.26) permit the formulation

$$y_{k+d} = F(z^{-1})w_{k+d} + \frac{G(z^{-1})}{C(z^{-1})}y_k \quad (101.28)$$

Let us, by  $\hat{y}_{k+d|k}$ , denote linear  $d$ -step predictors of  $y_{k+d}$  based upon the measured information available at time  $k$ . As the zero-mean term  $F(z^{-1})w_{k+d}$  of Eq. (101.28) is unpredictable at time  $k$ , it is natural to suggest the following  $d$ -step predictor

$$\hat{y}_{k+d|k} = \frac{G(z^{-1})}{C(z^{-1})}y_k \quad (101.29)$$

The prediction error satisfies

$$\begin{aligned} \varepsilon_{k+d} &= (\hat{y}_{k+d|k} - y_{k+d}) \\ &= \frac{G(z^{-1})}{C(z^{-1})}y_k - \frac{A(z^{-1})F(z^{-1}) + z^{-d}G(z^{-1})}{C(z^{-1})}y_{k+d} \\ &= -F(z^{-1})w_{k+d} \end{aligned} \quad (101.30)$$



Let  $\mathcal{E}\{\cdot \mid \mathcal{F}_k\}$  denote the *conditional mathematical expectation* relative to the measured information available at time  $k$ . The conditional mathematical expectation and the covariance of the  $d$ -step prediction relative to available information at time  $k$  is

$$\begin{aligned}\mathcal{E}\{\hat{y}_{k+d|k} - y_{k+d} \mid \mathcal{F}_k\} &= \mathcal{E}\{-F(z^{-1})w_{k+d} \mid \mathcal{F}_k\} = 0 \\ \mathcal{E}\{(\hat{y}_{k+d|k} - y_{k+d})^2 \mid \mathcal{F}_k\} &= \mathcal{E}\{[F(z^{-1})w_{k+d}]^2 \mid \mathcal{F}_k\} \\ &= \mathcal{E}\{(w_{k+d} + f_1 w_{k+d-1} + \cdots + f_{d-1} w_{k+1})^2 \mid \mathcal{F}_k\} \\ &= (1 + f_1^2 + \cdots + f_{n_F}^2) \sigma_w^2 = 0\end{aligned}\tag{101.31}$$

It follows that the predictor of Eq. (101.29) is unbiased and that the prediction error only depends on future, unpredictable noise components. It is straightforward to show that the predictor of Eq. (101.29) achieves the lower bound of Eq. (101.31) and that the predictor of Eq. (101.29) is optimal in the sense that the prediction error variance is minimized.

**Example 101.13** **An Optimal Predictor for a First-Order Model.** Consider for the first-order ARMA model

$$y_{k+1} = -a_1 y_k + w_{k+1} + c_1 w_k \tag{101.32}$$

The variance of a one-step-ahead predictor  $\hat{y}_{k+1|k}$  is

$$\begin{aligned}\mathcal{E}\{(\hat{y}_{k+1|k} - y_{k+1})^2 \mid \mathcal{F}_k\} &= \mathcal{E}\{(\hat{y}_{k+1|k} + a_1 y_k - c_1 w_k)^2 \mid \mathcal{F}_k\} \\ &\quad + \mathcal{E}\{w_{k+1}^2 \mid \mathcal{F}_k\} \\ &= \mathcal{E}\{(\hat{y}_{k+1|k} + a_1 y_k - c_1 w_k)^2 \mid \mathcal{F}_k\} + \sigma_w^2 \geq \sigma_w^2\end{aligned}\tag{101.33}$$

The optimal predictor satisfying the lower bound in Eq. (101.33) is obtained from Eq. (101.33) as

$$\hat{y}_{k+1|k}^o = -a_1 y_k + c_1 w_k \tag{101.34}$$

which, unfortunately, is not realizable as it stands because  $w_k$  is not available to measurement. Therefore, the noise sequence  $\{w_k\}$  has to be substituted by some function of the observed variable  $\{y_k\}$ . A linear predictor chosen according to Eq. (101.29) is

$$\hat{y}_{k+1|k} = \frac{G(z^{-1})}{C(z^{-1})} y_k = \frac{c_1 - a_1}{1 + c_1 z^{-1}} y_k \tag{101.35}$$

## 101.6 The Kalman Filter

---

Consider the linear state-space model

$$\begin{aligned} x_{k+1} &= \Phi x_k + v_k, & x_k &\in \mathcal{R}^n \\ y_k &= C x_k + w_k, & y_k &\in \mathcal{R}^m \end{aligned} \quad (101.36)$$

where  $\{v_k\}$  and  $\{w_k\}$  are assumed to be independent zero-mean white-noise processes with covariances  $\Sigma_v$  and  $\Sigma_w$ , respectively. It is assumed that  $\{y_k\}$  but not  $\{x_k\}$  is available to measurement and that it is desirable to predict  $\{x_k\}$  from measurements of  $\{y_k\}$ .

Introduce the state predictor,

$$\begin{aligned} \hat{x}_{k+1|k} &= \Phi \hat{x}_{k|k-1} - K_k (\hat{y}_k - y_k), & \hat{x}_{k|k-1} &\in \mathcal{R}^n \\ \hat{y}_k &= C \hat{x}_{k|k-1}, & y_k &\in \mathcal{R}^m \end{aligned} \quad (101.37)$$

The predictor of Eq. (101.37) has the same dynamics matrix  $\Phi$  as the state-space model of Eq. (101.36) and, in addition, there is a correction term  $K_k(\hat{y}_k - y_k)$  with a factor  $K_k$  to be chosen. The prediction error is

$$\tilde{x}_{k+1|k} = \hat{x}_{k+1|k} - x_{k+1} \quad (101.38)$$

The prediction-error dynamics is

$$\tilde{x}_{k+1} = (\Phi - K_k C) \tilde{x}_k + v_k - K_k w_k \quad (101.39)$$

The mean prediction error is governed by the recursive equation

$$\mathcal{E}\{\tilde{x}_{k+1}\} = (\Phi - K_k C) \mathcal{E}\{\tilde{x}_k\} \quad (101.40)$$

and the mean square error of the prediction error is governed by

$$\begin{aligned} \mathcal{E}\{\tilde{x}_{k+1} \tilde{x}_{k+1}^T\} &= \mathcal{E}\{[(\Phi - K_k C) \tilde{x}_k + v_k - K_k w_k][(\Phi - K_k C) \tilde{x}_k + v_k - K_k w_k]^T\} \\ &= (\Phi - K_k C) \mathcal{E}\{\tilde{x}_k \tilde{x}_k^T\} (\Phi - K_k C)^T + \Sigma_v + K_k \Sigma_w K_k \end{aligned} \quad (101.41)$$

If we denote

$$\begin{aligned} P_k &= \mathcal{E}\{\tilde{x}_k \tilde{x}_k^T\} \\ Q_k &= \Sigma_v + C P_k C^T \end{aligned} \quad (101.42)$$

then Eq. (101.41) is simplified to

$$P_{k+1} = \Phi P_k \Phi^T - K_k C P_k \Phi - \Phi^T P_k C^T K_k^T + \Sigma_v + K_k Q_k K_k^T \quad (101.43)$$

By completing squares of terms containing  $K_k$  we find

$$\begin{aligned} P_{k+1} &= \Phi P_k \Phi^T + \Sigma_v - \Phi P_k C^T Q_k^{-1} C P_k \Phi^T \\ &\quad + (K_k - \Phi P_k C^T Q_k^{-1}) Q_k (K_k - \Phi P_k C^T Q_k^{-1})^T \end{aligned} \quad (101.44)$$

where only the last term depends on  $K_k$ . Minimization of  $P_{k+1}$  can be done by choosing  $K_k$  such that the positive semidefinite  $K_k$ -dependent term in Eq. (101.44) disappears. Thus  $P_{k+1}$  achieves its lower bound for

$$K_k = \Phi P_k C^T (\Sigma_w + C P_k C^T)^{-1} \quad (101.45)$$

and the *Kalman filter* (or *Kalman-Bucy filter*) takes the form

$$\begin{aligned} \hat{x}_{k+1|k} &= \Phi \hat{x}_{k|k-1} - K_k (\hat{y}_k - y_k) \\ \hat{y}_k &= C \hat{x}_{k|k-1} \\ K_k &= \Phi P_k C^T (\Sigma_w + C P_k C^T)^{-1} \\ P_{k+1} &= \Phi P_k \Phi^T + \Sigma_v - \Phi P_k C^T (\Sigma_w + C P_k C^T)^{-1} C P_k \Phi^T \end{aligned} \quad (101.46)$$

which is the optimal predictor in the sense that the mean square error [Eq. (101.41)] is minimized in each step.

**Example 101.234 Kalman Filter for a First-Order System.** Consider the state-space model

$$\begin{aligned} x_{k+1} &= 0.95x_k + v_k \\ y_k &= x_k + w_k \end{aligned} \quad (101.47)$$

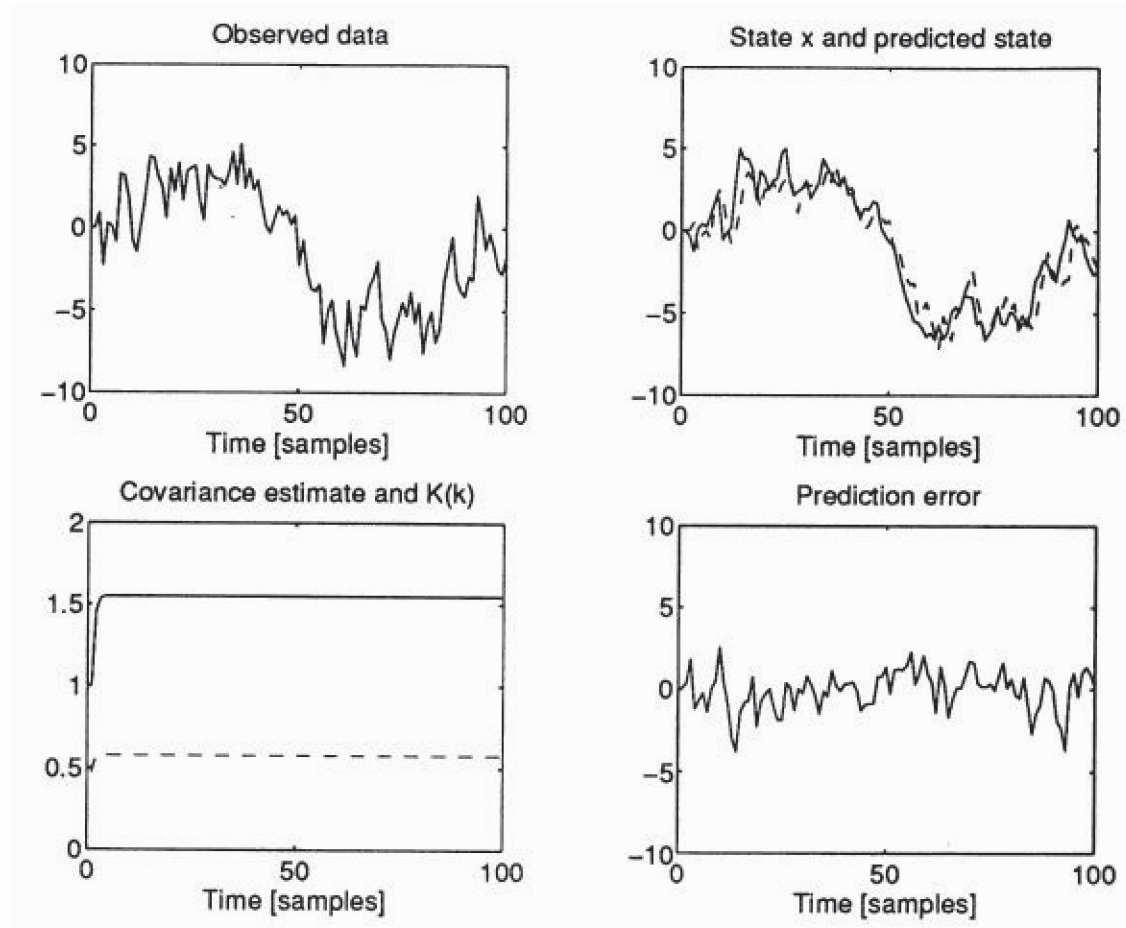
where  $\{v_k\}$  and  $\{w_k\}$  are zero-mean white-noise processes with covariances  $\mathcal{E}\{v_k^2\} = 1$  and  $\mathcal{E}\{w_k^2\} = 1$ , respectively.

The Kalman filter takes on the form

$$\begin{aligned} \hat{x}_{k+1|k} &= 0.95\hat{x}_{k|k-1} - K_k (\hat{x}_{k|k-1} - y_k) \\ K_k &= \frac{0.95P_k}{1 + P_k} \\ P_{k+1} &= 0.95^2 P_k + 1 - \frac{0.95^2 P_k^2}{1 + P_k} \end{aligned} \quad (101.48)$$

The result of one such realization is shown in Fig. 101.4.

**Figure 101.4** Kalman filter applied to one-step-ahead prediction of  $x_{k+1}$  in Eq. (101.47). The observed variable  $\{y_k\}$ , the state  $\{x_k\}$ , and the predicted state  $\{\hat{x}_k\}$ , the estimated variance  $\{P_k\}$  and  $\{K_k\}$ , and the prediction error  $\{\tilde{x}_k\}$  are shown in a 100-step realization of the stochastic process. (Source: Johansson, R. 1993. *System Modeling and Identification*. Prentice Hall, Englewood Cliffs, NJ.)



## Defining Terms

**Autoregressive model (AR):** An autoregressive time series of order  $n$  is defined via  $y_k = -\sum_{m=1}^n a_m y_{k-m} + w_k$ . The sequence  $\{w_k\}$  is usually assumed to consist of zero-mean identically distributed stochastic variables  $w_k$ .

**Autoregressive moving average model (ARMA):** An autoregressive moving average time series of order  $n$  is defined via  $y_k = -\sum_{m=1}^n a_m y_{k-m} + \sum_{m=0}^n c_m w_{k-m}$ . The sequence  $\{w_k\}$  is usually assumed to consist of zero-mean identically distributed stochastic variables  $w_k$ .

**Discrete Laplace transform:** The discrete Laplace transform is counterpart to the Laplace transform with application to discrete signals and systems. The discrete Laplace transform is obtained from the  $z$  transform by means of the substitution  $z = \exp(sT)$ , where  $T$  is the

sampling period.

**Moving average process (MA):** A moving average time series of order  $n$  is defined via  $y_k = \sum_{m=0}^n c_m w_{k-m}$ . The sequence  $\{w_k\}$  is usually assumed to consist of zero-mean identically distributed stochastic variables  $w_k$ .

**Rational model:** AR, MA, ARMA, and ARMAX are commonly referred to as rational models.

**Time series:** A sequence of random variable  $\{y_k\}$ , where  $k$  belongs to the set of positive and negative integers.

**$z$  transform:** A generating function applied to sequences of data and evaluated as a function of the complex variable  $z$  with interpretation of frequency.

## References

- Box, G. E. P. and Jenkins, G. M. 1970. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, CA.
- Hurewicz, W. 1947. Filters and servo systems with pulsed data. In *Theory of Servomechanisms*, ed. H. M. James, N. B. Nichols, and R. S. Phillips. McGraw-Hill, New York.
- Jenkins, G. M. and Watts, D. G. 1968. *Spectral Analysis and Its Applications*. Holden-Day, San Francisco, CA.
- Johansson, R. 1993. *System Modeling and Identification*. Prentice Hall, Englewood Cliffs, NJ.
- Jury, E. I. 1956. Synthesis and critical study of sampled-data control systems. *AIEE Trans.* 75: 141–151.
- Kalman, R. E. and Bertram, J. E. 1958. General synthesis procedure for computer control of single and multi-loop linear systems. *Trans. AIEE.* 77: 602–609.
- Kolmogorov, A. N. 1939. Sur l'interpolation et extrapolation des suites stationnaires. *C.R. Acad. Sci.* 208:2043–2045.
- Ragazzini, J. R. and Zadeh, L. A. 1952. The analysis of sampled-data systems. *AIEE Trans.* 71: 225–234.
- Tsytkin, Y. Z. 1950. Theory of discontinuous control. *Avtomatika i Telemekhanika*. Vol. 5.
- Wiener, N. 1949. *Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*. John Wiley & Sons, New York.

## Further Information

Early theoretical efforts developed in connection with servomechanisms and radar applications [Hurewicz, 1947]. Tsytkin [1950] introduced the discrete Laplace transform and the formal  $z$  transform definition was introduced by Ragazzini and Zadeh [1952] with further developments by Jury [1956]. Much of prediction theory was originally developed by Kolmogorov [1939] and Wiener [1949] whereas state-space methods were forwarded by Kalman and Bertram [1958]. Pioneering textbooks on time-series analysis and spectrum analysis are provided by Box and Jenkins [1970] and Jenkins and Watts [1968].

Detailed accounts of time-series analysis and the  $z$  transform and their application to signal processing are to be found in

- Oppenheim, A. V. and Schaffer, R. W. 1989. *Discrete-Time Signal Processing*. Prentice Hall,

Englewood Cliffs, NJ.

- Proakis, J. G. and Manolakis, D. G. 1989. *Introduction to Digital Signal Processing*. Maxwell MacMillan Int. Ed., New York.

Theory of time-series analysis and its application to discrete-time control is to be found in

- Aström, K. J. and Wittenmark, B. 1990. *Computer-Controlled Systems*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.

Theory of time-series analysis and methodology for determination and validation of discrete-time models and other aspects of system identification are to be found in

- Johansson, R. 1993. *System Modeling and Identification*. Prentice Hall, Englewood Cliffs, NJ.

Good sources to monitor current research are

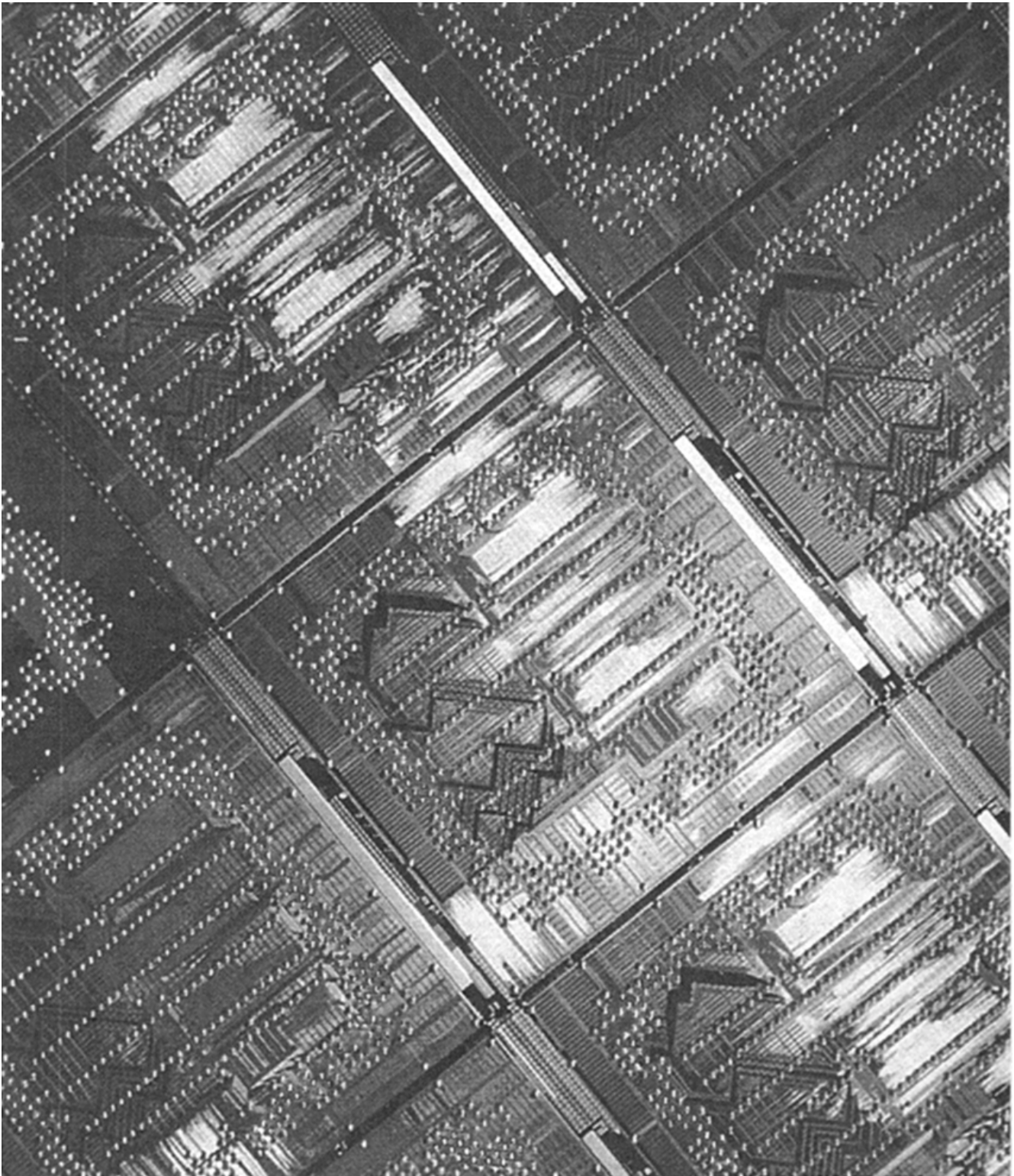
- *IEEE Transactions on Automatic Control*
- *IEEE Transactions on Signal Processing*

Examples of easy-to-read survey articles for signal processing applications are

- Cadzow, J. A. 1990. Signal processing via least-squares error modeling. *IEEE ASSP Magazine*. 7:12–31, October.
- Schroeder, M. R. 1984. Linear prediction, entropy, and signal analysis. *IEEE ASSP Magazine*. 1:3–11, July.

Wai-Kai Chen. "Circuits"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000





The PowerPC family of microprocessors is the product of an alliance between Apple, IBM and Motorola. Shown above is a 200-mm wafer section of the Power PC 601, which is the first in the Power PC family and is designed for a wide range of desktop computing applications. This wafer consists of thousands of integrated circuits built by layering several different materials in a well-defined fashion on a silicon base, thereby bringing workstation performance to desktop computers. (Photo by Tom Way. Courtesy of IBM.)



# XVII

## Circuits

---

**Wai-Kai Chen**

*University of Illinois, Chicago*

**102 Passive Components** *H. Domingos*

Resistors • Capacitors • Inductors

**103 RL, RC, and RLC Circuits** *M. D. Ciletti*

RL Circuits • RC Circuits • RLC Circuits

**104 Node Equations and Mesh Equations** *J. A. Svoboda*

Node Equations • Mesh Equations

**105 Sinusoidal Excitation and Phasors** *M. H. Rashid*

Sinusoidal Source • Phasor • Passive Circuit Elements in Phasors Domain

**106 Three-Phase Circuits** *N. Balabanian*

Relationships between Voltages and Currents • Line Voltages • Power Relationship • Balanced Source and Balanced Load • Other Types of Interconnections

**107 Filters (Passive)** *A. J. Rosa*

Fundamentals • Applications

**108 Power Distribution** *R. Broadwater, A. Sargent, and R. E. Lee*

Equipment • System Divisions and Types • Electrical Analysis, Planning, and Design • System Control • Operations

**109 Grounding and Shielding** *W. G. Duff*

Grounding • Shielding

CIRCUITS ARE FUNDAMENTAL to electrical engineering. The circuit problem deals with predicting the behavior of a system of interconnected physical elements such as resistors, capacitors, inductors, and independent and dependent sources. Kirchhoff made the first comprehensive investigation of the circuit problem in 1847 and provided a solution to a resistive circuit. Maxwell pointed out in 1892 that Kirchhoff's formulation omitted the concept of potential, and he devised two very effective methods for solving the circuit problem, now termed Maxwell's *mesh* and *nodal* methods. They are still widely used today.

The most important periodic signals, as far as electrical engineers are concerned, are the familiar sine and cosine functions known as the *sinusoids*. They are not only found throughout nature, as in the motion of a pendulum and in the vibrations of strings, but they also can approximate a generator voltage or current. In fact, any periodic signal can be represented as a series of sinusoids, the individual frequencies of which are multiples of the frequency known as the *fundamental frequency*.

Circuits are designed to perform certain functions. The most common functions are to separate, pass, or suppress a group of signals from a mixture of signals. A device with such a function is

referred to as a *filter*. On a larger scale, televisions and radios are typical examples of electrical filters. When a television is tuned to a particular channel, it will only pass those signals transmitted by that channel and will block all other signals. On a smaller scale, filters are basic electronic components used in the design of communication systems such as telephone, television, radio, radar, and computer. In fact, electrical filters permeate modern technology so much that it is difficult to find an electronic system that does not employ a filter in one form or another.

Electrical safety is a very broad and diverse topic. Everyone has experienced some form of electrical shock. Whether that shock was from a harmless electrostatic discharge or from accidental contact with an energized electrical circuit, the response was probably the same. For safe and sustained operation of an electrical device, grounding and shielding are two very important factors that must be considered in its design. The primary purposes of grounding and shielding circuits are to prevent shock hazard, to protect circuits, and to reduce electromagnetic interference.

A traditional problem in electrical engineering is to deliver electrical energy to consumers in a manner that is economic, reliable, and safe. In addition, it must conform to regulatory standards. An important use of electricity is to drive electric motors in the AC steady state. To reduce fluctuation, power is delivered to the load using what is known as a *three-phase* system. A single fault does not result in an interruption of power to the customers. A residential customer is supplied from a two-wire, one-phase, overhead lateral, whereas the commercial and industrial customers are supplied from the three-phase, four-wire overhead primary feeder.

In this section, we introduce the basics of these topics. No engineering graduate is equipped to understand the literature or advance practice without a knowledge of these fundamentals.

Domingos, H. "Passive Components"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

[102.1 Resistors](#)[102.2 Capacitors](#)[102.3 Inductors](#)**Henry Domingos***Clarkson University*

Passive components such as resistors, capacitors, and inductors can be an integral part of a more complex unit such as an operational amplifier, digital integrated circuit, or microwave circuit. However, this chapter deals with discrete components purchased as individual parts. Their importance lies in the fact that virtually every piece of electronic equipment incorporates a multitude of discrete passive components, and sale of these components parallels the fortunes of the electronics industry as a whole.

Although passive components have been in use for hundreds of years, one should not assume they represent a static industry: There is a continuing evolution in the design and fabrication of these parts to make them smaller and cheaper with better performance and better reliability. In fact, the useful life of modern components exceeds by far that of the electronic equipment they are used in. Failure is more often the result of misapplication. The continuous improvement in quality, price, performance, and general usefulness contributes to greater convenience and functionality for consumers at a lower and lower price.

---

## 102.1 Resistors

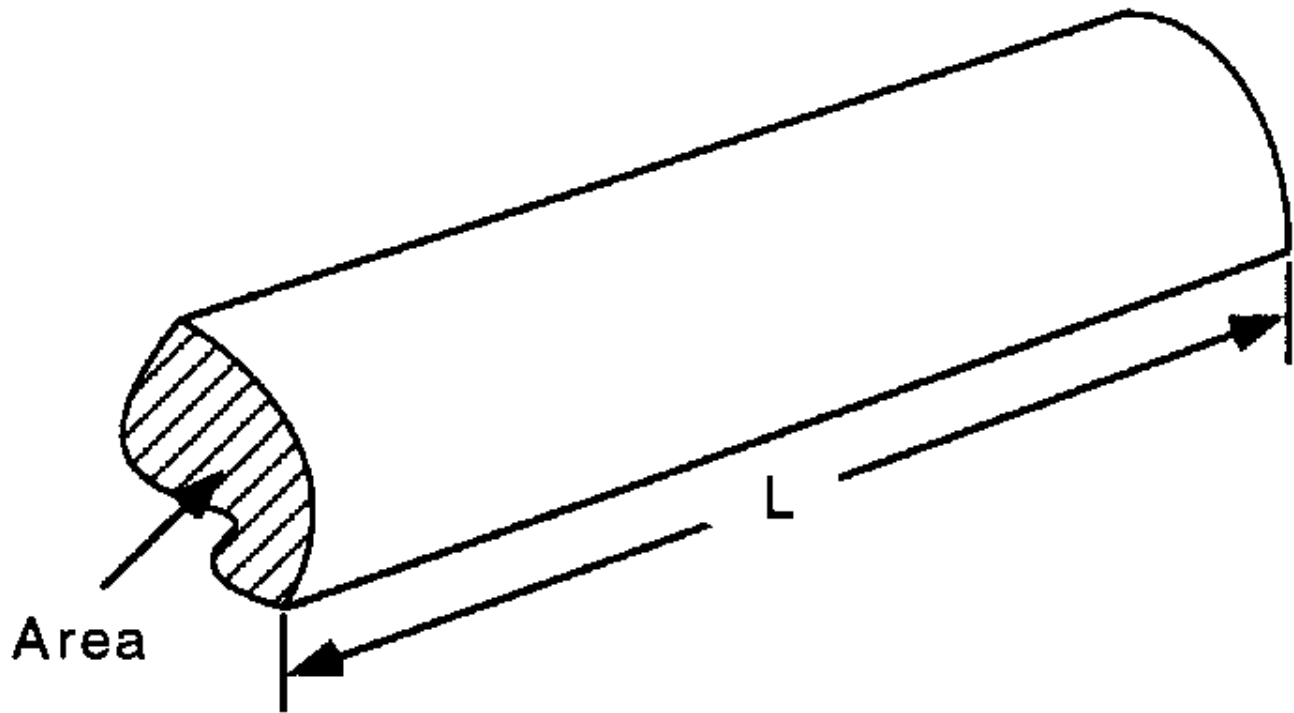
---

Resistance is a property of a device that relates the current through the device with the voltage or electric potential across it. It is a dissipative property in that energy is converted from electrical energy to heat. The resistance of a device is directly related to the **resistivity** of the material of which it is composed. The resistivity is a relative property that varies from that of good conductors such as aluminum or copper (resistivity  $\approx 10^{-6}$  ohm-cm) to semiconductors such as silicon ( $\approx 10^{-3} - 10^{+3}$  ohm-cm) to that of good insulators such as alumina ( $\approx 10^{14}$  ohm-cm). The resistance of a device is determined by its resistivity and its geometry. [Figure 102.1](#) and Eq. (102.1) show the relationship for objects of uniform resistivity and cross-sectional area.

$$R = \rho \frac{L}{A} \quad (102.1)$$

In Eq. (102.1)  $R$  is the resistance in ohms,  $\rho$  is the resistivity in ohm-cm,  $L$  is the length in cm, and  $A$  is the cross-sectional area in  $\text{cm}^2$ . For example, #12 AWG copper wire has a diameter of 0.0808 in. and a resistivity of  $1.724 \times 10^{-6}$  ohm-cm. Using Eq. (102.1) and converting units as appropriate, the resistance of 1000 feet of wire is 1.588 ohms.

**Figure 102.1** Resistance of a conductor with uniform resistivity and cross section is given by Eq. (102.1).



When resistive films are deposited on substrates or impurities are implanted in layers in silicon devices, the resistivity is not uniform and/or the thickness of the layer may not be known accurately. It is convenient then to define a **sheet resistance** that takes into account the average resistivity and the thickness so that the resistance is given by

$$R = R_s \frac{L}{W} \quad (102.2)$$

where  $R_s$  is the sheet resistance in ohms per square (the "square" is dimensionless),  $L$  is the length of the film, and  $W$  its width. Suppose, for example, a resistor ink is silk-screened onto a ceramic substrate and fired so that its sheet resistance is 100 ohms per square. If the width of the resistor is 10 mils (a mil is one thousandth of an inch), the length required to make a 1500-ohm resistor is, from Eq. (102.2),  $L = 150$  mils.

Using the concepts of resistivity and sheet resistance, a discrete resistor can be fabricated in a variety of ways. The common types are carbon composition resistors, wire-wound resistors, and film resistors. Carbon composition resistors were at one time the most popular type. They are low

priced, high in reliability, and available in a wide range of resistance values. The drawbacks are poor long-term stability and tolerance of only 5% or larger. The resistor is fabricated from a silica-loaded resin with resistivity controlled by the addition of carbon particles. Axial leads are inserted and the outer shell, of the same material as the core, minus the carbon granules, insulates the assembly.

Wire-wound resistors are power wire wounds or precision wire wounds. In either case the resistance wire is either a nickel-chromium alloy with a resistivity of  $1.33 \cdot 10^{-4}$  ohm-cm for high resistance values, or a copper-nickel alloy with a resistivity of  $5 \cdot 10^{-5}$  ohm-cm for lower resistances. Power wire wounds are intended to operate at higher power dissipation and are composed of single layers of bare wire with substrates and packages of high-temperature materials. Precision resistors are stable, highly accurate components. The wire is multilayered, and therefore must be insulated, and all materials can be low temperature. The main disadvantages of wire-wound resistors are the high cost and low operating frequency.

Film-type resistors are made from various resistive materials deposited on an insulating substrate. Film thicknesses range from 0.005 micrometers deposited by evaporation for precision film resistors to as much as 100 micrometers deposited as a resistive ink. Sheet resistances range from 10 to 10 000 ohms per square. The final resistance value can be adjusted by cutting a spiral path through the resistive film until the desired value is reached. The resistors are stable, accurate, and low cost and are the most popular types today. Resistor networks, chip resistors, and film resistors deposited directly on a hybrid substrate provide a wide selection of styles.

The important specifications for a resistor are the resistance value, the tolerance, and the power rating. For carbon composition resistors the resistance value and tolerance are indicated by the familiar color code in [Table 102.1](#). Values range from 1 ohm to 100 megohms. Power rating ranges from 1/20 watt to 2 watts. The power rating is somewhat misleading since the temperature rise is the crucial variable, not the power dissipation. A 2 W resistor cannot dissipate 2 W if it is conformably coated and situated where cooling air is restricted. On the other hand, if a suitable heat sink is provided and forced air cooling is available, the rating could be higher than specified. Under pulse conditions the instantaneous power can be orders of magnitude larger than its nominal rating.

**Table 102.1 Resistor Color Code**

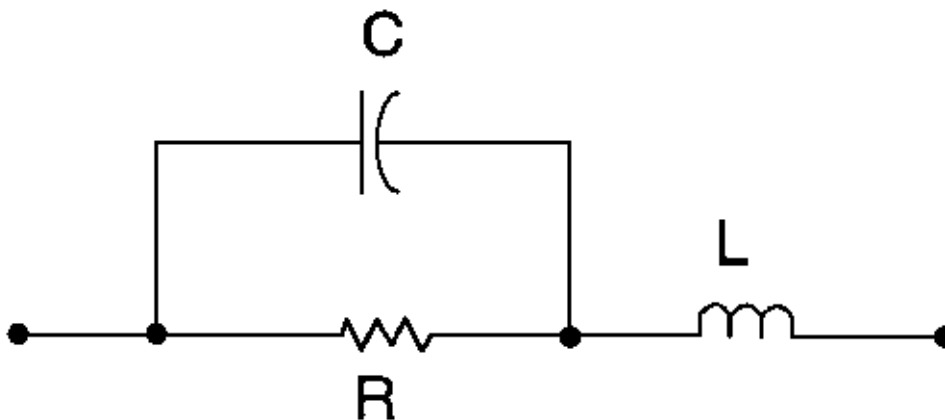
Color	First Band, <sup>a</sup> Significant Figure	Second Band, Significant Figure	Third Band, Multiplier	Fourth Band, <sup>b</sup> Tolerance (%)	Fifth Band, <sup>b</sup> Failure Rate (%/1000 h)
Black	0	0	1	—	—
Brown	1	1	10	—	1
Red	2	2	10 <sup>2</sup>	—	0.1
Orange	3	3	10 <sup>3</sup>	—	0.01
Yellow	4	4	10 <sup>4</sup>	—	0.001
Green	5	5	10 <sup>5</sup>	—	—
Blue	6	6	10 <sup>6</sup>	—	—
Violet	7	7	10 <sup>7</sup>	—	—
Gray	8	8	10 <sup>8</sup>	—	—
White	9	9	10 <sup>9</sup>	—	—
Silver	—	—	0.01	10	—
Gold	—	—	0.1	5	—
None	—	—	—	20	—

<sup>a</sup>The first band is the one closest to one end of the resistor. A first band wider than the others indicates a wire-wound resistor.

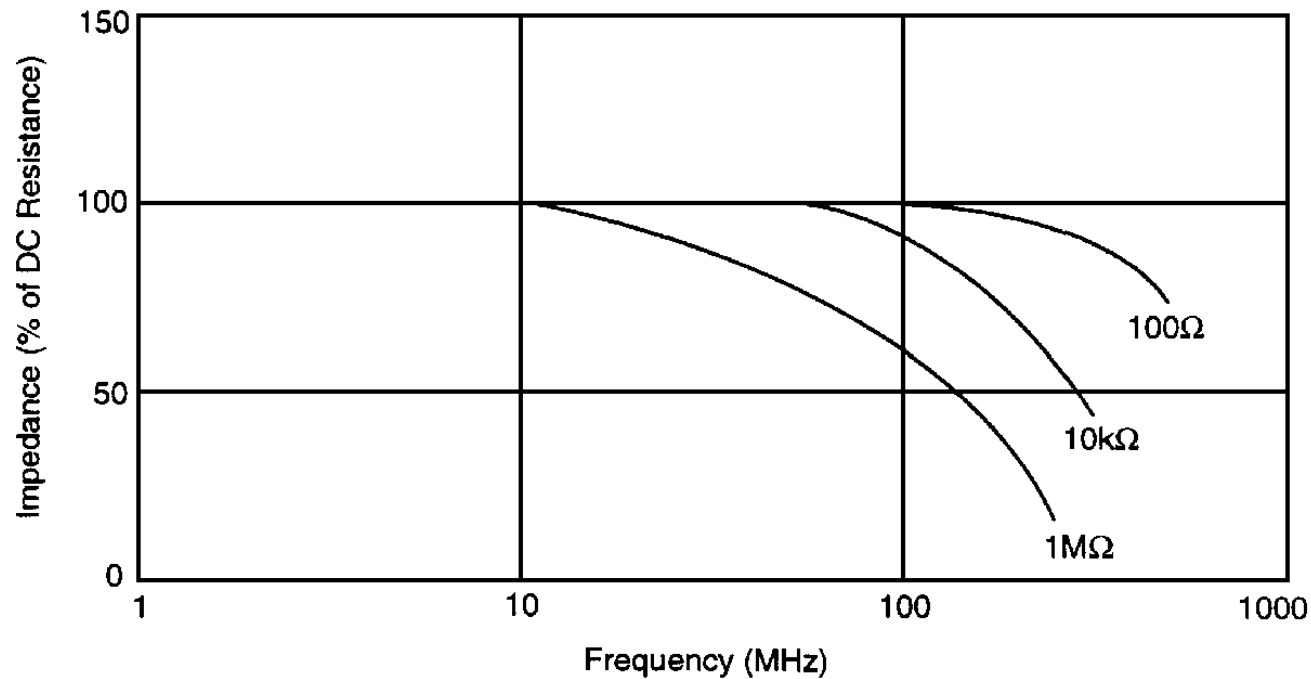
<sup>b</sup>Certain MIL parts.

No resistance is ideal, and over wide frequency ranges the parasitic inductance and capacitance limit the usefulness. [Figure 102.2](#) shows a common equivalent circuit for a resistor. The series inductance can be large for wire-wound resistors and limits their usefulness to frequencies less than about 50 to 100 kHz. For other resistors the leads constitute the main source of inductance, estimated to be of the order of 20 nH per inch. In general, low values of resistance and thin films tend to have lower inductance. Capacitance tends to dominate at high frequencies and with higher values of resistance. [Figure 102.3](#) illustrates typical variation of impedance of a resistor with frequency.

**Figure 102.2** At high frequencies the effects of parasitic capacitance and inductance of a resistor must be accounted for. This shows one possible equivalent representation.



**Figure 102.3** Variation of total impedance of a resistor with frequency for a typical film resistor.



Other resistor characteristics include temperature coefficient, voltage coefficient, stability, and noise. For a discussion of these the reader is referred to Dorf [1993] and Harper [1977].

## 102.2 Capacitors

**Capacitance** is a property that exists between two conductors separated by a **dielectric**. With equal but opposite charges on the two conductors a potential difference can be measured between them, and the capacitance is defined by the equation



$$C = \frac{Q}{V} \quad (102.3)$$

$C$  is the capacitance in farads,  $Q$  is the magnitude of the charge on either plate in coulombs, and  $V$  is the voltage between the plates in volts. The capacitance in the ideal case of two plane-parallel plates is given by

$$C = \varepsilon \frac{A}{d} \quad (102.4)$$

where  $\varepsilon$  is the **permittivity** of the dielectric in farads per meter,  $A$  is the area of each plate in meters squared, and  $d$  is the plate separation in meters.  $\varepsilon$  is equal to  $\varepsilon_r \varepsilon_0$ , where  $\varepsilon_r$  is the relative permittivity (sometimes referred to as the *dielectric constant*,  $k$ ) = 1 for vacuum,  $\approx 1$  for air, and ranging from 2 to 10 for common dielectrics.  $\varepsilon_0$  is the permittivity of free space, equal to  $8.854 \cdot 10^{-12}$  farads per meter. All **capacitors** use appropriate values of  $\varepsilon$ ,  $A$ , and  $d$  in Eq. (102.4) to achieve the wide range of values available in commercial capacitors.

The constitutive relationship between voltage and current in a capacitor is

$$v(t) = v_0 + \frac{1}{C} \int_0^t i \, dt \quad (102.5)$$

where  $v(t)$  is the instantaneous voltage across the capacitor,  $v_0$  is the initial voltage on the capacitor at  $t = 0$ , and  $i$  is the instantaneous current through the capacitor. By differentiating Eq. (102.5),

$$i(t) = C \frac{dv}{dt} \quad (102.6)$$

Unlike a resistor, an ideal capacitor does not dissipate energy. Instead, it accumulates energy during the charging process and releases energy to the electrical circuit as it discharges. The energy stored on a capacitor is given by

$$W = \frac{1}{2} CV^2 = \frac{1}{2} \frac{Q^2}{C} \quad (102.7)$$

where  $W$  is in joules.

There are two main classes of capacitors—electrolytic and electrostatic. Electrolytic capacitors include aluminum and tantalum types, used in applications where large capacitance values are needed. Electrostatic capacitors include plastic, ceramic disk, ceramic chip, mica, and glass. Aluminum electrolytics are constructed from high-purity aluminum foils that are chemically etched to increase the surface area, then anodized to form the dielectric. The thickness of this anodized layer determines the voltage rating of the capacitor. If only one foil is anodized, the capacitor is a polarized unit, and the instantaneous voltage cannot be allowed to reverse polarity.

Porous paper separates the two foils and is saturated with a liquid electrolyte; therefore, the unit must be sealed, and leakage is a common failure mode. During extended periods of storage the anodized layer may partially dissolve, requiring the unit to be reformed before rated voltage can be applied. Aluminum electrolytics exhibit a large series inductance, which limits the useful range of frequencies to about 20 kHz. They also have a large leakage current. Nevertheless, because of low cost and very large values of capacitance (up to 1 F), they are a popular choice for filtering applications.

Tantalum electrolytics are available in foil, wet slug, and solid varieties. All types require a tantalum electrode to be anodized, but the means used to make electrical contact to the plates differs in each case. The overall characteristics are superior to those of aluminum, but the cost is greater. Because of their small physical size and large capacitance they are used for the same applications as aluminum electrolytics.

The common type of plastic capacitor is the metalized film type, made by vacuum deposition of aluminum a fraction of a micron thick directly onto a plastic dielectric film. Common film types include polyester, polycarbonate, polypropylene, and polysulfone. Although the relative permittivity is low, the films can be made extremely thin, resulting in a large capacitance per unit volume. This type of capacitor exhibits an interesting self-healing feature. If breakdown should happen to occur—for instance, at a defect in the film or during a momentary electrical overstress—the metalization around the breakdown site will act like a fuse and vaporize, clearing the breakdown and restoring operation to an equal or higher voltage capability.

Plastic film capacitors have good electrical characteristics and are capable of operation at higher voltages and much higher frequencies than electrolytic capacitors. They have found wide application in filtering, bypassing, coupling, and noise suppression.

Ceramic capacitors employ ferroelectric dielectrics—commonly, barium titanate—that have an extremely high relative permittivity, 200 to 100 000. For disk capacitors the appropriate combination of powders is mixed, pressed into disks, and sintered at high temperatures. Electrodes are silk-screened on, leads are attached, and a protective coating applied. By varying the area, thickness (limited, to maintain mechanical strength), and especially the dielectric formulation, wide ranges of capacitance value are achieved. Very high voltage capability is also an option by increasing the thickness.

Ceramic chip capacitors are made from a slurry containing the dielectric powder and cast onto stainless-steel or plastic belts. After drying, electrodes are printed on the sheets, which are then arranged in a stack and fired. Electrodes are attached and the unit is encapsulated. By adjusting the number of plates and the plate area a wide range of values can be obtained. The extremely small size makes this type especially useful in hybrid integrated circuits and on printed circuit boards.

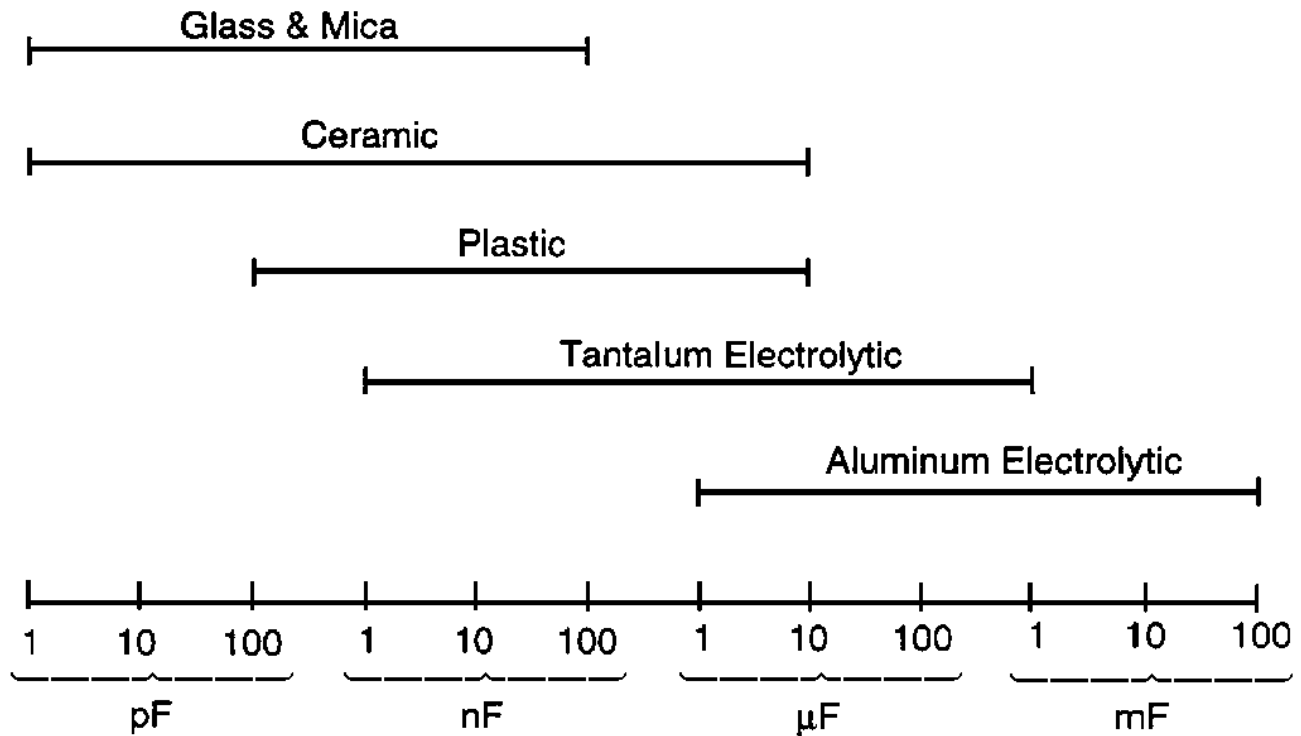
Mica and glass capacitors are used whenever high quality and excellent stability with respect to temperature and aging are required, and where high frequency operation is required. They are used in tuning circuits and for coupling where high performance and reliability are essential. Mica capacitors are made from sheets of mica that are alternately stacked with foils. Alternate foils are extended and folded over the ends. After leads are attached the unit is encapsulated. Glass capacitors are made from glass ribbons that are stacked alternately with foils. Leads are attached and the entire assembly is sealed in glass at high temperature and pressure.

Capacitors used in power system applications employ oil-impregnated paper as a dielectric.

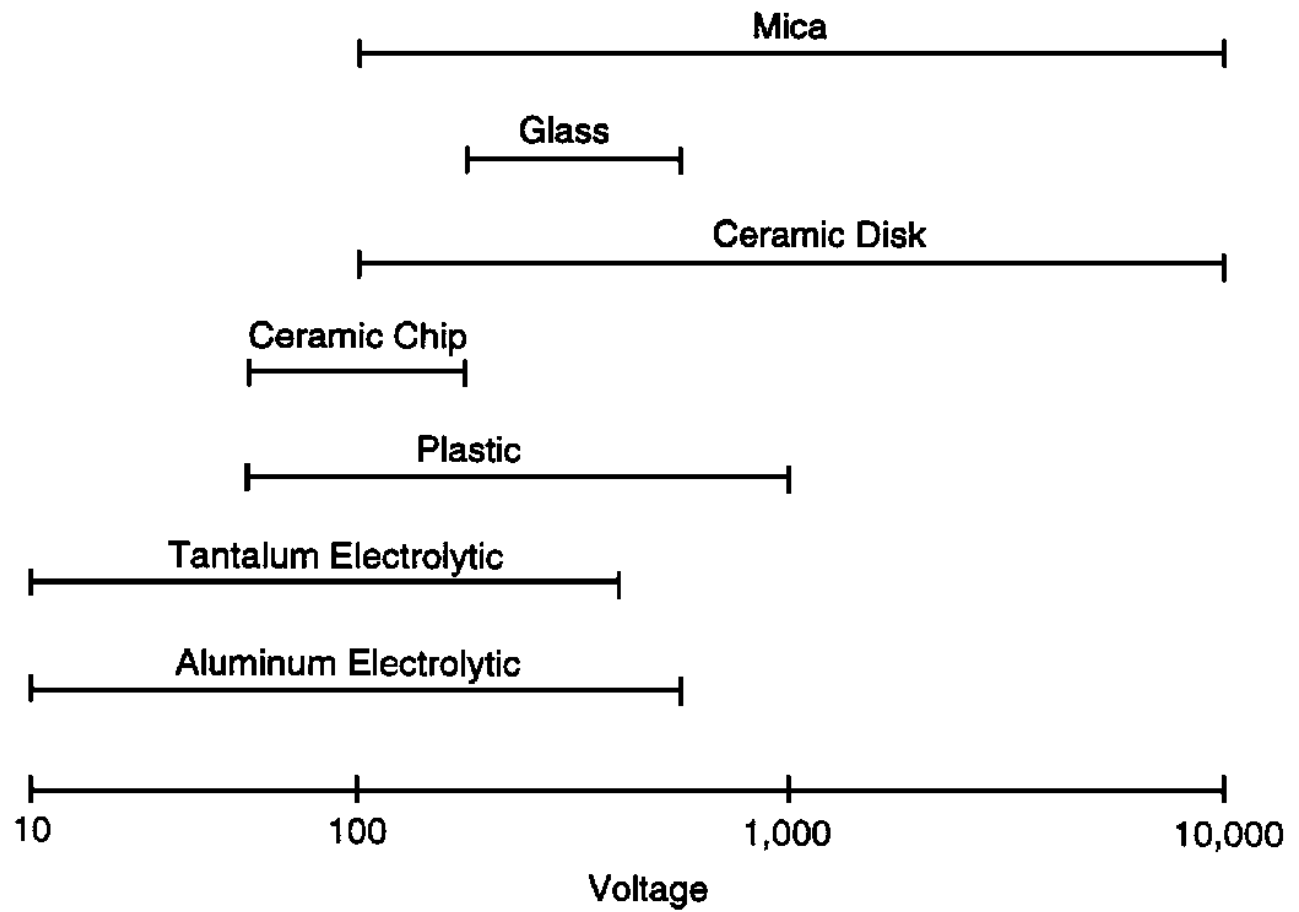
Because of the very high voltage ratings, these capacitors are shipped and stored with a shorting bar across the terminals as a safety precaution. Even though a capacitor has been discharged after high voltage operation or testing, it is possible for charge and voltage to re-accumulate by a phenomenon known as *dielectric absorption*.

Figures 102.4, 102.5, and 102.6 show ranges of capacitance values, voltage ratings, and frequency ranges for the different capacitors described in this section.

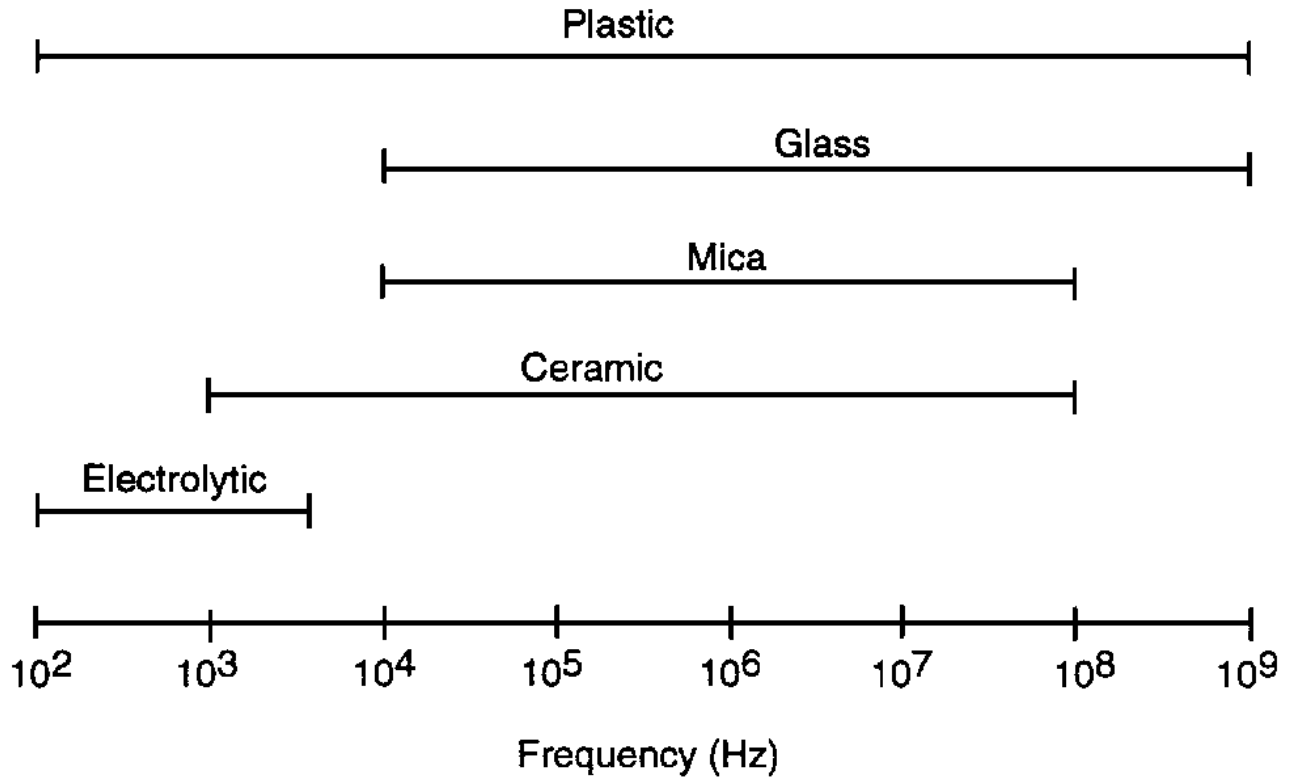
**Figure 102.4** Range of values for various types of capacitors.



**Figure 102.5** Voltage ratings of various capacitors.



**Figure 102.6** Frequency range of various capacitors.



All capacitors dissipate a small amount of power due to resistive losses in the conductors, leakage current across the dielectric, and losses in the dielectric under AC operation. If the capacitor is represented by its capacitance value in series with a resistor to simulate these losses, the resistance is called the **equivalent series resistance (ESR)**. Capacitance bridges often measure the **dissipation factor (DF)**, which is given by

$$DF = 2\pi f C_s R_s \quad (102.8)$$

where  $C_s$  is the series capacitance,  $f$  is the frequency in hertz, and  $R_s$  is the ESR. The reciprocal of the dissipation factor is the **quality factor (Q)** of the capacitor.

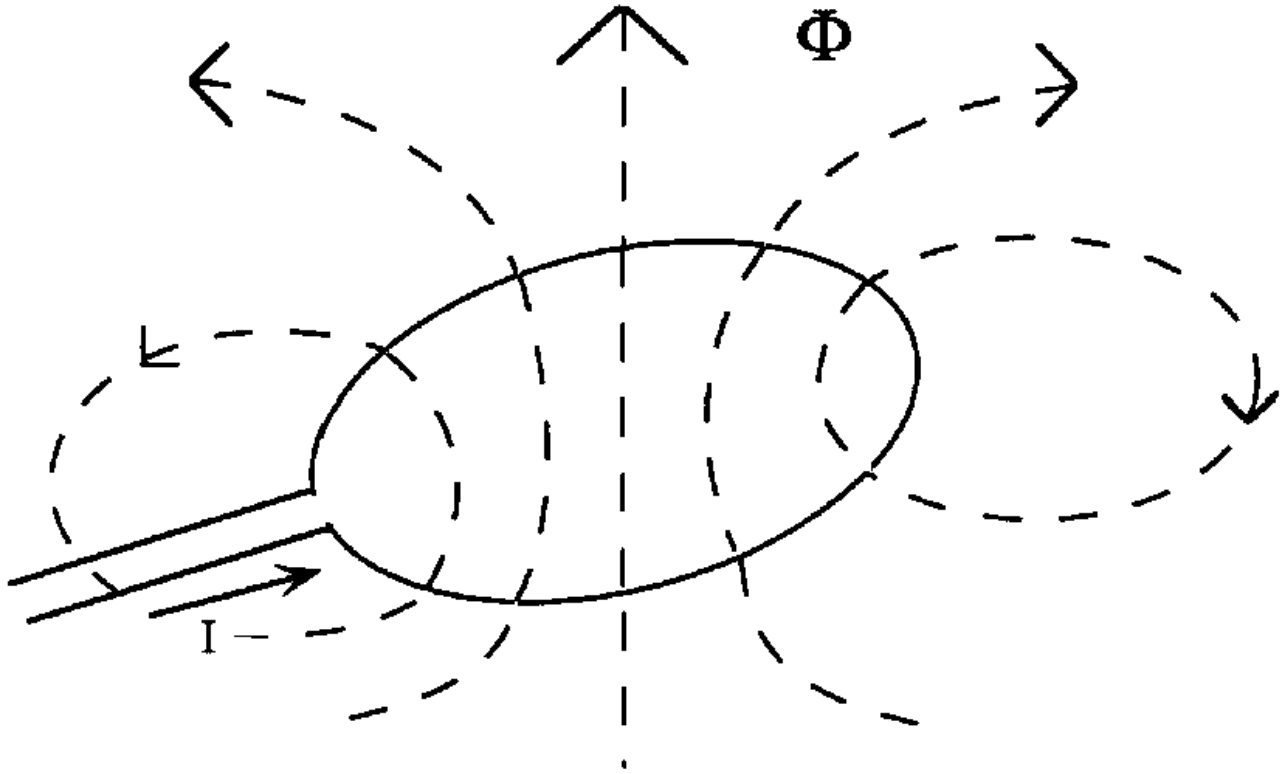
## 102.3 Inductors

**Inductance** is the property of a device that relates a magnetic field to the current that produces it. In Fig. 102.7, the current  $I$  flowing in a counterclockwise direction in the coil produces, by the right-hand rule, a magnetic flux. The inductance of the coil is defined by

$$L = \frac{\Phi}{I} \quad (102.9)$$

where  $L$  is the inductance in henrys,  $\Phi$  is the flux in webers, and  $I$  is the current in amperes.

**Figure 102.7** Magnetic flux produced by a current-carrying conductor. The inductance is defined by Eq. (102.9).



In an **inductor** the instantaneous voltage and current are related by the equation

$$v(t) = L \frac{di}{dt} \quad (102.10)$$

where  $v(t)$  is the instantaneous voltage in volts and  $di/dt$  is the rate of change of current with respect to time in amperes per second. Alternatively,

$$i(t) = i_0 + \frac{1}{L} \int_0^t v \, dt \quad (102.11)$$

where  $i_0$  is the initial current through the inductor at  $t = 0$ .

The energy stored in an inductor is

$$W = \frac{1}{2} I^2 L \quad (102.12)$$

Since the finite resistance of the wire in an inductor causes some energy loss, the quality factor, or  $Q$ , of an inductor—relating energy stored and energy dissipated—is an important quantity. It is

given by

$$Q = \frac{2\pi f L}{R} \quad (102.13)$$

where  $f$  is the frequency in hertz and  $R$  is the series resistance of the inductor.

In general, it is not possible to find an exact analytical expression for the inductance of a configuration. A number of empirical formulas are given as follows, and some general principles can be stated.

1. The inductance is proportional to the size of the coil. Thus larger loops have larger inductance.
2. Inductance is proportional to the number of turns squared. In Fig. 102.7, if there are two turns, the flux is doubled; if the flux is changing, then twice the induced voltage will appear in each turn, for a total of four times the voltage in a one-turn loop.
3. The inductance is proportional to the permeability of the surrounding medium. Most materials have a permeability  $\mu_0$  very close to that of free space;  $\mu_0 = 4 \cdot 10^{-4}$  henrys per meter. Magnetic materials have a permeability ranging from about 100 to 100 000 times  $\mu_0$ . They are used to concentrate and intensify the flux in transformers, motors, relays, and inductors.

Approximate formulas for the inductance of several configurations can be given. The internal inductance per unit length of a round solid wire is

$$L = \frac{\mu_0}{8\pi} \quad (102.14)$$

The inductance per unit length of two parallel wires carrying current in opposite directions is

$$L = 0.4 \cdot 10^{-6} \ln \left( \frac{2D}{d} \right) \quad (102.15)$$

where  $D$  is the center-to-center spacing of the wires and  $d$  is the diameter of the wire. For a coaxial cable the inductance per unit length is

$$L = \frac{\mu_0}{2\pi} \ln \left( \frac{r_b}{r_a} \right) \quad (102.16)$$

where the subscripts  $b$  and  $a$  refer to the radius of the outer and inner conductors, respectively. For a circular loop of wire,

$$L = \mu_0 a \left( \ln \left( \frac{8a}{r} \right) - 1.75 \right) \quad (102.17)$$

where  $a$  is the radius of the loop and  $r$  is the radius of the wire. For a solenoid consisting of a single layer of turns,

$$L = \frac{\mu_0 N^2 A}{0.45d + l} \quad (102.18)$$

where  $N$  is the number of turns,  $A$  is the cross-sectional area of the solenoid,  $d$  is the diameter of the solenoid, and  $l$  is its length.

In many cases the objective is to minimize the parasitic inductance of a layout. To achieve this one should minimize the area of current-carrying loops and use rectangular or flat conductors.

## Defining Terms

**Capacitance:** A property of an arrangement of two conductors equal to the charge on a conductor divided by the voltage between them.

**Capacitor:** An electronic component that accumulates charge and energy in an electric circuit.

**Dielectric:** An insulating material, for example, between the plates of a capacitor.

**Dissipation factor (DF):** The ratio of the energy dissipated per cycle to two times the maximum energy stored at a given frequency. The lower DF is, the more nearly ideal the capacitor is.

**Equivalent series resistance (ESR):** The resistance in series with a capacitor that accounts for energy loss in a nonideal capacitor.

**Inductance:** The property of a conductor arrangement that relates the magnetic flux to the current in the conductor.

**Inductor:** A passive component that stores energy in its magnetic field whenever it is conducting a current.

**Permittivity:** The property of a material that relates the electric field intensity to the electric flux density. The capacitance between two conductors is directly proportional to the permittivity of the dielectric between them.

**Quality factor (Q):** A value equal to two times the ratio of the maximum stored energy to the energy dissipated per cycle at a given frequency in a capacitor or inductor. The larger the value of  $Q$ , the more nearly ideal the component is. For a capacitor,  $Q$  is the inverse of DF.

**Resistivity:** The property of a material that relates current flow to electric field. The current density is equal to the electric field divided by the resistivity.

**Sheet resistance:** Sheet resistance (or sheet resistivity) is the average resistivity of a layer of material divided by its thickness. If the layer has the same length as width, the resistance between opposite edges of the "square" is numerically equal to the sheet resistance, regardless of the size of the square.

## References

Dorf, R. C. 1993. *The Electrical Engineering Handbook*. CRC Press, Boca Raton, FL.

Harper, C. C. 1977. *Handbook of Components for Electronics*. McGraw-Hill, New York.

Meeldijk, V. 1995. *Electronic Components: Selection and Application Guidelines*. John Wiley



& Sons, New York.

## **Further Information**

The *IEEE Transactions on Components, Packaging, and Manufacturing Technology* has a number of technical articles on components, and the CPMT Society sponsors the Electronic Components and Technology Conference. Both the transactions and the conference proceedings are the main sources for the latest developments in components.

The Capacitor and Resistor Technology Symposium is devoted entirely to the technology of electronic components. For further information on CARTS, contact: Components Technology Institute, Inc., 904 Bob Wallace Ave., Suite 117, Huntsville, AL 35801.

Ciletti, M. D. "RL, RC, and RLC Circuits"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## RL, RC, and RLC Circuits

---

### 103.1 RL Circuits

### 103.2 RC Circuits

### 103.3 RLC Circuits

Case 1: Overdamped Circuit • Case 2: Critically Damped Circuit • Case 3: Underdamped Circuit • RLC Circuit—Frequency Response

### Michael D. Ciletti

*University of Colorado, Colorado Springs*

Circuits that contain only resistive elements can be described by a set of algebraic equations that are obtained by systematically applying Kirchhoff's current and voltage laws to the circuit. When a circuit contains energy storage elements—that is, inductors and capacitors—Kirchhoff's laws are still valid, but their application leads to a differential equation (DE) model of the circuit instead of an algebraic model. A DE model can be solved by classical DE methods, by time-domain convolution, and by Laplace transform methods.

### 103.1 RL Circuits

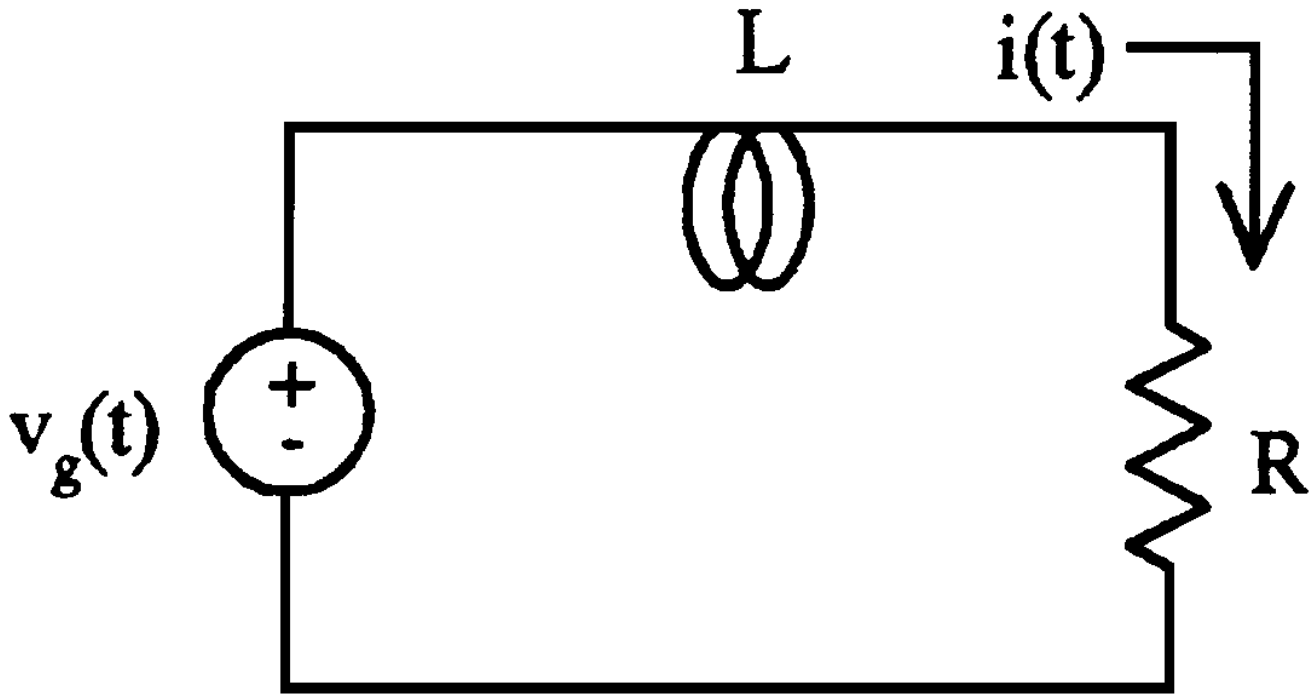
---

A series RL circuit can be analyzed by applying Kirchhoff's voltage law to the single loop that contains the elements of the circuit. For example, writing Kirchhoff's voltage law for the circuit in [Fig. 103.1](#) leads to the following first-order DE stating that the voltage drop across the inductor and the resistor must equal the voltage drop across the voltage source:

$$L \frac{di}{dt} + i(t)R = v_g(t) \quad (103.1)$$

This DE belongs to a family of linear, constant-coefficient, ordinary differential equations. Various approaches can be taken to find the solution for  $i(t)$ , the current through the elements of the circuit. The first is to solve the DE by classical (time-domain) methods; the second is to solve the equation by using Laplace transform theory. [A third approach, waveform convolution, will not be considered here (see [Ziemer et al., 1993](#)).]

**Figure 103.1** Series RL circuit.



The complete time-domain solution to the DE is formed as the sum of two parts, called the **natural solution** and the **particular solution**:

$$i(t) = i_N(t) + i_P(t) \quad (103.2)$$

The natural solution,  $i_N(t)$ , is obtained by solving the (homogeneous) DE with the sources (or forcing functions) turned off. The circuit in this example has the following homogeneous DE:

$$L \frac{di}{dt} + i(t)R = 0 \quad (103.3)$$

The solution to this equation is the exponential form:  $i_N(t) = K e^{st}$ , where  $K$  is an arbitrary constant. To verify this solution, substitute  $K e^{st}$  into Eq. (103.3) and form

$$LK s e^{st} + RK e^{st} = 0 \quad (103.4)$$

Rearranging gives:

$$[sL + R]K e^{st} = 0 \quad (103.5)$$

The factor  $K e^{st}$  can be canceled because the factor  $e^{st}$  is nonzero, and the constant  $K$  must be nonzero for the solution to be nontrivial. This reduction leads to the characteristic equation

$$sL + R = 0 \quad (103.6)$$

The **characteristic equation** of this circuit specifies the value of  $s$  for which  $i_N(t) = K e^{st}$  solves Eq. (103.3). By inspection,  $s = -R/L$  solves Eq. (103.6), and  $i_N(t) = K e^{-t/\tau}$  satisfies Eq. (103.3), and  $\tau = L/R$  is called the *time constant* of the circuit. The value  $s = -R/L$  is called the **natural frequency** of the circuit.

In general, higher-order circuits (those with higher-order differential equation models and corresponding higher-order algebraic characteristic equations) will have several natural frequencies, some of which may have a complex value—that is,  $s = \sigma + j\omega$ —where  $\sigma$  is an exponential damping factor and  $\omega$  is the undamped frequency of oscillation. The natural frequencies of a circuit play a key role in governing the dynamic response of the circuit to an input by determining the form and the duration of the transient waveform of the response [Ciletti, 1988].

The particular solution of this circuit's DE model is a function  $i_P(t)$  that satisfies Eq. (103.1) for a given source  $v_g(t)$ . For example, the constant input signal  $v_g(t) = V$  has the particular solution  $i_P(t) = V/R$ . [This can be verified by substituting this expression into Eq. (103.1).] The complete solution to the differential equation when  $v_g(t) = V$  is given by

$$i(t) = K e^{(-R/L)t} + V/R = K e^{-t/\tau} + V/R \quad (103.7)$$

The sources that excite physical circuits are usually modeled as being turned off before being applied at some specific time, say  $t_0$ , and the objective is to find a solution to its DE model for  $t \geq t_0$ . This leads to consideration of boundary conditions that constrain the behavior of the circuit and determine the unknown parameters in the solution of its DE model. The boundary conditions of the DE model of a physical circuit are determined by the energy that is stored in the circuit when the source is initially applied. For example, the inductor current in Fig. 103.1 could have the constraint given as  $i(t_0) = i_0$ , where  $i_0$  is a constant. If a circuit is modeled by a constant-coefficient DE, the time of application can be taken to be  $t_0 = 0$  without any loss in generality. (Note: The physical conditions that created  $i_0$  are not of concern.)

The value of the parameter  $K$  in Eq. (103.7) is specified by applying the given boundary condition to the waveform of Eq. (103.7), as follows:

$$i(0^+) = i_0 = K + V/R \quad (103.8)$$

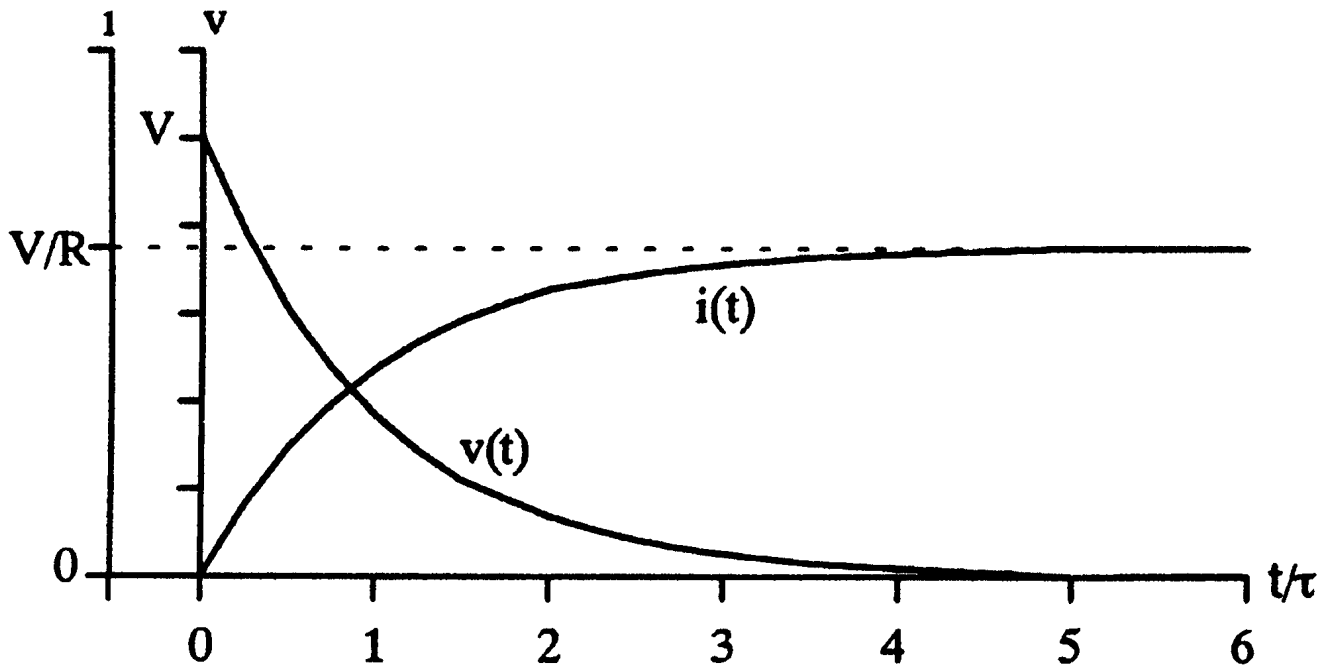
and so  $K = i_0 - V/R$ .

The response of the circuit to the applied input signal is the complete solution of the DE for  $t \geq 0$  evaluated with coefficients that conform to the boundary conditions. In the RL circuit example the complete response to the applied step input is

$$i(t) = [i_0 - V/R] e^{-t/\tau} + V/R, \quad t \geq 0 \quad (103.9)$$

The waveforms of  $i(t)$  and  $v(t)$ , the inductor voltage, are shown in Fig. 103.2 for the case in which the circuit is initially at rest, that is,  $i_0 = 0$ . The time axis in the plot has been normalized by  $\tau$ .

**Figure 103.2** Waveforms of  $v(t)$  and  $i(t)$  in the series RL circuit.



The physical effect of the inductor in this circuit is to provide inertia to a change in current in the series path that contains the inductor. An inductor has the physical property that its current is a continuous variable whenever the voltage applied across the terminals of the inductor has a bounded waveform (that is, no impulses). The current flow through the inductor is controlled by the voltage that is across its terminals. This voltage causes the accumulation of magnetic flux, which ultimately determines the current in the circuit. The initial voltage applied across the inductor is given by  $v(0^+) = V - i(0^+)R$ . If the inductor is initially relaxed—that is,  $i(0^-) = 0$ —the continuity property of the inductor current dictates that  $i(0^+) = i(0^-) = 0$ . So  $v(0^+) = V$ . All of the applied voltage initially appears across the inductor. When this voltage is applied for an interval of time, a magnetic flux accumulates and current is established through the inductor. Mathematically, the integration of this voltage causes the current in the circuit. The current waveform in [Fig. 103.2](#) exhibits exponential growth from its initial value of 0 to its final (steady state) value of  $V/R$ . The inductor voltage decays from its initial value to its steady state value of 0, and the inductor appears to be a "short circuit" to the steady state DC current.

## 103.2 RC Circuits

A capacitor acts like a reservoir of charge, thereby preventing rapid changes in the voltage across its terminals. A capacitor has the physical property that its voltage must be a continuous variable when its current is bounded. The DE model of the parallel RC circuit in [Fig. 103.3](#) is described according to Kirchhoff's current law by

$$C \frac{dv}{dt} + \frac{v}{R} = i_g(t) \quad (103.10)$$

Taking the one-sided Laplace transform [Ziemer, 1993] of this differential equation gives

$$sCV(s) - Cv(0^-) + \frac{V(s)}{R} = I_g(s) \quad (103.11)$$

where  $V(s)$  and  $I_g(s)$  denote the Laplace transforms of the related time-domain variables. (Note that the Laplace transform of the derivative of a variable explicitly incorporates the initial condition of the variable into the model of the circuit's behavior.) Rearranging this algebraic equation gives the following:

$$V(s) = \frac{I_g(s)}{sC + 1/R} + \frac{Cv(0)}{sC + 1/R} \quad (103.12)$$

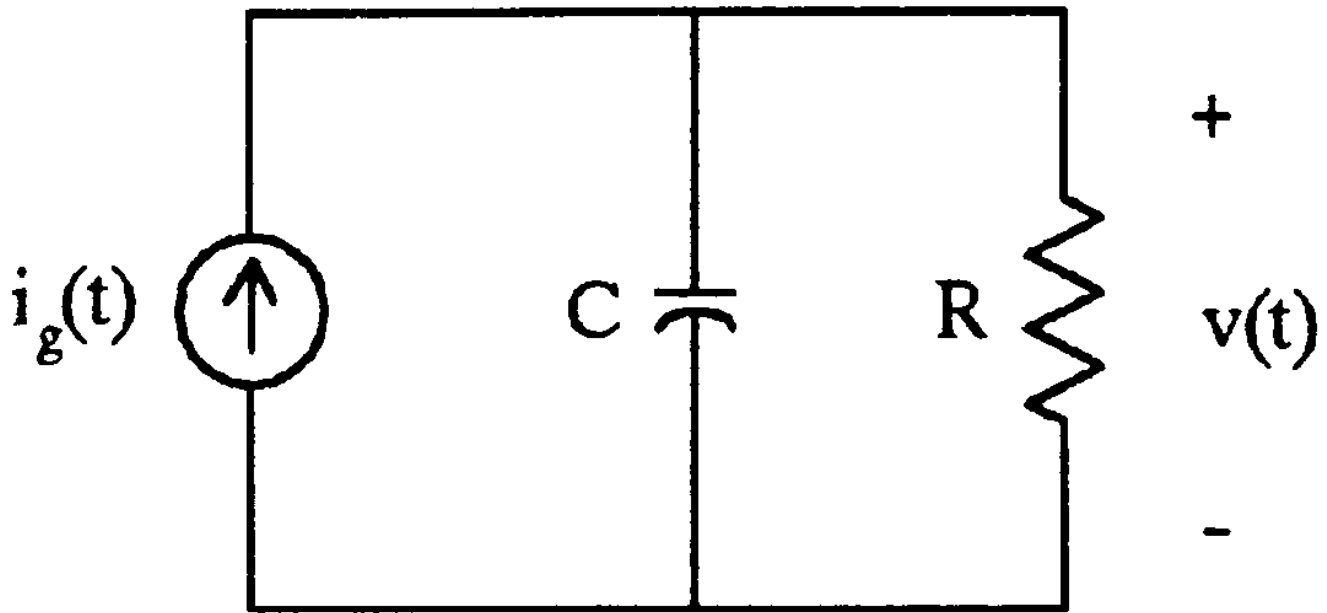
$$V(s) = I_g(s)H(s) + \frac{v(0)}{s + 1/(RC)} \quad (103.13)$$

where the  $s$ -domain function

$$H(s) = \frac{1}{sC + (1/R)} = \frac{1/C}{s + (1/RC)} \quad (103.14)$$

is called the input/output **transfer function** of the circuit. The expression in Eq. (103.13) illustrates an important property of linear circuits: the response of a linear RLC circuit variable is the superposition of the effect of the applied source and the effect of the initial energy stored in the circuit's capacitors and inductors. Another important fact is that the roots of the denominator polynomial of  $H(s)$  are the natural frequencies of the circuit (assuming no cancellation between numerator and denominator factors).

**Figure 103.3** Parallel RC circuit.



If a circuit is initially relaxed—that is, all capacitors and inductors are de-energized [set  $v(0^-) = 0$  in Eq. (103.13)], then the transfer function defines the ratio of the Laplace transform of the circuit's response to the Laplace transform of its stimulus. Alternatively, the transfer function and the Laplace transform of the input signal determine the Laplace transform of the output signal according to the simple product

$$V(s) = I_g(s)H(s) \quad (103.15)$$

The transfer function of a circuit can define a voltage ratio, a current ratio, a voltage-to-current ratio (impedance) or a current-to-voltage ratio (admittance). In this circuit  $H(s)$  relates the output (response of the capacitor voltage) to the input (i.e., the applied current source). Thus,  $H(s)$  is actually a generalized impedance function. If a circuit is initially relaxed its transfer function contains all of the information necessary to determine the response of the circuit to any given input signal.

Suppose that the circuit has an initial capacitor voltage and that the applied source in Eq. (103.13) is given by  $i_g(t) = Iu(t)$ , a step of height  $I$ . The Laplace transform [Ciletti, 1988] of the step waveform is given by  $I_g(s) = I/s$ , so the Laplace transform of  $v(t)$ , denoted by  $V(s)$ , is given by

$$V(s) = \frac{I/(sC)}{s + 1/(RC)} + \frac{v(0^-)}{s + 1/(RC)} \quad (103.16)$$

This expression for  $V(s)$  can be expanded algebraically into partial fractions as

$$V(s) = \frac{IR}{s} - \frac{IR}{s + 1/(RC)} + \frac{v(0^-)}{s + 1/(RC)} \quad (103.17)$$



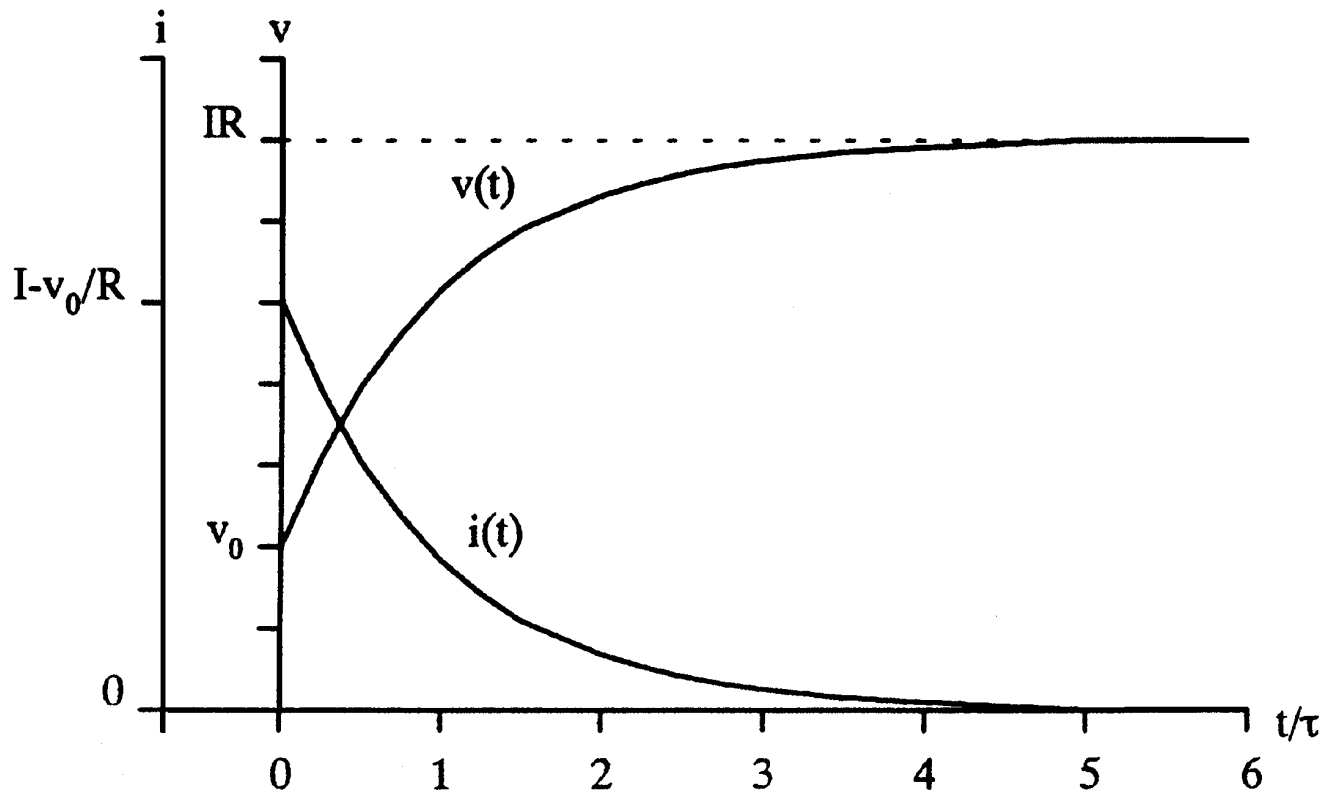
Associating the  $s$ -domain Laplace transform factor  $1/(s + a)$  with the time-domain function  $e^{-at}u(t)$  and taking the inverse Laplace transform of the individual terms of the expansion gives

$$v(t) = IR + [v(0^-) - IR]e^{-t/(RC)}, \quad t \geq 0 \quad (103.18)$$

When the source is applied to this circuit, the capacitor charges from its initial voltage to the steady state voltage given by  $v(\infty) = IR$ . The response of  $v(t)$  is shown in Fig. 103.4 with  $\tau = RC$  and  $v_o = v(0^-)$ . The capacitor voltage follows an exponential transition from its initial value to its steady state value at a rate determined by its time constant,  $\tau$ . A similar analysis would show that the response of the capacitor current is given by

$$i(t) = [I - v(0^-)/R]e^{-t/RC}, \quad t \geq 0 \quad (103.19)$$

**Figure 103.4** Step response of an initially relaxed RC circuit.



Capacitors behave like short circuits to sudden changes in current. The initial capacitor voltage determines the initial resistor current by Ohm's law:  $v(0^+)/R$ . Any initial current supplied by the source in excess of this amount will pass through the capacitor as though it were a short circuit. As the capacitor builds voltage, the resistor draws an increasing amount of current and ultimately conducts all of the current supplied by the constant source. In steady state the capacitor looks like

an open circuit to the constant source— $i(\infty) = 0$ —and it conducts no current.

### 103.3 RLC Circuits

---

Circuits that contain inductors and capacitors exhibit dynamical effects that combine the inductor's inertia to sudden changes in current with the capacitor's inertia to sudden changes in voltage. The topological arrangement of the components in a given circuit determines the behavior that results from the interaction of the currents and voltages associated with the individual circuit elements.

The parallel RLC circuit in Fig. 103.5(a) has an  $s$ -domain counterpart [see Fig. 103.5(b)], sometimes referred to as a *transformed circuit*, that is obtained by replacing each time-domain variable by its Laplace transform, and each physical component by a Laplace transform model of the component's voltage-current relationship. Here, for example, the physical capacitor is replaced by a model that accounts for the impedance of the capacitor and the capacitor's initial voltage. The additional sources account for the possibly nonzero initial values of capacitor voltage and inductor current. Algebraic expressions of Kirchhoff's laws are written from the Laplace model of the circuit. Applying Kirchhoff's current law to the circuit of Fig. 103.5 gives

$$sCV(s) + \frac{V(s)}{sL} + \frac{V(s)}{R} = I_g(s) + Cv(0^-) - i(0^-) \quad (103.20)$$

Algebraic manipulation of this expression gives

$$V(s) = \frac{(s/C)I_g(s)}{s^2 + (s/RC) + (1/LC)} + \frac{s/C[Cv(0^-) - i(0^-)]}{s^2 + (s/RC) + (1/LC)} \quad (103.21)$$

The transfer function relating the source current to the capacitor voltage is obtained directly from Eq. (103.21), with  $v(0^-) = 0$  and  $i(0^-) = 0$ :

$$H(s) = \frac{s/C}{s^2 + (s/RC) + (1/LC)} \quad (103.22)$$

Because  $V(s)$  represents the response of a second-order circuit, the form of its partial fraction expansion depends on the natural frequencies of the circuit. These are obtained by solving for the roots of the characteristic equation:

$$s^2 + \frac{s}{RC} + \frac{1}{LC} = 0 \quad (103.23)$$

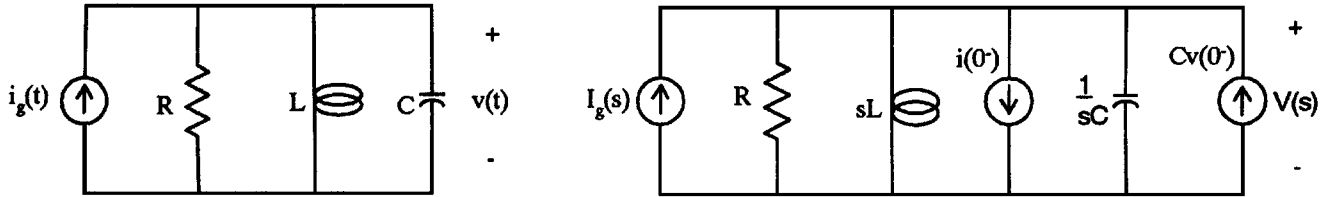
The roots are

$$s_1 = \frac{-1}{2RC} + \sqrt{\left[\frac{1}{2RC}\right]^2 - \frac{1}{LC}} \quad (103.24)$$

$$s_2 = \frac{-1}{2RC} - \sqrt{\left[\frac{1}{2RC}\right]^2 - \frac{1}{LC}} \quad (103.25)$$

To illustrate the possibilities of the second-order response, we let  $i_g(t) = u(t)$ , and  $I_g(s) = 1/s$ .

**Figure 103.5** Parallel RLC circuit.



### Case 1: Overdamped Circuit

If both roots of the second-order characteristic equation are real valued the circuit is said to be *overdamped*. The physical significance of this term is that the circuit response to a step input exhibits exponential decay and does not oscillate. The form of the step response of the circuit's capacitor voltage is

$$v(t) = K_1 e^{s_1 t} + K_2 e^{s_2 t} \quad (103.26)$$

and  $K_1$  and  $K_2$  are chosen to satisfy the initial conditions imposed by  $v(0^-)$  and  $i(0^-)$ .

### Case 2: Critically Damped Circuit

When the two roots of a second-order characteristic equation are identical the circuit is said to be *critically damped*. The circuit in this example is critically damped when  $s_1 = s_2 = -1/(2RC)$ . In this case, Eq. (103.21) becomes

$$V(s) = \frac{I/C}{[s + (1/2RC)]^2} + \frac{s}{C} \frac{[Cv(0^-) - i(0^-)]}{[s + (1/2RC)]^2} \quad (103.27)$$

The partial fraction expansion of this expression is

$$V(s) = \frac{1/C}{[s + (1/2RC)]^2} + \frac{1}{C} \frac{[Cv(0^-) - i(0^-)]}{s + (1/2RC)} - \frac{1}{2RC^2} \frac{[Cv(0^-) - i(0^-)]}{[s + (1/2RC)]^2} \quad (103.28)$$

Taking the inverse Laplace transform of  $V(s)$  gives

$$v(t) = \frac{1}{C}te^{-t/(2RC)} + \frac{1}{C}[Cv(0^-) - i(0^-)]e^{-t/(2RC)} - \frac{1}{2RC^2}[Cv(0^-) - i(0^-)]te^{-t/(2RC)} \quad (103.29)$$

The behavior of the circuit in this case is called *critically damped* because a reduction in the amount of damping in the circuit would cause the circuit response to oscillate.

### Case 3: Underdamped Circuit

The component values in this case are such that the roots of the characteristic equation are a complex conjugate pair of numbers. This leads to a response that is oscillatory, having a damped frequency of oscillation,  $\omega_d$ , given by

$$\omega_d = \sqrt{\frac{1}{LC} + \left[\frac{1}{2RC}\right]^2} \quad (103.30)$$

and a damping factor,  $\alpha$ , given by

$$\alpha = \frac{1}{2RC} \quad (103.31)$$

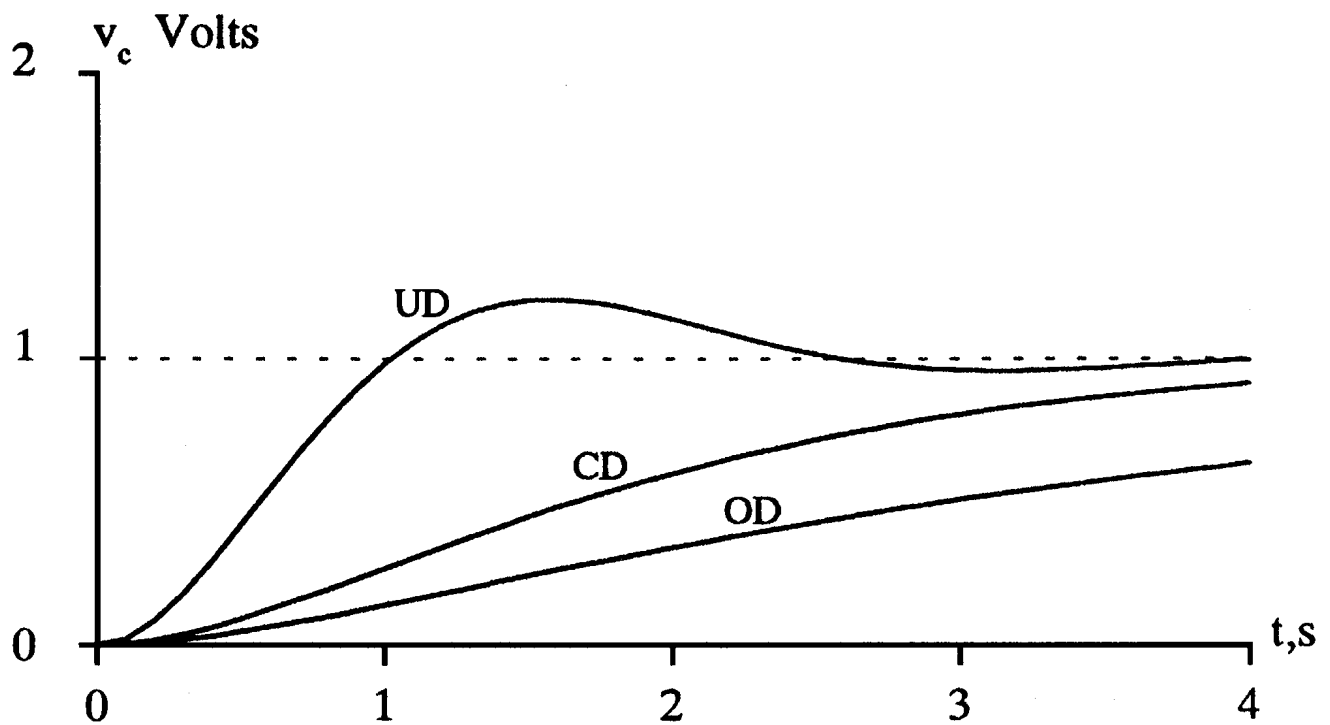
The form of the response of the capacitor voltage to a unit step input current source is

$$v(t) = 2|K|e^{-\alpha t} \sin(\omega_d t + \varphi) \quad (103.32)$$

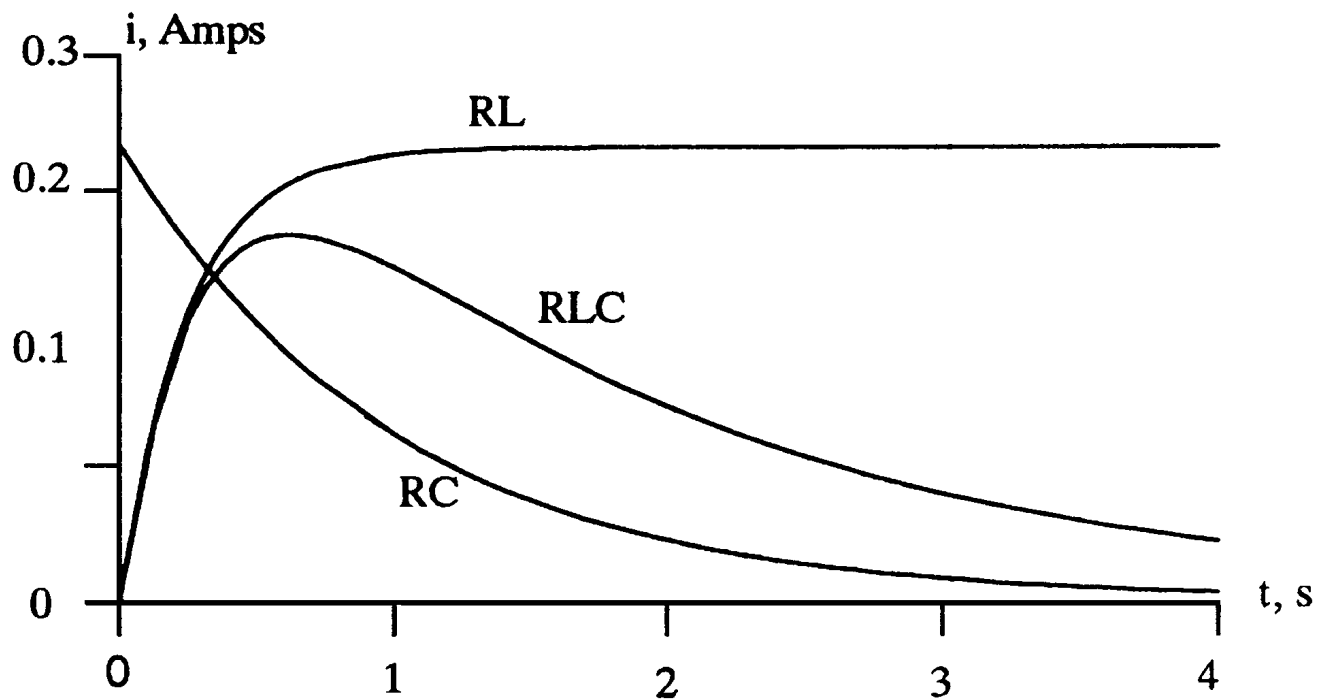
where the parameters  $K$  (a nonnegative constant) and  $\varphi$  are chosen to satisfy initial conditions.

A comparison of overdamped, critically damped, and underdamped responses of the capacitor voltage is given in [Fig. 103.6](#) for a unit step input, with the circuit initially relaxed. Additional insight into the behavior of RL, RC, and RLC circuits can be obtained by examining the response of the current in the series RL, RC, and RLC circuits for corresponding component values. For example, [Fig. 103.7](#) shows the waveforms of the step response of  $i(t)$  for the following component values:  $R = 4 \, \Omega$ ,  $L = 1 \, \text{H}$ , and  $C = 1 \, \text{F}$ . The RL circuit has a time constant of 0.25 s, the RC circuit has a time constant of 2.00 s, and the overdamped RLC circuit has time constants of 0.29 s and 1.69 s. The inductor in the RL circuit blocks the initial flow of current, the capacitor in the RC circuit blocks the steady state flow of current, and the RLC circuit exhibits a combination of both behaviors. Note that the time constants of the RLC circuit are bounded by those of the RL and RC circuits.

**Figure 103.6** Overdamped, critically damped, and underdamped responses of a parallel RLC circuit with a step input.



**Figure 103.7** A comparison of current in series RL, RC, and RLC circuits.



## RLC Circuit – Frequency Response

A circuit's transfer function defines a relationship between the frequency-domain (spectral) characteristics of any input signal and the frequency-domain characteristics of its corresponding output signal. In many engineering applications circuits are used to shape the spectral characteristics of a signal. For example, in a communications application the frequency response could be chosen to eliminate noise from a signal. The frequency response of a circuit consists of the graphs of  $|H(j\omega)|$  and  $\theta(j\omega)$ , the magnitude response, and the phase response, respectively, of the transfer function of a given circuit voltage or current. The term  $|H(j\omega)|$  determines the ratio of the amplitude of the *sinusoidal steady state response* of the circuit to a sinusoidal input, and  $\theta(j\omega)$  determines the phase shift (manifest as a time-axis translation) between the input and the output waveforms. Values of  $|H(j\omega)|$  and  $\theta(j\omega)$  are obtained by taking the magnitude and angle, respectively, of the complex value  $H(j\omega)$ .

The magnitude and phase responses play an important role in filter theory, where for distortionless transmission (i.e., the filter output waveform is a scaled and delayed copy of the input waveform) it is necessary that  $|H(j\omega)| = K$ , a constant (flat response), and  $\theta(j\omega)$  be linear over the pass band of a signal that is to be passed through a filter.

In the parallel RC circuit of [Fig. 103.3](#), the transfer function relating the capacitor voltage to the current source has

$$H(j\omega) = \frac{R}{1 + j\omega RC} \quad (103.33)$$

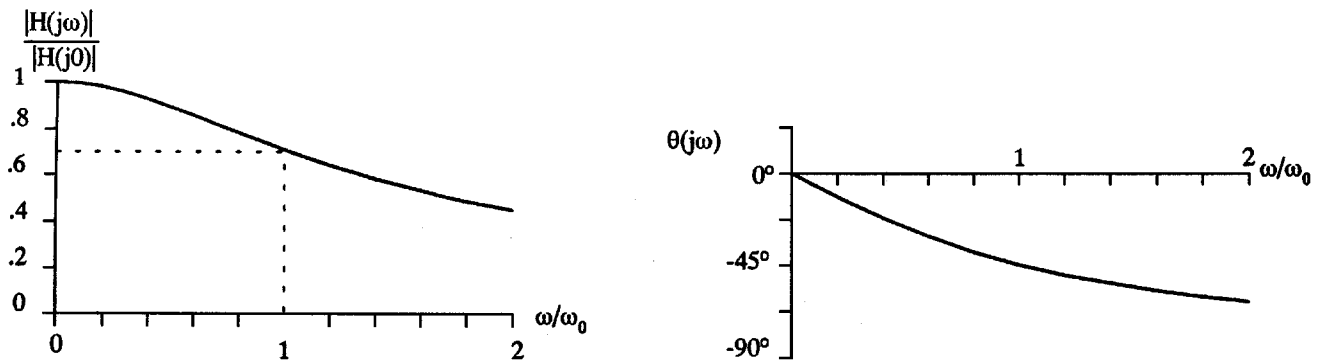
The arithmetic of complex numbers gives the following

$$|H(j\omega)| = \frac{R}{\sqrt{1 + \omega^2 R^2 C^2}} \quad (103.34)$$

$$\theta(j\omega) = \angle H(j\omega) = -\tan^{-1}(\omega RC) \quad (103.35)$$

The graphs of  $|H(j\omega)|$  and  $\theta(j\omega)$  are shown in [Fig. 103.8](#). The capacitor voltage in the parallel RC circuit has a "low pass" filter response, meaning that it will pass low frequency sinusoidal signals without significant attenuation provided that  $\omega < \omega_o$  [ $\omega_o = 1/(RC)$ ], the cutoff frequency of the filter. The approximate linearity of the phase response within the passband is also evident.

**Figure 103.8** Magnitude and phase response of a parallel RC circuit.



A filter's component values determine its cutoff frequency. An important design problem is to determine the values of the components so that a specified cutoff frequency is realized by the circuit. Here, increasing the size of the capacitor lowers the cutoff frequency, or, alternatively, reduces the bandwidth of the filter. Other typical filter responses that can be formed by RL, RC, and RLC circuits (with and without op amps) are high-pass, band-pass, and band-stop filters.

## Defining Terms

**Characteristic equation:** The equation obtained by setting the denominator polynomial of a transfer function equal to zero. The equation defines the natural frequencies of a circuit.

**Generalized impedance:** An  $s$ -domain transfer function in which the input signal is a circuit's current and the output signal is a voltage in the circuit.

**Natural frequency:** A root of the characteristic polynomial. A natural frequency corresponds to a mode of exponential time-domain behavior.

**Natural solution:** The solution to the unforced differential equation.

**Particular solution:** The solution to the differential equation for a particular forcing function.

**Steady state response:** The response of a circuit after sufficient time has elapsed to allow the transient response to become insignificant.

**Transient response:** The response of a circuit prior to its entering the steady state.

**Transfer function:** An  $s$ -domain function that determines the relationship between an exponential forcing function and the particular solution of a circuit's differential equation model. It also describes a relationship between the  $s$ -domain spectral description of a circuit's input signal and the spectral description of its output signal.

## References

- Ciletti, M. D., 1988. *Introduction to Circuit Analysis and Design*. Holt, Rinehart and Winston, New York.
- Ziemer, R. E., Tranter, W. H., and Fannin, D. R. 1993. *Signals and Systems: Continuous and Discrete*. Macmillan, New York.

## Further Information

For further information on the basic concepts of RL, RC, and RLC circuits, see *Circuits, Devices,*

*and Systems* by R. J. Smith and R. C. Dorf. For a treatment of convolution methods, Fourier transforms, and Laplace transforms, see *Signals and Systems: Continuous and Discrete* by R. E. Ziemer *et al.* For a treatment of RL, RC, and RLC circuits with op amps, see *Introduction to Circuit Analysis and Design* by Ciletti.



Svoboda, J. A. "Node Equations and Mesh Equations"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Node Equations and Mesh Equations

---

## 104.1 Node Equations

## 104.2 Mesh Equations

**James A. Svoboda**

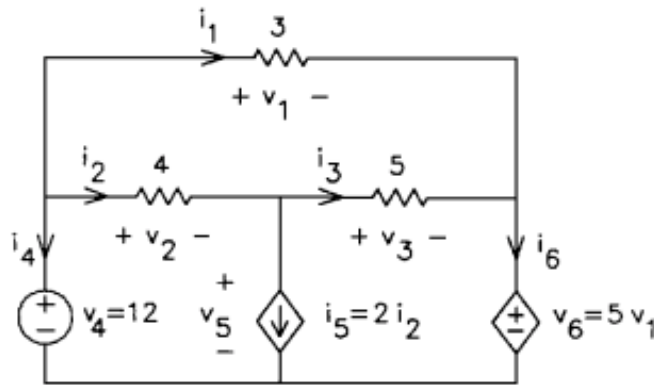
*Clarkson University*

Node equations and mesh equations are sets of simultaneous equations that are used to analyze electric circuits. Engineers have been writing and solving node equations and mesh equations for a long time. For example, procedures for formulating node equations and mesh equations are found in textbooks from the late 1950s and 1960s [Seshu and Balabanian, 1959; Seshu and Reed, 1961; Desoer and Kuh, 1969]. This longevity is likely due to two facts. First, node equations and mesh equations are easy to write. Second, the node equations and the mesh equations are both relatively small sets of simultaneous equations.

Electric circuits are interconnections of electrical devices, for example, resistors, independent and dependent sources, capacitors, inductors, op amps, and so on. Circuit behavior depends both on how the devices work and on how the devices are connected together. **Constitutive equations** describe how the devices in the circuit work. Ohm's law is an example of a constitutive equation. The Kirchhoff's law equations describe how the devices are connected together to form the circuit. Both the node equations and the mesh equations efficiently organize the information provided by the constitutive equations and the Kirchhoff's law equations.

In order to better appreciate the advantages of using node and mesh equations, an example will be done without them. The constitutive equations and Kirchhoff's law equations comprise a set of simultaneous equations. In this first example, illustrated in Fig. 104.1, these equations are used to analyze an electric circuit. This example also illustrates the use of MathCAD to solve simultaneous equations. The availability of such a convenient method of solving equations influences the way that node equations and mesh equations are formulated.

**Figure 104.1** Example illustrating the use of MathCAD to solve simultaneous equations.



branch.MCD1

0 79 auto

### Solving a Circuit using Branch Equations

Guess the values of the branch voltages and currents. Any guess will do. A good guess is not required. In this example,  $V_4$  is known to be 12. Zero is a suitable guess for the other branch voltages and currents.

$I_1 := 0$	$V_1 := 0$	$I_2 := 0$	$V_2 := 0$
$I_3 := 0$	$V_3 := 0$	$I_4 := 0$	$V_4 := 12$
$I_5 := 0$	$V_5 := 0$	$I_6 := 0$	$V_6 := 0$

Enter the equations describing the branches of the network. The word "Given" marks the beginning of these equations.

Given

Enter the Kirchhoff's Current Law (KCL) equations. (Use <ALT>= for the wiggly equal signs.)

$I_1 + I_2 + I_4 = 0$	$-I_2 + I_3 + I_5 = 0$	$-I_1 - I_3 + I_6 = 0$
-----------------------	------------------------	------------------------

Enter the Kirchhoff's Voltage Law (KVL) equations.

$V_1 - V_3 - V_2 = 0$	$V_2 + V_5 - V_4 = 0$	$V_3 + V_6 - V_5 = 0$
-----------------------	-----------------------	-----------------------

Enter the Constitutive Equations.

$V_1 = 3 \cdot I_1$	$V_2 = 4 \cdot I_2$	$V_3 = 5 \cdot I_3$	$V_4 = 12$
$I_5 = 2 \cdot I_2$	$V_6 = 5 \cdot V_1$		

Ask MathCAD to solve these equations. The word "Find" marks the end of the equations.

Find( $I_1, I_2, I_3, I_4, I_5, I_6, V_1, V_2, V_3, V_4, V_5, V_6$ ) =

	0.667
	-2
	2
	1.333
	-4
	2.667
	2
	-8
	10
	12
	20
	10

The circuit shown in Fig. 104.1 consists of six devices. Two variables—a branch current and a voltage—are associated with each branch. There are 12 variables associated with this small circuit! The constitutive equations describe each of the six devices:

$$\begin{aligned} v_1 &= 3i_1, & v_2 &= 4i_2, & v_3 &= 5i_3, \\ v_4 &= 12, & i_5 &= 2i_2, & v_6 &= 5v_1 \end{aligned} \quad (104.1)$$

Kirchhoff's laws provide six more equations:

$$\begin{aligned} i_1 + i_2 + i_4 &= 0, & -i_2 + i_3 + i_5 &= 0, & -i_1 - i_3 + i_6 &= 0, \\ v_1 - v_3 - v_2 &= 0, & v_2 + v_5 - v_4 &= 0, & v_3 + v_6 - v_5 &= 0 \end{aligned} \quad (104.2)$$

Although it is not practical to solve 12 equations in 12 unknowns by hand, these equations can easily be solved using a personal computer with appropriate software. Figure 104.1 includes a screen dump that illustrates the use of MathCAD [Wieder, 1992] to solve these equations.

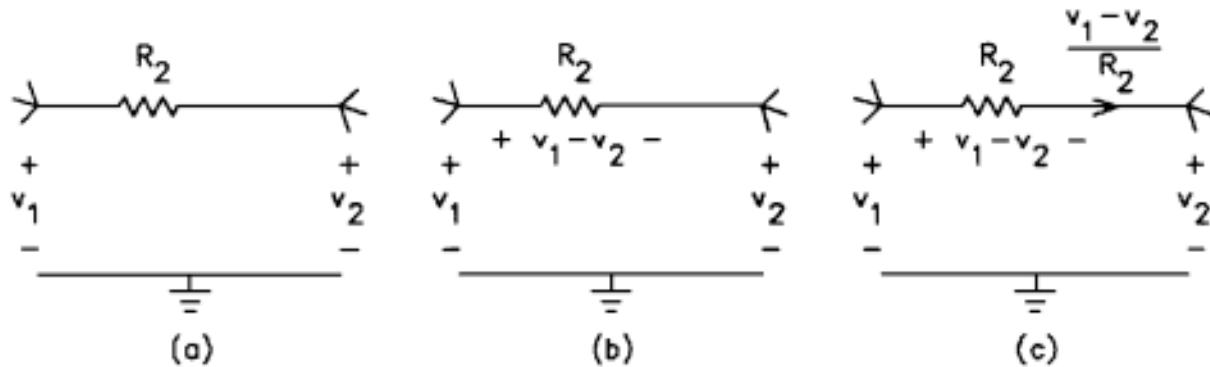
Writing all of these equations is tedious and the process becomes more tedious as the size of the circuit increases. It would be convenient to use a smaller set of simultaneous equations, hence the interest in node equations and mesh equations. The size of the set of simultaneous equations is very important when the equations must be solved by hand. Most contemporary treatments of node equations and mesh equations [Dorf, 1993; Irwin, 1993; Nilsson, 1993] describe procedures that result in as small a set of equations as possible. The availability of the personal computer with appropriate software reduces, but does not eliminate, the importance of the size of the set of equations.

## 104.1 Node Equations

This section describes a procedure for obtaining node equations to represent a connected circuit. This procedure is based on the observation that an independent set of equations can be obtained by applying Kirchhoff's current law (KCL) at all of the nodes of the circuit except for one node [Seshu and Balabanian, 1959; Seshu and Reed, 1961; Desoer and Kuh, 1969]. The node at which KCL is not applied is called the **reference node**.

The voltage at any node, with respect to the reference node, is called the **node voltage** at that node. Figure 104.2 shows that the current in a resistor can be expressed in terms of the voltages at the nodes of the resistor. This is accomplished in three steps. First, in Fig. 104.2(a), the node voltages corresponding to the nodes of the resistor are labeled. Next, in Fig. 104.2(b), the branch voltage is expressed in terms of the node voltages (notice the polarities of the voltages). Finally, in Fig. 104.2(c), Ohm's law is used to express the branch current in terms of the node voltages (notice the polarities of the branch voltage and current).

**Figure 104.2** Expressing branch voltages and currents as functions of node voltages.

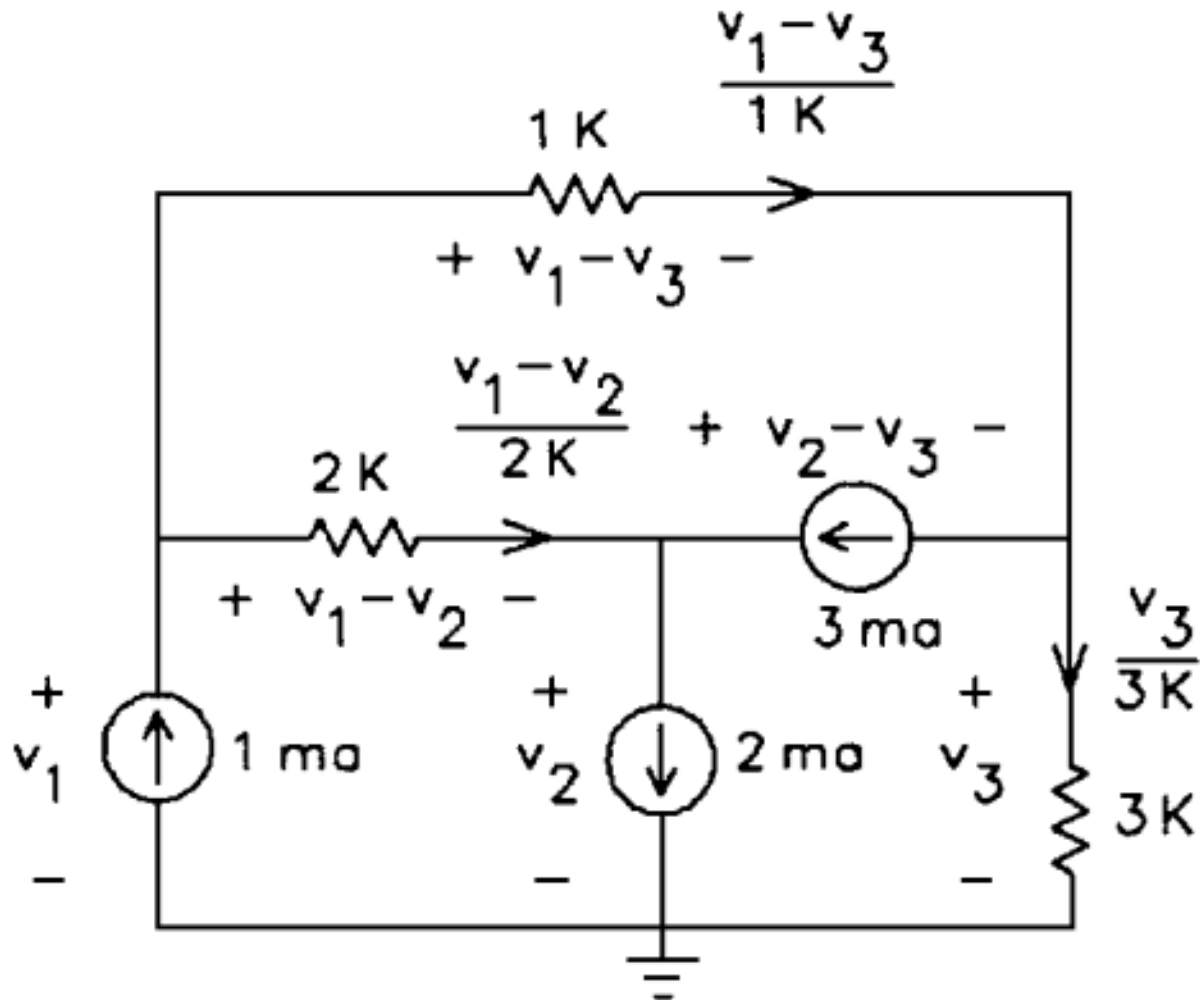


Consider writing simultaneous equations to represent the circuit shown in Fig. 104.3. When the reference node is selected as shown in Fig. 104.3, the node voltages are  $v_1$ ,  $v_2$ , and  $v_3$ . The resistor currents have been expressed in terms of the node voltages using the technique described in Fig. 104.2. Since all of the branch currents have been labeled, a set of simultaneous equations can now be obtained by applying KCL at all of the nodes except for the reference node.

$$\begin{aligned} .001 &= \frac{v_1 - v_2}{2000} + \frac{v_1 - v_3}{1000} \\ .002 &= .003 + \frac{v_1 - v_2}{2000} \\ \frac{v_1 - v_3}{1000} &= .003 + \frac{v_3}{3000} \end{aligned} \quad (104.3)$$

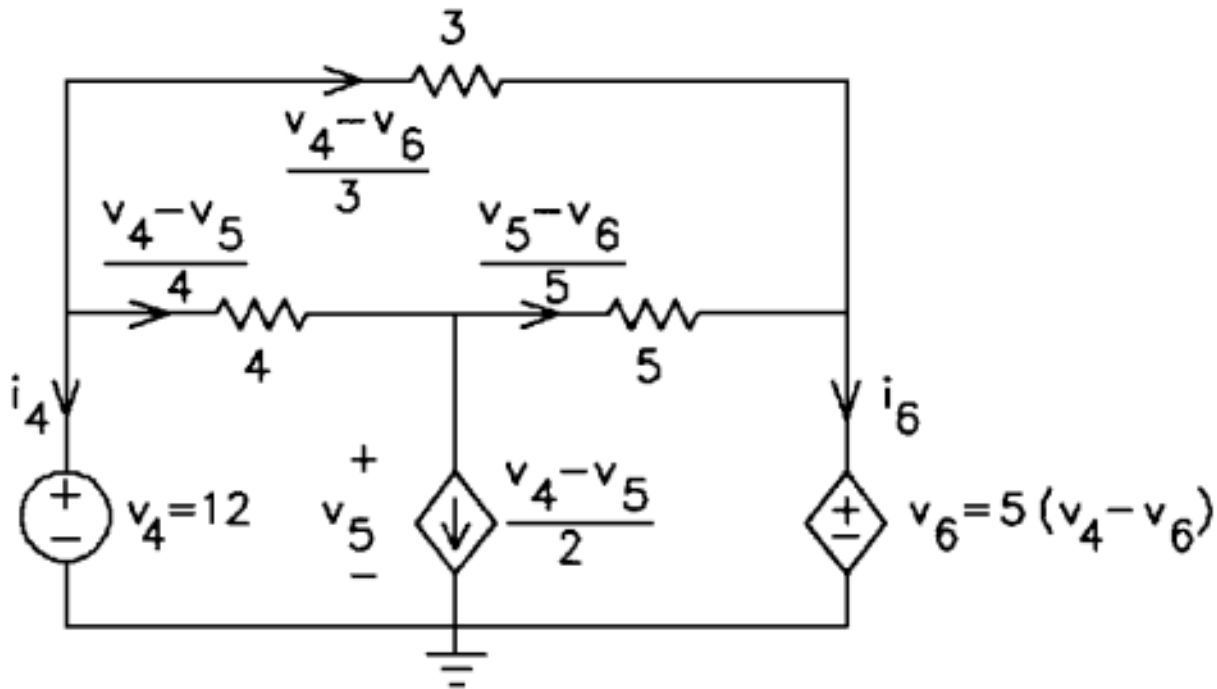
The independent variables in these equations are node voltages and so this set of simultaneous equations is called the *node equations*. Representing the branch voltages and branch currents as functions of the node voltages made it possible to represent this circuit by a smaller set of variables, the node voltages.

**Figure 104.3** A circuit consisting of resistors and current sources.



Consider again the circuit shown in [Fig. 104.1](#). This circuit contains two voltage sources (one independent and one dependent). In general, there is no easy way to express the current in a voltage source as a function of the node voltages. Since the size of the set of simultaneous equations is not of critical importance, it is appropriate to add the currents of the voltage sources to the list of independent variables. The circuit has been redrawn in [Fig. 104.4](#). A reference node has been selected and labeled. The independent variables, that is, the node voltages and currents in the voltage sources, have been labeled.

**Figure 104.4** Writing node equations for the circuit from Fig. 104.1.



The voltage across each voltage source can be expressed in two ways. The source voltage is given by the constitutive equation. The branch voltage is the difference of the node voltages at the nodes of the voltage source. These expressions must be equivalent so the branch voltage can be equated to the source voltage. In Fig. 104.4 this means that

$$v_4 = 12 \quad \text{and} \quad v_6 = 5(v_4 - v_6) \quad (104.4)$$

The circuit in Fig. 104.4 contains two dependent sources. Dependent sources will not be a problem if the controlling variables of the dependent sources are first expressed as functions of the independent variables of the circuit. In Fig. 104.4 this means that  $i_2$  and  $v_1$  must be expressed as functions of  $v_4, v_5, v_6, i_4$ , and  $i_6$ .

$$i_2 = \frac{v_4 - v_5}{4} \quad \text{and} \quad v_1 = v_4 - v_6 \quad (104.5)$$

Next the controlled variables of the dependent source can be expressed as functions of the independent variables of the circuit:

$$i_5 = 2 \left( \frac{v_4 - v_5}{4} \right) = \frac{v_4 - v_5}{2} \quad \text{and} \quad v_6 = 5(v_4 - v_6) \quad (104.6)$$

Now all of the branch currents have been labeled. Apply KCL at all of the nodes except the reference node to get

$$\begin{aligned}\frac{v_4 - v_6}{3} + \frac{v_4 - v_5}{4} + i_4 &= 0 \\ -\frac{v_4 - v_5}{4} + \frac{v_4 - v_5}{2} + \frac{v_5 - v_6}{5} &= 0 \quad (104.7) \\ -\frac{v_5 - v_6}{5} - \frac{v_4 - v_6}{3} + i_6 &= 0\end{aligned}$$

Equations (104.4) and (104.7) comprise the node equations representing this circuit.

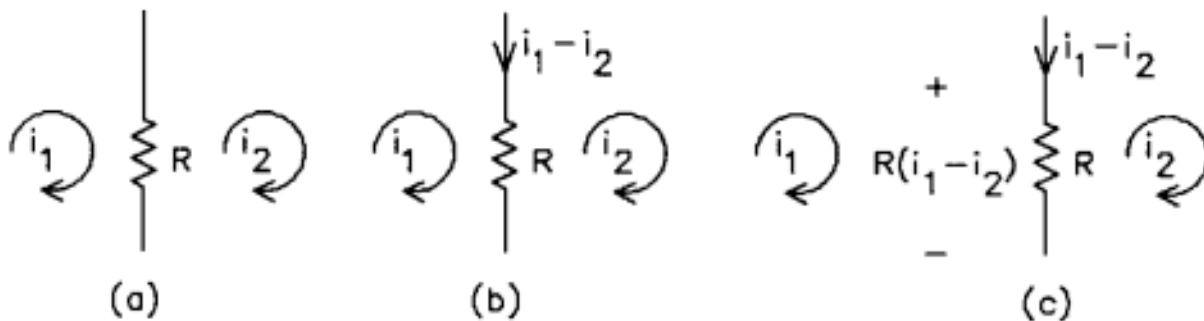
In summary, the following procedure is used to write node equations:

1. Choose and label the reference node. Label the independent variables: node voltages and voltage source currents.
2. Express branch currents as functions of the independent variables and the input variables.
3. Express the branch voltage of each voltage source as the difference of the node voltages at its nodes. Equate the branch voltage to the voltage source voltage.
4. Apply KCL at each node except for the reference node.

## 104.2 Mesh Equations

Mesh equations are formulated by applying Kirchhoff's voltage law (KVL) to each **mesh** of a circuit. A **mesh current** is associated with each mesh. [Figure 104.5](#) shows how to express the branch voltage and branch current of a resistor as functions of the mesh currents.

**Figure 104.5** Expressing branch voltages and currents as functions of mesh currents.



The procedure for formulation of mesh equations is analogous to the procedure for formulating node equations:

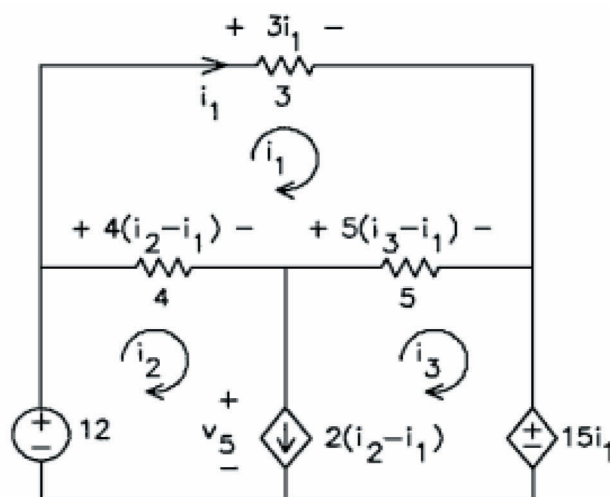
1. Label the independent variables: mesh currents and current source voltages.
2. Express branch voltages as functions of the independent variables and the input variables.
3. Express the branch current of each current source as the difference of the mesh currents of its meshes. Equate the branch current to the current source current.



4. Apply KVL to each mesh.

As an example, consider again the circuit shown in Fig. 104.1. This circuit is redrawn in Fig. 104.6 to show the mesh currents  $i_1$ ,  $i_2$ , and  $i_3$ . The branch voltage of the current source,  $v_5$ , is also labeled, because it will be added to the list of independent variables. In Fig. 104.6, the branch voltages have been expressed as functions of the independent variables  $i_1$ ,  $i_2$ ,  $i_3$ , and  $v_5$ .

**Figure 104.6** Writing mesh equations for the circuit from Fig. 104.1.



mesh\_eqn.MCD4

23 77

### Solving a Circuit using Mesh Equations

Guess the values of the mesh currents and current source voltage. Any guess will do. A good guess is not required.

$i1 := 0$        $i2 := 0$        $i3 := 0$        $v5 := 0$

Given

branch current =  $i2 - i3 = 2 \cdot (i2 - i1)$  = current source current

Apply KVL to each mesh.

$$-4 \cdot (i2 - i1) + 3 \cdot i1 - 5 \cdot (i3 - i1) \approx 0 \quad -12 + 4 \cdot (i2 - i1) + v5 \approx 0$$

$$-v5 + 5 \cdot (i3 - i1) + 15 \cdot i1 \approx 0$$

$$\text{Find}(i1, i2, i3, v5) = \begin{bmatrix} 0.667 \\ -1.333 \\ 2.667 \\ 20 \end{bmatrix}$$

The current in the dependent current source can be expressed in two ways. This source current is given by the constitutive equation. This branch current is also the difference of two mesh currents. These expressions must be equal so

$$i2 - i3 = 2(i2 - i1) \quad (104.8)$$

Next, apply KVL to each mesh to get

$$\begin{aligned}-4(i_2 - i_1) + 3i_1 - 5(i_3 - i_1) &= 0 \\ -12 + 4(i_2 - i_1) + v_5 &= 0 \\ -v_5 + 5(i_3 - i_1) + 15i_1 &= 0\end{aligned}\quad (104.9)$$

Equations (104.8) and (104.9) comprise the mesh equations. [Figure 104.6](#) illustrates the use of MathCAD to solve the circuit using mesh equations.

## Defining Terms

**Constitutive equations:** Equations that describe the relationship between the branch current and branch voltage of an electric device. Ohm's law is an example of a constitutive equation.

**Mesh:** A loop that does not contain any branch in its interior. Only planar circuits have meshes. Redrawing a planar circuit can change the meshes [[Seshu and Reed, 1961](#); [Desoer and Kuh, 1969](#)].

**Mesh current:** A current associated with a mesh. This current circulates around the mesh.

**Node voltage:** The voltage at any node, with respect to the reference node.

**Reference node:** An independent set of equations can be obtained by applying KCL at all of the nodes of the circuit except for one node. The node at which KCL is not applied is called the *reference node*. Any node of the circuit can be selected to be the reference node. In electronic circuits, where one node of the network is the ground node of the power supplies, the reference node is almost always selected to be the ground node.

## References

- Desoer, C. A. and Kuh, E. S. 1969. *Basic Circuit Theory*. McGraw-Hill, New York.
- Dorf, R. C. 1993. *Introduction to Electric Circuits*. John Wiley & Sons, New York.
- Irwin, J. D. 1993. *Basic Engineering Circuit Analysis*. Macmillan, New York.
- Nilsson, J. W. 1993. *Electric Circuits*. Addison-Wesley, Reading, MA.
- Seshu, S. and Balabanian, N. 1959. *Linear Network Analysis*. John Wiley & Sons, New York.
- Seshu, S. and Reed, M. B. 1961. *Linear Graphs and Electric Networks*. Addison-Wesley, Reading, MA.
- Wieder, S. 1992. *Introduction to MathCAD for Scientists and Engineers*. McGraw-Hill, New York.

## Further Information

*Computer Methods for Circuit Analysis and Design* by Jiri Vlach and Kishore Singhal describes procedures for formulating circuit equations that are well suited to computer-aided analysis.

Rashid, M. H. "Sinusoidal Excitation and Phasors"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Sinusoidal Excitation and Phasors

---

### 105.1 Sinusoidal Source

### 105.2 Phasor

### 105.3 Passive Circuit Elements in Phasors Domain

Resistor • Inductor • Capacitor • Sinusoidal Responses

**Muhammad H. Rashid**

*Purdue University*

A sinusoidal forcing function known as a **sinusoid** is one of the most important excitations. In electrical engineering, the carrier signals for communications are sinusoids, and the sinusoid is also the dominant signal in the power industry. Sinusoids abound in nature, as, for example, in the motion of a pendulum, in the bouncing of a ball, and in the vibrations of strings and membranes.

Because of the importance of sinusoids, the output response of a circuit due to a sinusoidal input signal is an important criterion in determining the performance of a circuit. In this section, we will define a sinusoidal function, represent it in a phasor form, and then illustrate the techniques for determining the sinusoidal responses.

### 105.1 Sinusoidal Source

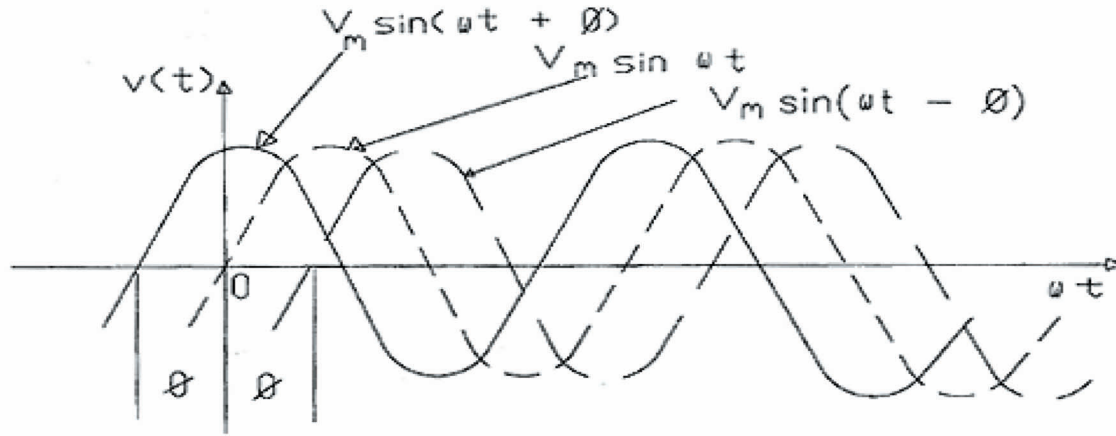
---

A sinusoidal source (independent or dependent) produces a signal that varies sinusoidally with time. In electrical engineering, it is normally a voltage or a current. A sinusoid can be expressed as a sine function or a cosine function. There is no clear-cut choice for the use of either function. However, the sine function is more commonly used. Using a sine function, the instantaneous voltage, which is shown in [Fig. 105.1](#), can be represented by

$$v(t) = V_m \sin \omega t \quad (105.1)$$

where  $V_m$  is the maximum amplitude, and  $\omega$  is the *angular* or *radian frequency* in rad/s.

**Figure 105.1** Three sinusoids with different phase angles.



A sinusoid is a *periodic* function defined generally by the property

$$v(t + T) = v(t) \quad (105.2)$$

where  $T$  is the period, the time for one complete cycle. That is, the function goes through one complete cycle every  $T$  seconds and is then repeated. The period  $T$  is related to the angular frequency  $\omega$  by

$$T = \frac{2\pi}{\omega} \quad (105.3)$$

In 1 s, the function goes through  $1/T$  cycles, or periods. The *frequency* is then

$$f = \frac{1}{T} = \frac{\omega}{2\pi} \quad (105.4)$$

The number of cycles per second, or *hertz* (abbreviated Hz), named for the German physicist Heinrich R. Hertz (1857–1894), is the standard unit of frequency. The frequency and the angular frequency are related by

$$\omega = 2\pi f \quad (105.5)$$

So far, we have assumed that a sinusoid starts at  $t = 0$ . However, it could be phase shifted by an angle  $\phi$  and has the more general expression given by

$$v(t) = V_m \sin(\omega t + \phi) \quad (105.6)$$

where  $\phi$  is the *phase angle*, or simply *phase*. Since  $\omega t$  is in radians,  $\phi$  should also be expressed in radians. However, it is often convenient to specify  $\phi$  in degrees. That is,

$$v(t) = V_m \sin\left(\omega t + \frac{\pi}{3}\right) \quad (105.7)$$

or

$$v(t) = V_m \sin(\omega t + 60^\circ) \quad (105.8)$$

is acceptable, although Eq. (105.8) is mathematically inconsistent. While computing the value of  $\sin(\omega t + \phi)$ , one should use the same units (radians or degrees) for both  $\omega t$  and  $\phi$ . If  $\phi$  has a positive value, the sinusoid is said to have a *leading phase angle*. If  $\phi$  has a negative value, it is said to have a *lagging phase angle*.

## 105.2 Phasor

---

A **phasor** is a complex number that represents the amplitude and phase angle of a sinusoidal function. It transforms a sinusoidal function from time domain to the complex number domain. The phasor concept, which is generally credited to electrical engineer Charles Proteus Steinmetz (1865–1923), is based on the Euler's exponential function of the trigonometric function

$$e^{j\theta} = \cos \theta + j \sin \theta \quad (105.9)$$

Thus, a cosine function may be regarded as the real part of the exponential function and a sine function as the imaginary part of the exponential function. That is,

$$\cos \theta = \operatorname{Re} (e^{j\theta}) \quad (105.10)$$

$$\sin \theta = \operatorname{Im} (e^{j\theta}) \quad (105.11)$$

Therefore, we can write the sinusoidal voltage function of Eq. (105.6) as

$$v(t) = V_m \sin(\omega t + \phi) \quad (105.12)$$

$$= V_m \operatorname{Im} \left\{ e^{j(\omega t + \phi)} \right\} = \operatorname{Im} \left\{ (V_m e^{j\phi}) e^{j\omega t} \right\} \quad (105.13)$$

which indicates that the coefficient of the exponential  $e^{j\omega t}$  is a complex number. It is the *phasor representation*, or *phasor transform*, of the sinusoidal function and is represented by a boldface letter. Thus, the phasor  $\mathbf{V}$  becomes

$$\mathbf{V} = V_m e^{j\phi} \quad (\text{exponential form}) \quad (105.14)$$

$$= V_m \cos \theta + j V_m \sin \theta \quad (\text{polar form}) \quad (105.15)$$

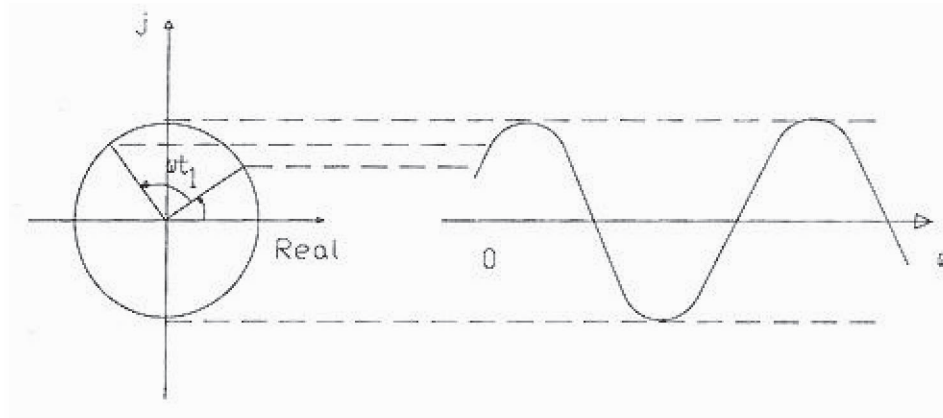
Since phasors are used extensively in analysis of electrical engineering circuits, the exponential function  $e^{j\phi}$  is abbreviated in a shorthand notation for the sake of simplicity

as

$$e^{j\phi} = 1 \angle \phi \quad (105.16)$$

A graphical relationship between a phasor and sinusoid is shown in Fig. 105.2. A unit phasor may be regarded as a unit vector having an initial phase displacement of  $\phi$  and rotating in the counterclockwise direction at an angular speed of  $\omega$ .

**Figure 105.2** Rotating phasor.



## 105.3 Passive Circuit Elements in Phasors Domain

If we want to find the response of an electrical circuit due to a sinusoidal forcing function, first we need to establish the relationship between the phasor voltage across and the phasor current through passive elements such as the resistor, inductor, and capacitor.

### Resistor

Let us assume that a sinusoidal current of

$$i(t) = I_m \sin(\omega t + \theta) \quad (105.17)$$

where  $I_m$  is the maximum amplitude of the current in amperes, and  $\theta$  is the phase angle of the current.

Using Ohm's law for the resistor in Fig. 105.3(a), the instantaneous voltage across the terminals of the resistor is related to its instantaneous current by

$$v(t) = Ri(t) = R \{ I_m \sin(\omega t + \theta) \} = RI_m \sin(\omega t + \theta) \quad (105.18)$$

which can be represented as a phasor by

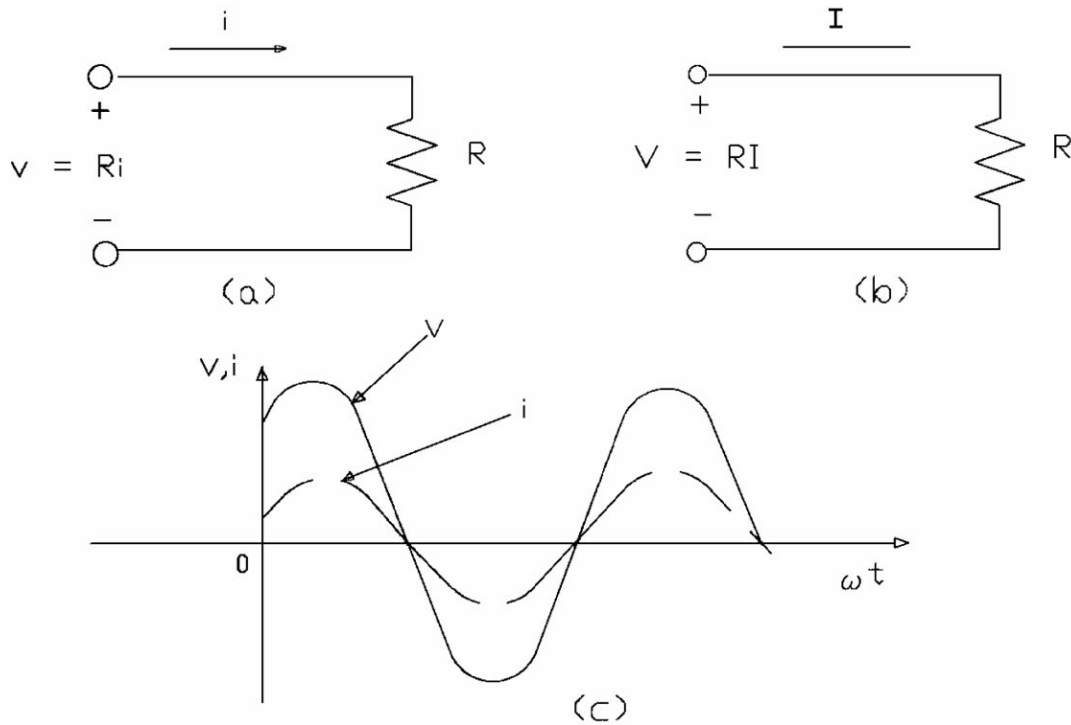
$$V = RI_m e^{j\theta} = RI_m \angle \theta \quad (105.19)$$

Since  $I_m \angle \theta = \mathbf{I}$  is the phasor representation of the current, we can write

$$V = RI \quad (105.20)$$

That is, the phasor voltage across the terminals of a resistor is the resistor times the phasor current. [Figure 105.3\(b\)](#) shows the  $\mathbf{V}$  and  $\mathbf{I}$  relationship of a resistor. A resistor has no phase shift between the voltage and its current. The current is said to be in time phase with the voltage, as shown in [Fig. 105.3\(c\)](#).

**Figure 105.3** Voltage-current relationships of a resistor.



## Inductor

Assuming a sinusoidal current of  $i = I_m \sin(\omega t + \theta)$ , the voltage across an inductor shown in [Fig. 105.4\(a\)](#) is related to its current by

$$v = L \frac{di}{dt} = \omega L I_m \cos(\omega t + \theta) \quad (105.21)$$

Using the trigonometric identity of  $\cos A = \sin(A + \pi/2)$ , Eq. (105.21) can be rewritten as

$$v = \omega L I_m \sin\left(\omega t + \theta + \frac{\pi}{2}\right) \quad (105.22)$$

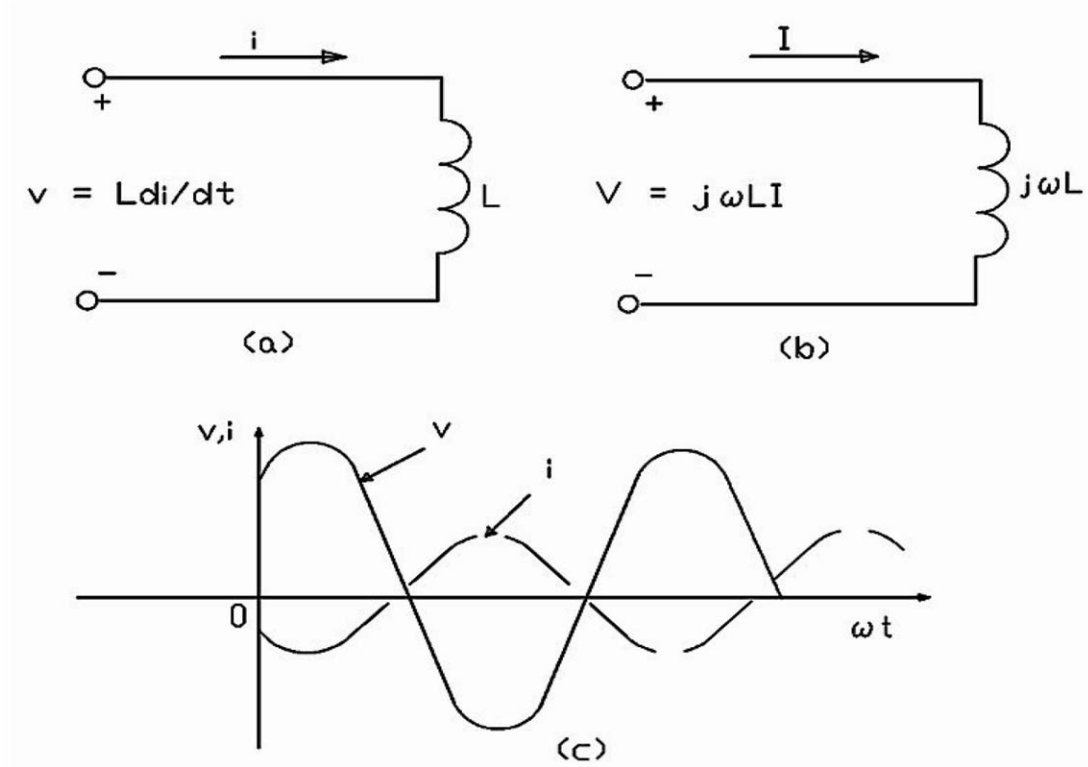
The phasor representation of the voltage given by Eq. (105.22) is



$$\begin{aligned}
 V &= \omega L I_m e^{j(\theta + \pi/2)} = \omega L I_m \angle(\theta + \pi/2) \\
 &= \omega L I_m e^{j\theta} e^{j\pi/2} \\
 &= \omega L I_m e^{j\theta} (\cos \pi/2 + j \sin \pi/2) \\
 &= j\omega L I_m e^{j\theta} = j\omega L I
 \end{aligned}
 \tag{105.23}$$

Thus, the phasor voltage across the terminals of an inductor equals  $j\omega L$  times the phasor current. Since the operator  $j$  gives a phase shift of  $+90^\circ$ , then  $jI = jI \angle \theta = I_m \angle(\theta + \pi/2)$ . That is, the current lags behind the voltage by  $90^\circ$ , or the voltage leads the current by  $90^\circ$ . The voltage and current relationship in the phasor domain is shown in Fig. 105.4(b) and in the time domain in Fig. 105.4(c).

**Figure 105.4** Voltage-current relationships of an inductor.



## Capacitor

Assuming a sinusoidal voltage of  $v = V_m \sin(\omega t + \phi)$ , the current through a capacitor shown in Fig. 105.5(a) is related to its voltage by

$$i = C \frac{dv}{dt} = j\omega C V_m \cos(\omega t + \phi) \tag{105.24}$$

Using the trigonometric identity, Eq. (105.24) can be rewritten as

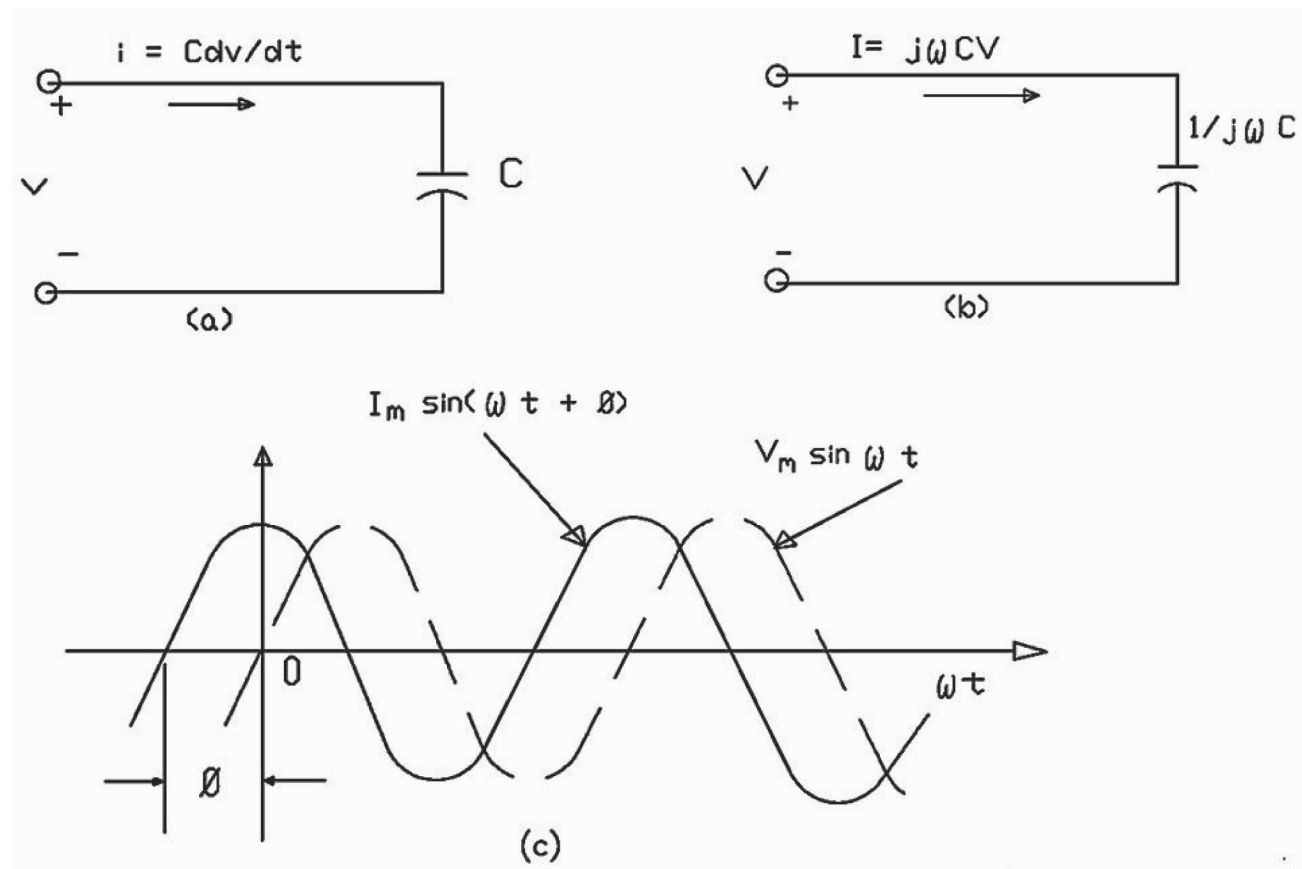
$$i = \omega C V_m \sin\left(\omega t + \phi + \frac{\pi}{2}\right) \tag{105.25}$$

The phasor representation of the current given by Eq. (105.25) is

$$I = j\omega CV \quad (105.26)$$

Thus, the phasor current through a capacitor equals  $j\omega C$  times the phasor voltage. That is, the voltage lags behind the current by  $90^\circ$ , or the current leads the voltage by  $90^\circ$ . The voltage and current relationship in the phasor domain is shown in Fig. 105.5(b) and in the time domain in Fig. 105.5(c).

**Figure 105.5** Voltage-current relationships of a capacitor.



## Sinusoidal Responses

We can conclude from the previous discussions that the phasor relationship between the voltage and the current of an element takes the general form of

$$V = ZI \quad (105.27)$$

where  $Z$  is the **impedance** of the circuit element. That is, the impedance of a resistor is  $R$ , the impedance of an inductor is  $j\omega L$ , and the impedance of a capacitor is  $1/j\omega C$ . Thus, for a circuit having  $L$  and/or  $C$ , the impedance  $Z$  will be a complex number with a real part  $R$  and an imaginary

part  $X$  such that

$$\mathbf{Z} = R + jX = Z\angle\theta \quad (105.28)$$

where

$$Z = [R^2 + X^2]^{1/2} \quad (105.29)$$

and

$$\theta = \tan^{-1}(X/R) \quad (105.30)$$

If the Kirchhoff's voltage law for a set of  $n$  sinusoidal voltages in the time domain is given by

$$v_1 + v_2 + v_3 + v_4 + \cdots + v_n = 0 \quad (105.31)$$

then the equivalent statement in the phasor domain can be written as

$$\mathbf{V}_1 + \mathbf{V}_2 + \mathbf{V}_3 + \mathbf{V}_4 + \cdots + \mathbf{V}_n = 0 \quad (105.32)$$

Similarly, if the Kirchhoff's current law for a set of  $n$  sinusoidal currents in the time domain is given by

$$i_1 + i_2 + i_3 + i_4 + \cdots + i_n = 0 \quad (105.33)$$

then the equivalent statement in the phasor domain can be written as

$$\mathbf{I}_1 + \mathbf{I}_2 + \mathbf{I}_3 + \mathbf{I}_4 + \cdots + \mathbf{I}_n = 0 \quad (105.34)$$

Let us consider the  $RLC$  circuit shown in [Fig. 105.6](#). We will use the phasor concept in finding its current in response to an input source voltage of  $v_s = V_m \sin(\omega t + \phi)$ . We use the phasor notation

$$\mathbf{V}_s = V_m \angle \phi \quad (105.35)$$

Applying the Kirchhoff's voltage law, we get

$$\mathbf{V}_s = (R + j\omega L + 1/j\omega C)\mathbf{I} = \mathbf{Z}\mathbf{I} \quad (105.36)$$

where  $\mathbf{Z}$  is the total impedance of the series loop formed by  $R$ ,  $L$ , and  $C$ . That is,

$$\begin{aligned} Z &= R + j(\omega L - 1/\omega C) = [R^2 + (\omega L - 1/\omega C)^2]^{1/2} \angle \theta \\ &= Z \angle \theta \end{aligned} \quad (105.37)$$

where  $Z$  is the impedance magnitude and is given by

$$Z = [R^2 + (\omega L - 1/\omega C)^2]^{1/2} \quad (105.38)$$

and  $\theta$  is the impedance angle given by

$$\theta = \tan^{-1} \left( \frac{\omega L - 1/\omega C}{R} \right) \quad (105.39)$$

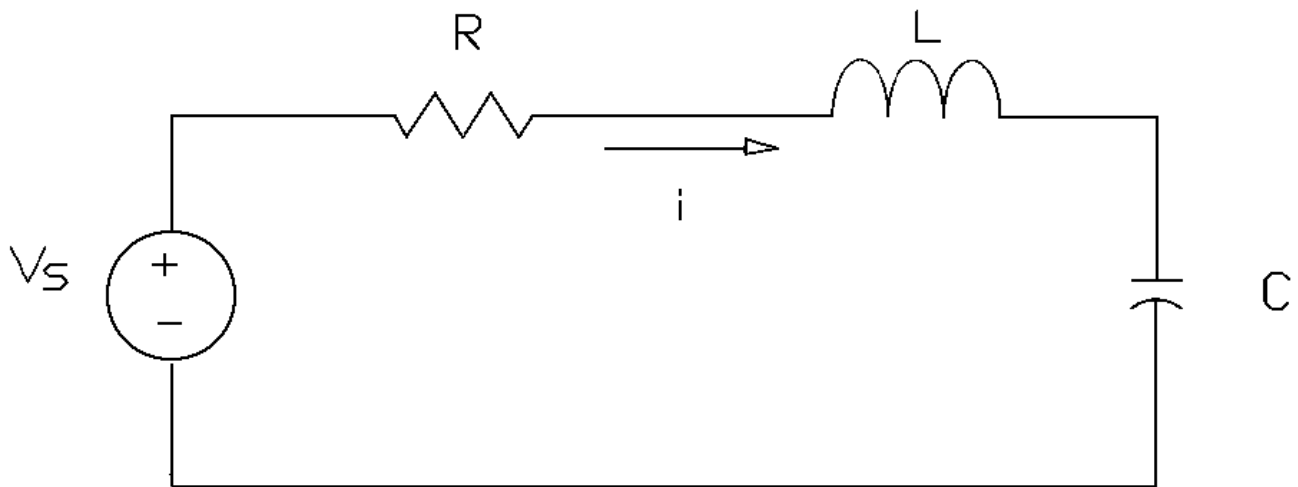
Dividing the phasor voltage by the phasor impedance gives the phasor current as

$$I = \frac{V_s}{Z} = \frac{V_s \angle \phi}{Z \angle \theta} = \frac{V_s \angle \phi - \theta}{Z} \quad (105.40)$$

which indicates that the current lags the input voltage by an angle of  $\theta$ . Converting to the time domain, the current is given by

$$i = \frac{V_m}{\sqrt{R^2 + (\omega L - 1/\omega C)^2}} \sin(\omega t + \phi - \theta) \quad (105.41)$$

**Figure 105.6** A series *RLC* circuit.



Let us consider the parallel *RLC* circuit shown in [Fig. 105.7](#). Applying the Kirchhoff's current

law at the node  $a$ , we get

$$I = I_1 + I_2 + I_3 \quad (105.42)$$

Since the three impedances are:  $Z_R = R$ ,  $Z_1 = R_1 + j\omega L$ , and  $Z_C = 1/j\omega C$ , we can substitute for  $I$  by using Ohm's law. That is,

$$I = \frac{V_s \angle \phi}{R} + \frac{V_s \angle \phi}{R_1 + j\omega L} + \frac{V_s \angle \phi}{1/j\omega C}$$

$$I = V_s \angle \phi \left[ \frac{1}{R} + \frac{1}{\sqrt{R_1^2 + (\omega L)^2}} \angle \theta_1 + j\omega C \right] \quad (105.43)$$

where  $\theta_1$  is the impedance angle of  $R_1$  and  $L$ , and it is given by

$$\theta_1 = \tan^{-1} \frac{\omega L}{R_1} \quad (105.44)$$

Equation (105.42) can be written as

$$I = V_s \angle \phi Y \angle -\theta = V_s Y \angle (\phi - \theta) \quad (105.45)$$

where  $Y \angle -\theta$  is the equivalent admittance of  $R$ ,  $R_1$ , and  $L$  such that it is related to the equivalent impedance  $Z$  by

$$Y \angle -\theta = \frac{1}{Z \angle \theta} = \frac{1}{R} + \frac{1}{R_1 + j\omega L} + j(\omega C) \quad (105.46)$$

Equation (105.43) can be written in the time domain as

$$i(t) = V_m Y \sin(\omega t + \phi - \theta) \quad (105.47)$$

where  $\theta$  is the impedance angle of  $(R_1 + j\omega L)$ , which is in parallel with  $R$  and  $1/j\omega C$ . In converting from the phasor domain to the time domain, we have assumed that the maximum amplitude of a sinusoid is the magnitude of its phasor representation. However, in practice, the magnitude of a sinusoid is normally quoted in its root-mean-square (rms) value, which is defined by

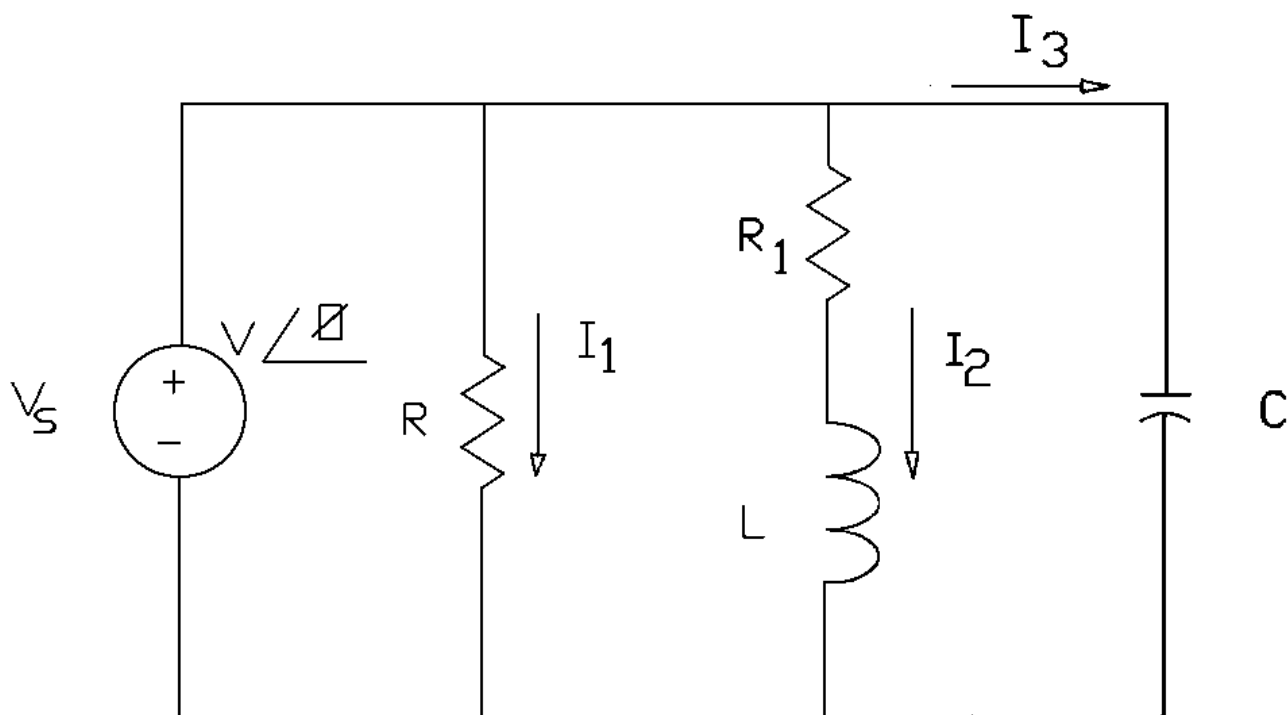
$$V_{\text{rms}} = \sqrt{\frac{1}{T} \int V_m^2 \cos^2(\omega t + \phi) dt} \quad (105.48)$$

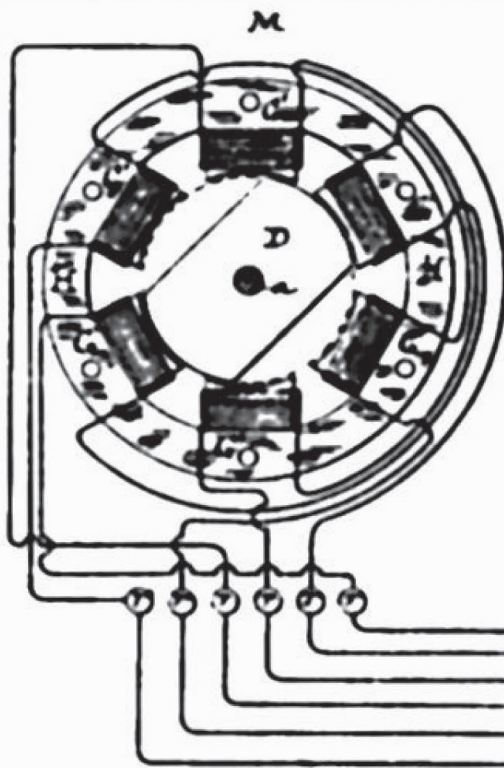
which, after completing the integration and simplification, becomes

$$V_{\text{rms}} = \frac{V_m}{\sqrt{2}} \quad (105.49)$$

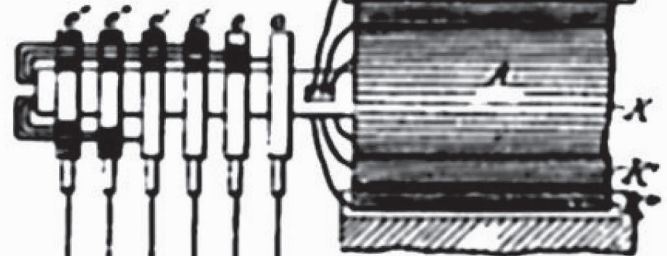
Thus,  $v = V_m \sin(\omega t + \phi) = \sqrt{2}V_{\text{rms}} \sin(\omega t + \phi)$ , and it will be represented in a phasor form by  $V_{\text{rms}} \angle \phi$ . For example,  $170 \sin(\omega t + \phi) \equiv (170/\sqrt{2}) \angle \phi = 120 \angle \phi$ .

**Figure 105.7** A parallel  $RLC$  circuit.

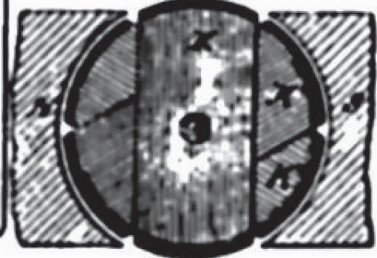




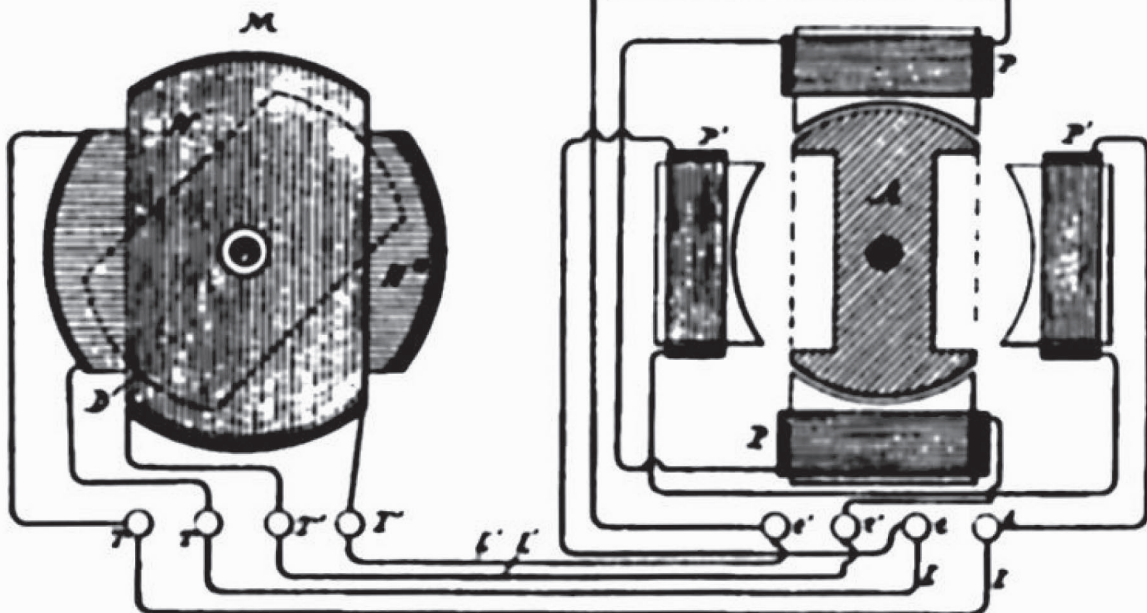
*Fig. 13.*



*Fig. 14.*



*Fig. 15.*



## ELECTRO-MAGNETIC MOTOR

*Nikola Tesla*

*Patented May 1, 1888*

*#381,968*

Tesla was an electrical genius from Serbo-Croatia. He is credited with over 700 inventions; this one perhaps his most important. He invented what is now known as the induction motor. It ran from alternating current (AC) and had no commutating brushes to burn up like the direct current (DC) motors used up to that time. An AC induction motor was inherently synchronous and therefore ran at a constant speed regardless of load; a great advantage in industrial applications.

Westinghouse bought the rights to this invention in July 1888. Through the 1890's, the induction motor's application to industry helped force the eventual world-wide domination by AC power systems which up to that time had been DC systems like those promoted by Edison. (© 1995, DewRay Products, Inc. Used with permission.)

## Defining Terms

**Impedance:** An impedance is a measure of the opposition to a current flow due to a sinusoidal voltage source. It is a complex number and has a real and an imaginary part.

**Phasor:** A phasor is the vector representation of a sinusoid in the complex domain. It represents the amplitude and the phase angle of a sinusoid.

**Sinusoid:** A sinusoid is a sinusoidal time-dependent periodic forcing function which has a maximum amplitude and can also have a phase delay.

## References

Balabania, N. 1994. *Electric Circuits*. McGraw-Hill, New York.

Dorf, R. C. 1993. *Introduction to Electric Circuits*. John Wiley & Sons, New York, NY.

Hayt, W. H., Jr., and Kemmerley, J. E. 1993. *Engineering Circuit Analysis*. McGraw-Hill, New York.

Irwin, D. J. 1989. *Basic Engineering Circuit Analysis*. Macmillan, New York.

Jackson, H. W. 1986. *Introduction to Electric Circuits*. Prentice Hall, Englewood Cliffs, NJ.

Johnson, D. E., Hilbert, J. L., and Johnson, J. R. 1989. *Basic Electric Circuit Analysis*. Prentice Hall, Englewood Cliffs, NJ.

Nilson, J. W. 1993. *Electric Circuits*. Addison-Wesley, Reading, MA.

## Further Information

A general review of linear circuits is presented in *Linear Bircuits* by M. E. Van Valkenburgh, Prentice Hall (Chapters 9–11).

The applications of phasors in determining frequency responses are presented in *Basic Network Theory* by J. Vlach, Van Nostrand Reinhold (Chapter 8).



Many examples of computer-aided simulations by the Industry Standard Circuit Simulator SPICE are given in *SPICE for Circuits and Electronics Using PSpice* by M. H. Rashid, Prentice Hall.

The monthly journal *IEEE Transactions on Circuits and Systems* <sup>3/4</sup>*Fundamental Theory and Applications* reports advances in the techniques for analyzing electrical and electronic circuits. For subscription information contact IEEE Service Center, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331. Phone (800) 678-IEEE.

Balabanian, N. "Three-Phase Circuits"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## 106.1 Relationships between Voltages and Currents

## 106.2 Line Voltages

## 106.3 Power Relationship

## 106.4 Balanced Source and Balanced Load

## 106.5 Other Types of Interconnections

**Norman Balabanian***University of Florida, Gainesville*

A very important use of electricity is the driving of industrial equipment, such as electric motors, in the AC steady state. Suppose that the instantaneous AC voltage and current of such a load is given by

$$\begin{aligned}v(t) &= \sqrt{2}|V| \cos(\omega t + \alpha) \\i(t) &= \sqrt{2}|I| \cos(\omega t + \beta)\end{aligned}\quad (106.1)$$

Then the power to the load at any instant of time is

$$p(t) = |V||I|[\cos(\alpha - \beta) + \cos(2\omega t + \alpha + \beta)] \quad (106.2)$$

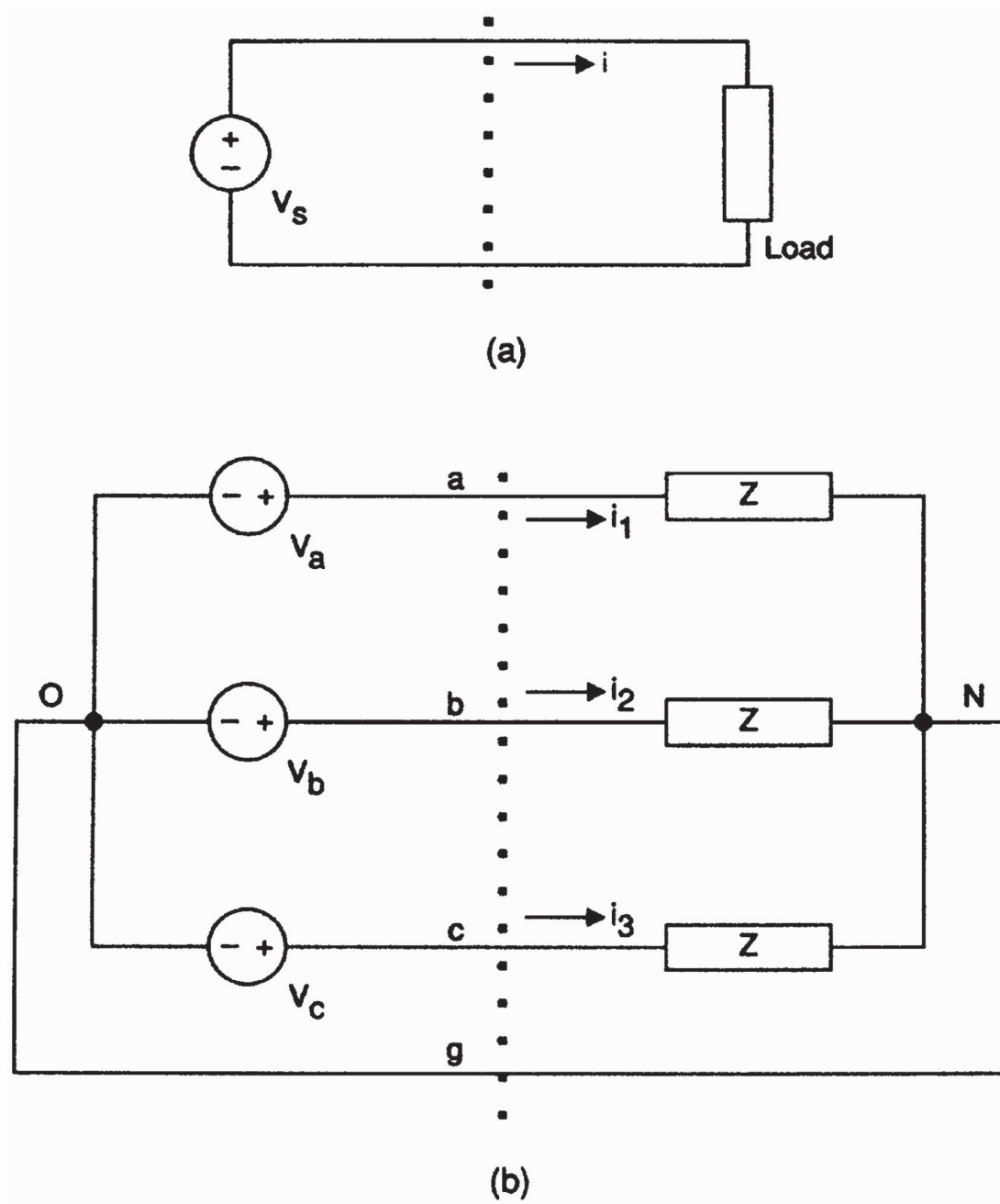
The instantaneous power has a constant term and a sinusoidal term at twice the frequency. The quantity in brackets fluctuates between a minimum value of  $\cos(\alpha - \beta) - 1$  and a maximum value of  $\cos(\alpha - \beta) + 1$ . This fluctuation of power delivered to the load has a great disadvantage when the load is an electric motor. A motor operates by receiving electric power and transmitting mechanical (rotational) power at its shaft. If the electric power is delivered to the motor in spurts, the motor is likely to vibrate. For satisfactory operation in such a case, a physically larger motor, with a larger shaft and flywheel, will be needed to provide more inertia for smoothing out the fluctuations than would be the case if the delivered power were constant.

This problem is overcome in practice by the use of what is called a *three-phase* system. This chapter will provide a brief discussion of three-phase power systems.

Consider the circuit in [Fig. 106.1](#). This is an interconnection of three AC sources to three loads connected in such a way that each source/load combination shares the return connection from O to N. The three sources can be viewed collectively as a single source, and the three loads—which are assumed to be identical—can be viewed collectively as a single load. Each of the individual

sources and loads is referred to as one *phase* of the three-phase system.

**Figure 106.1** Flow of power from source to load.



## 106.1 Relationships between Voltages and Currents

The three sources are assumed to have the same frequency; for this reason, they are said to be *synchronized*. It is also assumed that the three phase voltages have the same rms values and that the phase difference between each pair of voltages is  $\pm 120^\circ$  ( $2\pi/3$  rad). Thus, the voltages can be written:

$$\begin{aligned} v_a &= \sqrt{2}|V| \cos(\omega t + \alpha_1) & \leftrightarrow & V_a = |V|e^{j0^\circ} \\ v_b &= \sqrt{2}|V| \cos(\omega t + \alpha_2) & \leftrightarrow & V_b = |V|e^{-j120^\circ} \\ v_c &= \sqrt{2}|V| \cos(\omega t + \alpha_3) & \leftrightarrow & V_c = |V|e^{j120^\circ} \end{aligned} \quad (106.3)$$

The **phasors** representing the sinusoids have also been shown. For convenience the angle  $v_a$  has been chosen as the reference for angles;  $v_b$  *lags*  $v_a$  by  $120^\circ$  and  $v_c$  *leads*  $v_a$  by  $120^\circ$ . ■

Observe that the principle value of the angle lying between  $\pm 180^\circ$  is used. One could add  $360^\circ$  to the negative angle and use the value  $240^\circ$  instead.

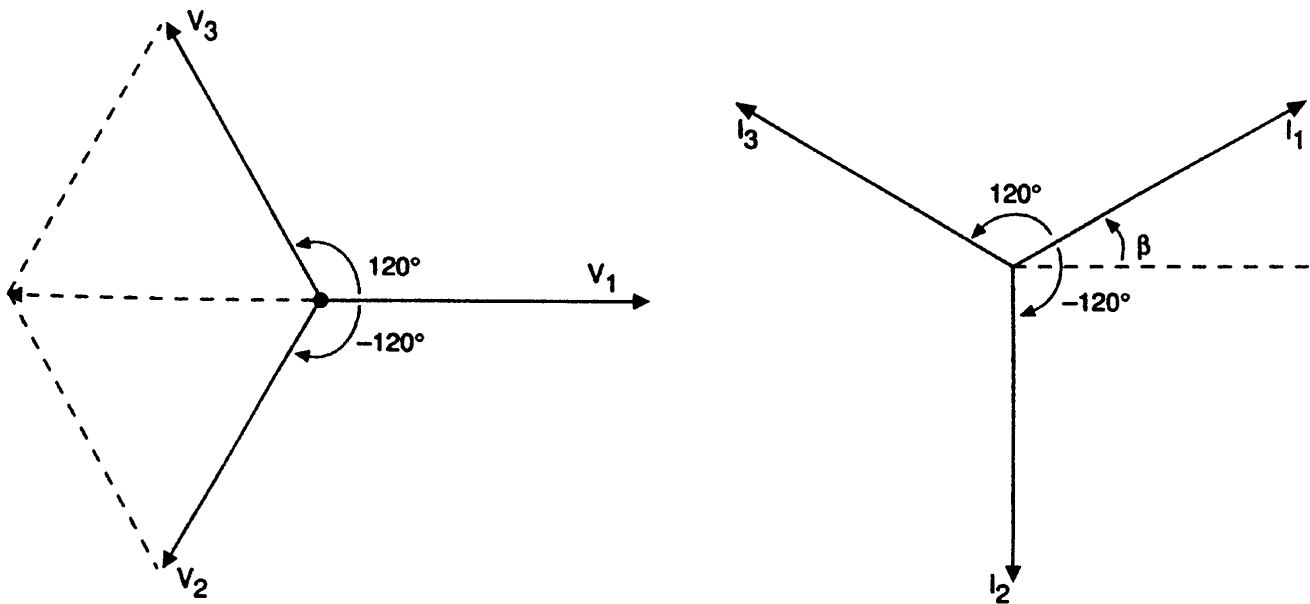
There are two options for choosing the sequence of the phases. Once the particular phase that is to be the reference for angles is chosen and named "a," there are two possible sequences for the other two: either "abc" or "acb." This fact is hardly earthshaking; all it means is that the leading and lagging angles can be interchanged. Obviously, nothing fundamental is different in the second sequence. Hence, the discussion that follows is limited to the *abc phase sequence*.

Because the loads are identical, the rms values of the three currents shown in Fig. 106.1 will also be the same and the phase difference between each pair of them will be  $\pm 120^\circ$ . Thus, the currents can be written as

$$\begin{aligned} i_1 &= \sqrt{2}|I| \cos(\omega t + \beta_1) & \leftrightarrow & I_1 = |I|e^{j\beta_1} \\ i_2 &= \sqrt{2}|I| \cos(\omega t + \beta_2) & \leftrightarrow & I_2 = |I|e^{j(\beta_1 - 120^\circ)} \\ i_3 &= \sqrt{2}|I| \cos(\omega t + \beta_3) & \leftrightarrow & I_3 = |I|e^{j(\beta_1 + 120^\circ)} \end{aligned} \quad (106.4)$$

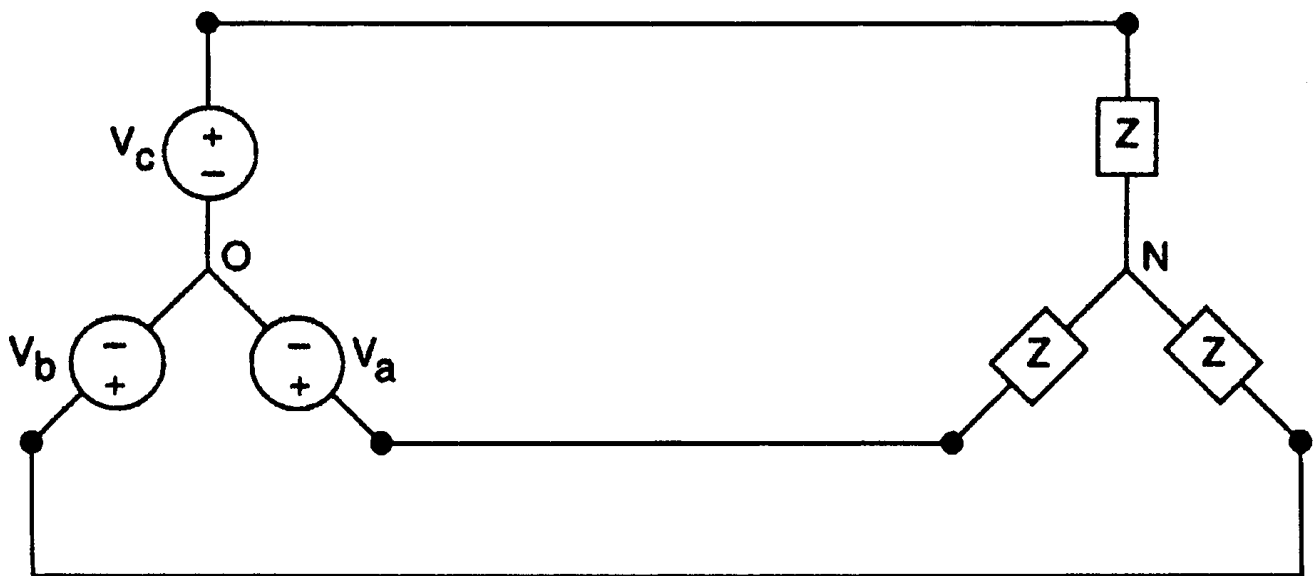
Perhaps a better form of visualizing the voltages and currents is a graphical one. Phasor diagrams for the voltages and the currents are shown separately in Fig. 106.2. The value of angle  $\beta_1$  will depend on the load. Something significant is clear from these diagrams. First,  $V_2$  and  $V_3$  are each the other's conjugate. So if they are added, the imaginary parts cancel and the sum will be real, as illustrated by the construction in the voltage diagram. Furthermore, the construction shows this sum to be negative and equal in magnitude to  $V_1$ . Hence, *the sum of the three voltages is zero*. The same is true of the sum of the three currents, as can be established graphically by a similar construction. The same results can be confirmed analytically by converting the phasor voltages and currents into rectangular form.

**Figure 106.2** Voltage and current phasor diagrams.



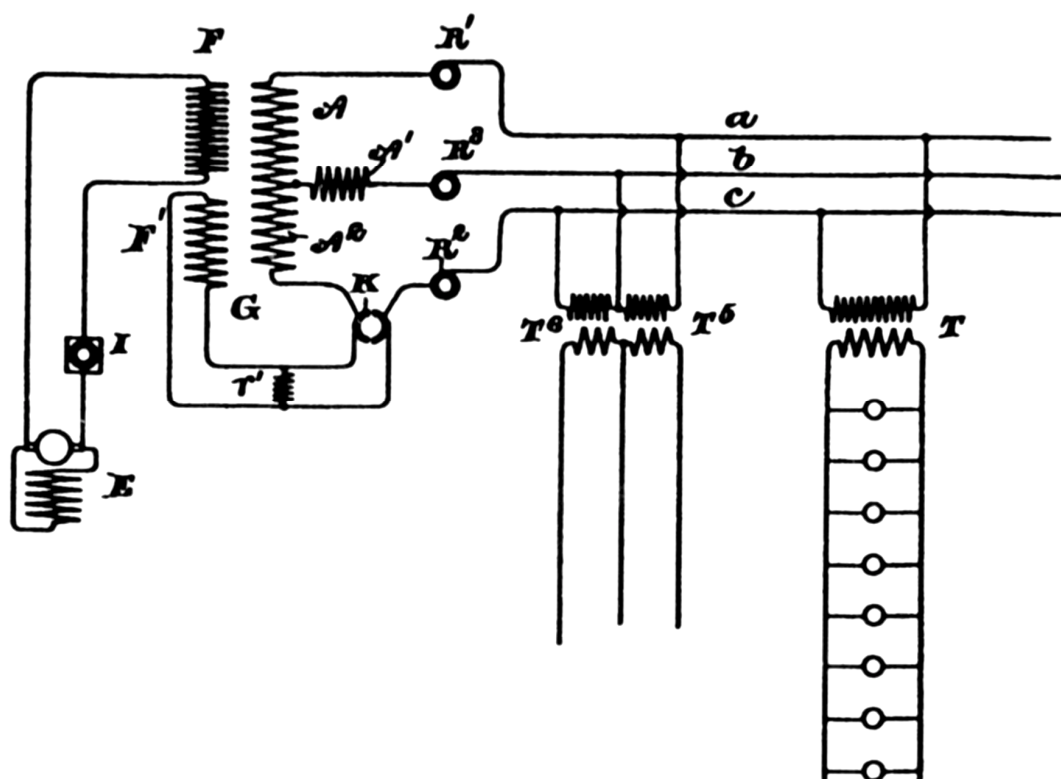
By Kirchhoff's current law applied at node N in Fig. 106.1, we find that the current in the return line is the sum of the three currents in Eq. (106.4). But since this sum was found to be zero, *the return line carries no current*. Hence, it can be removed entirely without affecting the operation of the system. The resulting circuit is redrawn in Fig. 106.3. It can be called a *three-wire* three-phase system. Because of its geometrical form, this connection of both the sources and the loads is said to be a **wye (Y) connection**, even though it is an upside-down Y.

**Figure 106.3** Wye-connected three-phase system.



Notice that the circuit in Fig. 106.3 is planar, with no lines crossing any other lines. That simplicity has been achieved at a price. Notice how the sequence (abc) of sources has been laid out

geometrically. Clearly, with the connections shown, the sequence of the loads is not the same as that of the sources. Having the same sequence would require interchanging the connections of the b and c sources with the bottom two loads. Doing that would result in one branch crossing another. However, nothing fundamental would change with either connection, assuming equal loads.



# SYSTEM OF DISTRIBUTION BY ALTERNATING CURRENTS

Charles P. Steinmetz

Patented January 29, 1895

#533,244

By the 1890's, Edison's light bulb had taken over the lighting of big cities, but his high-voltage DC method of supplying power had not. Alternating current (AC) was becoming the standard power source. It was cheaper and more efficient to transmit AC over long distances because it could be transformed up to high voltages and low currents (thinner wires, less losses) for transmission and then transformed back down to lower voltages for household current.

Steinmetz, while at General Electric's laboratory in Schenectady, New York, developed the system for 3-phase AC. This permitted polyphase industrial motors and machines to operate on the same system as single-phase electric light. This system of power distribution is substantially the same one in use throughout the world today. (©1993, DewRay Products, Inc. Used with permission.)

## 106.2 Line Voltages

---

In the three-wire three-phase system in Fig. 106.3, the neutral point O is not accessible, so phase voltages cannot be measured. The voltages that *are* available for measurement are the *line-to-line* or simply the *line* voltages:  $V_{ab}$ ,  $V_{bc}$ , and  $V_{ca}$ . By Kirchhoff's voltage law,

$$\begin{aligned} V_{ab} &= V_a - V_b = |V| - |V|e^{-j120^\circ} = \sqrt{3}|V|e^{j30^\circ} \\ V_{bc} &= V_b - V_c = |V|e^{-j120^\circ} - |V|e^{j120^\circ} = \sqrt{3}|V|e^{-j90^\circ} \\ V_{ca} &= V_c - V_a = |V|e^{j120^\circ} - |V| = \sqrt{3}|V|e^{j150^\circ} \end{aligned} \quad (106.5)$$

The interesting result is that all the line-voltage magnitudes are equal at  $\sqrt{3}$  times the phase-voltage magnitude. Thus, a 220 V line voltage corresponds to a phase voltage of 127 V. The line-voltage angles have the same mutual relationships as the phase-voltage angles; they are separated by  $\pm 120^\circ$ .

## 106.3 Power Relationship

---

The instantaneous power delivered by each of the sources has the form given in Eq. (106.2), consisting of a constant term representing the average power and a double-frequency sinusoidal term. The latter, being sinusoidal, can be represented by a phasor also. The only caveat is that a different frequency is involved here, so this power phasor should not be mixed with the voltage and current phasors in the same diagram or calculations. Let  $|S| = |V||I|$  be the apparent power delivered by each of the three sources, and let the three power phasors be  $S_a$ ,  $S_b$ , and  $S_c$ , respectively. Then,

$$\begin{aligned} S_a &= |S|e^{j(\alpha_1 + \beta_1)} = |S|e^{j\beta_1} \\ S_b &= |S|e^{j(\alpha_2 + \beta_2)} = |S|e^{j(-120^\circ + \beta_1 - 120^\circ)} = |S|e^{j(\beta_1 + 120^\circ)} \\ S_c &= |S|e^{j(\alpha_3 + \beta_3)} = |S|e^{j(120^\circ + \beta_1 + 120^\circ)} = |S|e^{j(\beta_1 - 120^\circ)} \end{aligned} \quad (106.6)$$

It is evident that the phase relationships between these three phasors are the same as the ones between the voltages and the currents. That is, the second leads the first by  $120^\circ$  and the third lags the first by  $120^\circ$ . Hence, just as with the voltages and the currents, the sum of these three power phasors will also be zero. This is a very significant result. It constitutes the motivation for using three-phase power over the pulsating power of a single-phase system. Although the instantaneous power delivered by each load has a constant component and a sinusoidal component, when the



three powers are added, the sinusoidal components add to zero, leaving only the constants. Thus, the total power delivered to the three loads is constant.

To determine the value of this constant power, let's use Eq. (106.2) as a model. The contribution of the  $k$ th source to the total (constant) power is  $|S| \cos(\alpha_k - \beta_k)$ . It can be easily verified that  $\alpha_k - \beta_k = \alpha_1 - \beta_1 = -\beta_1$ . The first equality follows from the relationships between the  $\alpha$  values from Eq. (106.3) and between the  $\beta$  values from Eq. (106.4). The choice of  $\alpha_1 = 0$  leads to the last equality. Hence, each phase contributes an equal amount to the total average power. If  $P$  is the total average power, then

$$P = P_a + P_b + P_c = 3P_a = 3|V||I| \cos(\alpha_1 - \beta_1) \quad (106.7)$$

Although the angle  $\alpha_1$  has been set equal to zero, it is shown in this equation for the sake of generality.

A similar result can be obtained for the reactive power. The reactive power of the  $k$ th phase is  $|S| \sin(\alpha_k - \beta_k) = |S| \sin(\alpha_1 - \beta_1)$ . If  $Q$  is the total reactive power, then

$$Q = 3|S| \sin(\alpha_1 - \beta_1)$$

## 106.4 Balanced Source and Balanced Load

What has just been described is a *balanced* three-phase three-wire power system. The three sources in practice are not three independent sources but consist of three different parts of the same generator. The same is true of the loads. ■

An AC power generator consists of (a) a *rotor* that is rotated by a *prime mover* (say a turbine) and produces a magnetic field that also rotates, and (b) a *stator* on which is wound one or more coils of wire. In three-phase systems the number of coils is three. The rotating magnetic field induces a voltage in each of the coils. The frequency of the induced voltage depends on the number of magnetic poles created on the rotor and the speed of rotation. These are fixed so as to "synchronize" with the 60 Hz frequency of the power system. The 120° leading and lagging phase relationships between these voltages is obtained by distributing the conductors of the coils around the circumference of the stator so that they are separated geometrically by 120°. Thus, the three sources described in the text are in reality a single physical device, a single generator. Similarly, the three loads might be the three windings on a three-phase motor, again a single physical device. Or they might be the windings of a three-phase transformer.

What has been described is ideal in a number of ways. First, the circuit can be *unbalanced* for example, by the loads being somewhat unequal. Second, since the real devices whose ideal model is a voltage source are coils of wire, each source should be accompanied by a branch consisting of the coil inductance and resistance. Third, since the power station (or the distribution transformer at some intermediate point) may be at some distance from the load, the parameters of the physical line carrying the power (the line inductance and resistance) must also be inserted in series between the source and the load.

The analysis of this chapter does not apply to an unbalanced system. An entirely new analytical

technique is required to do full justice to such a system.■

The technique for analyzing unbalanced circuits utilizes what are called *symmetrical components*.

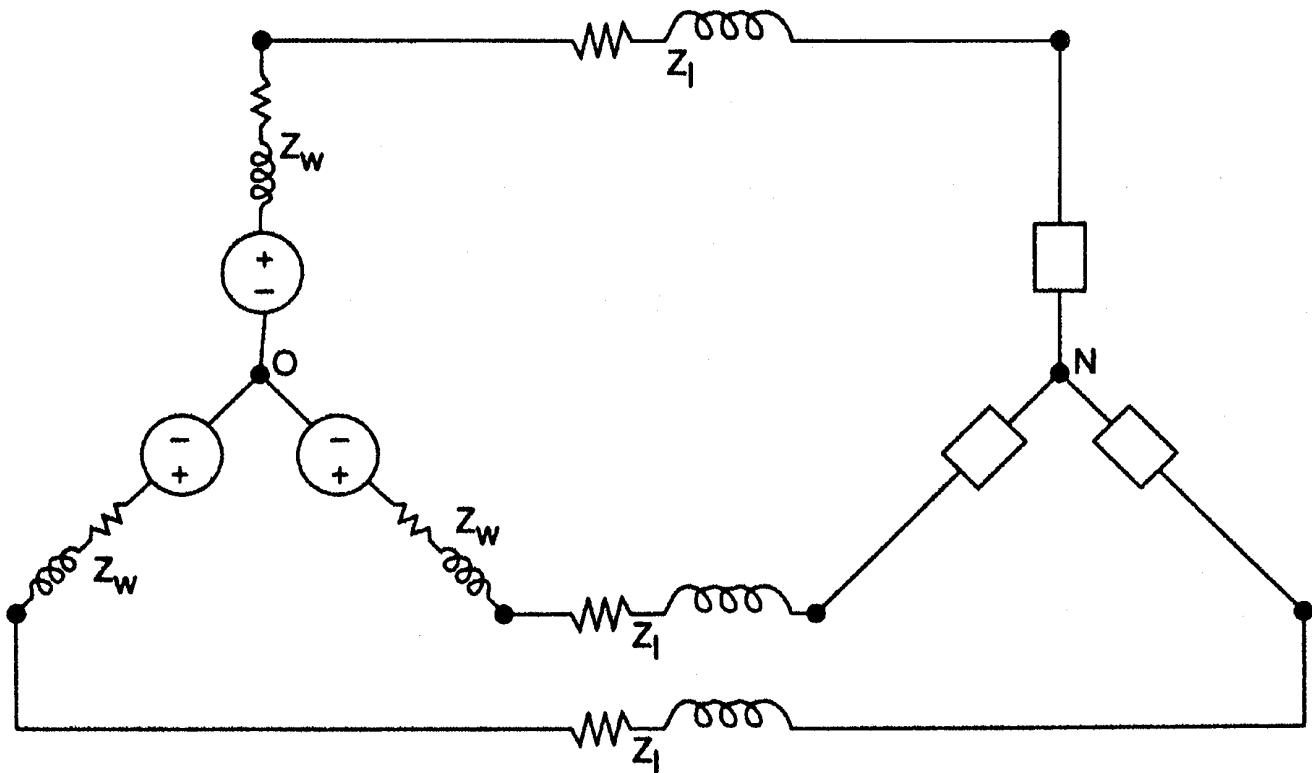
An understanding of balanced circuits is a prerequisite to tackling the unbalanced case.

The last two of the conditions that make the circuit less than ideal (winding and line impedances) introduce algebraic complications but change nothing fundamental in the preceding theory. If these two conditions are taken into account, the appropriate circuit takes the form shown in Fig. 106.4. Here the internal impedance of a source (the winding impedance labeled  $Z_w$ ) and the line impedance  $Z_l$  connecting that source to its load are both connected in series with the corresponding load. Thus, instead of the impedance in each phase being  $Z$ , it is  $Z + Z_w + Z_l$ . Hence, the rms value of each current is

$$|I| = \frac{|V|}{|Z + Z_w + Z_l|} \quad (106.8)$$

instead of  $|V|/|Z|$ . All other previous results remain unchanged—namely, that the sum of the phase currents add to zero and that the sum of the phase powers is a constant. The detailed calculations just become a little more complicated.

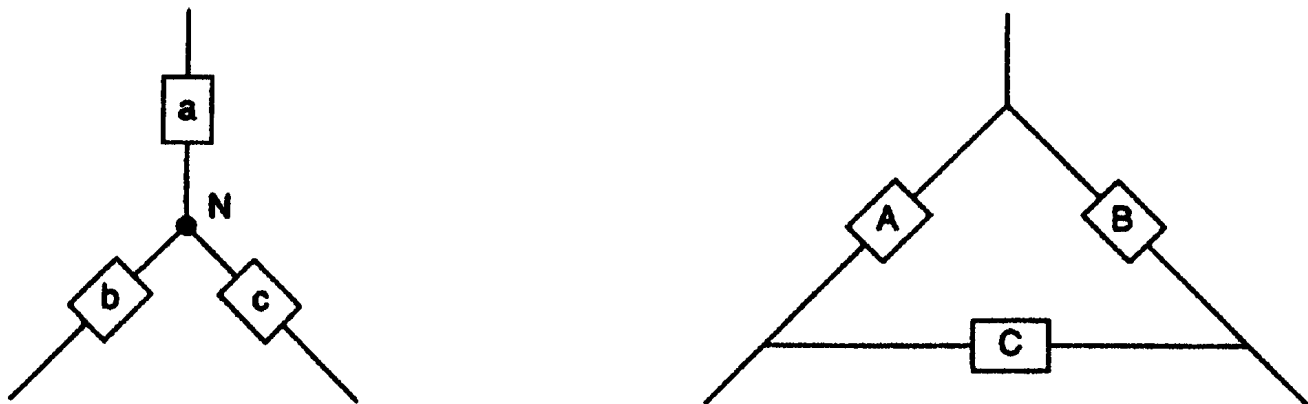
**Figure 106.4** Three-phase circuit with nonzero winding and line impedances.



## 106.5 Other Types of Interconnections

All of the preceding development was based on both the sources and the loads being connected in a wye connection. Although the upside-down Y structure looks geometrically a little different from an upside-down tee circuit, electrically, the two are exactly the same. The wye is not, however, the only possible way to connect the phases of a three-phase system. Another possibility, the **delta connection**, so named because it looks like the Greek letter  $\Delta$ , is shown in Fig. 106.5. (In this figure the boxes can represent either sources or impedances.)

**Figure 106.5** Wye connection and delta connection.



By proper choice of the branch parameters, a tee can be made equivalent to a pi ( $\Pi$ ) at the terminals. We note that the delta is just an upside-down pi. As a pi, the junction between A and B is usually extended as the common terminal of a two-port.

If the structures in Fig. 106.5 are to be equivalent, the line voltages  $V_{ab}$ ,  $V_{bc}$ , and  $V_{ca}$  should be the same in both circuits. Similarly, the currents into the terminals should be the same in both. Note that, in the delta connection, the phase voltages are not evident; the only voltages available are the line voltages. Thus the voltages in the delta are the line voltages given in Eq. (106.3). In the wye the phase currents are also the currents in the lines. For the delta, however, the line currents are the difference between two phase currents, as noted in Fig. 106.5. For the line currents, a set of equations similar to Eq. (106.5) can be written in terms of the phase currents. Since the same  $120^\circ$  difference of angle exists between the phase currents as between the phase voltages, we would expect that the result for currents would be similar to the result for voltages in Eq. (106.5)—namely, that the line-current magnitudes in a delta connection would be  $\sqrt{3}$  times the phase-current magnitudes.

In a three-phase circuit the sources, the loads, or both, can be replaced by a delta equivalent; four different three-phase circuits can therefore be imagined: wye-wye, wye-delta, delta-wye, and delta-delta. There are no fundamental differences in analyzing each of these four circuits.

**Example.** A balanced, 120 V, three-wire three-phase transmission system in a wye-wye connection is represented by the circuit in Fig. 106.4. Assume that the winding impedances are

negligible but that the line impedances are given by  $Z_l = 0.1 + j0.2$  . Each load impedance is  $Z = 20 + j5$  . The following quantities are to be determined: (a) the line current magnitude; (b) the magnitude of the voltage across each load; (c) the average power, reactive power, and apparent power delivered to the load by each phase; (d) the average power, reactive power, and apparent power delivered by each source; and (e) the fraction of the power delivered by the system that is lost in the lines.

**Solution.** The solution is completely straightforward. First, the line current is found by dividing the phase voltage by the sum of the load and line impedances; the load voltage follows from the product of the load impedance by the line current. Thus,

$$|I| = \frac{120}{\sqrt{(20 + 0.1)^2 + (5 + 0.2)^2}} = 5.78 \text{ A}$$

$$|V_L| = |I||Z| = 5.78\sqrt{20^2 + 5^2} = 119.16 \text{ V}$$

The power calculations then follow:

$$|S_L| = |V_L||I| = 119.16(5.78) = 688.7 \text{ VA or}$$

$$|S_L| = |I|^2|Z_L| = 5.78^2\sqrt{20^2 + 5^2} = 688.7 \text{ VA}$$

$$P_L = R_L|I|^2 = 20(5.78)^2 = 668.2 \text{ W}$$

$$Q_L = X_L|I|^2 = 5(5.78)^2 = 167.0 \text{ VAR}$$

$$= \sqrt{|S_L|^2 - P_L^2} = \sqrt{688.7^2 - 668.2^2} = 166.8 \text{ VAR}$$

Perhaps the best way to find the power delivered by the sources is to determine the power lost in the line and then add this to the load power. Carrying out this approach leads to the following result:

$$P_l = 0.1|I|^2 = 3.34 \text{ W} \qquad P_s = 3.34 + 668.2 = 671.5 \text{ W}$$

$$Q_l = 0.2|I|^2 = 6.68 \text{ VAR} \qquad Q_s = 6.68 + 167.0 = 173.7 \text{ VAR}$$

Finally, the fraction of the source power that is lost in the line is  $3.34/671.5 = 0.005$  or 0.5%.

## Defining Terms

**Delta connection:** The sources or loads in a three-phase system connected end-to-end, forming a closed path, like the Greek letter  $\Delta$ .

**Phasor:** A complex number representing a sinusoid; its magnitude and angle are the rms value and the phase of the sinusoid, respectively.

**Wye connection:** The three sources or loads in a three-phasor system connected to have one common point, like the letter Y.

## References

- del Toro, V. 1992. *Electric Power Systems*. Prentice Hall, Englewood Cliffs, NJ.
- Gungor, B. R. 1988. *Power Systems*. Harcourt Brace Jovanovich, San Diego, CA.
- Peebles, P. Z. and Giuma, T. A. 1991. *Principles of Electrical Engineering*. McGraw-Hill, New York.

Rosa, A. J. "Filters (Passive)"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

[107.1 Fundamentals](#)[107.2 Applications](#)Simple RL and RC Filters • Compound Filters • Constant- $k$  Filters •  $m$ -Derived Filters**Albert J. Rosa***University of Denver*

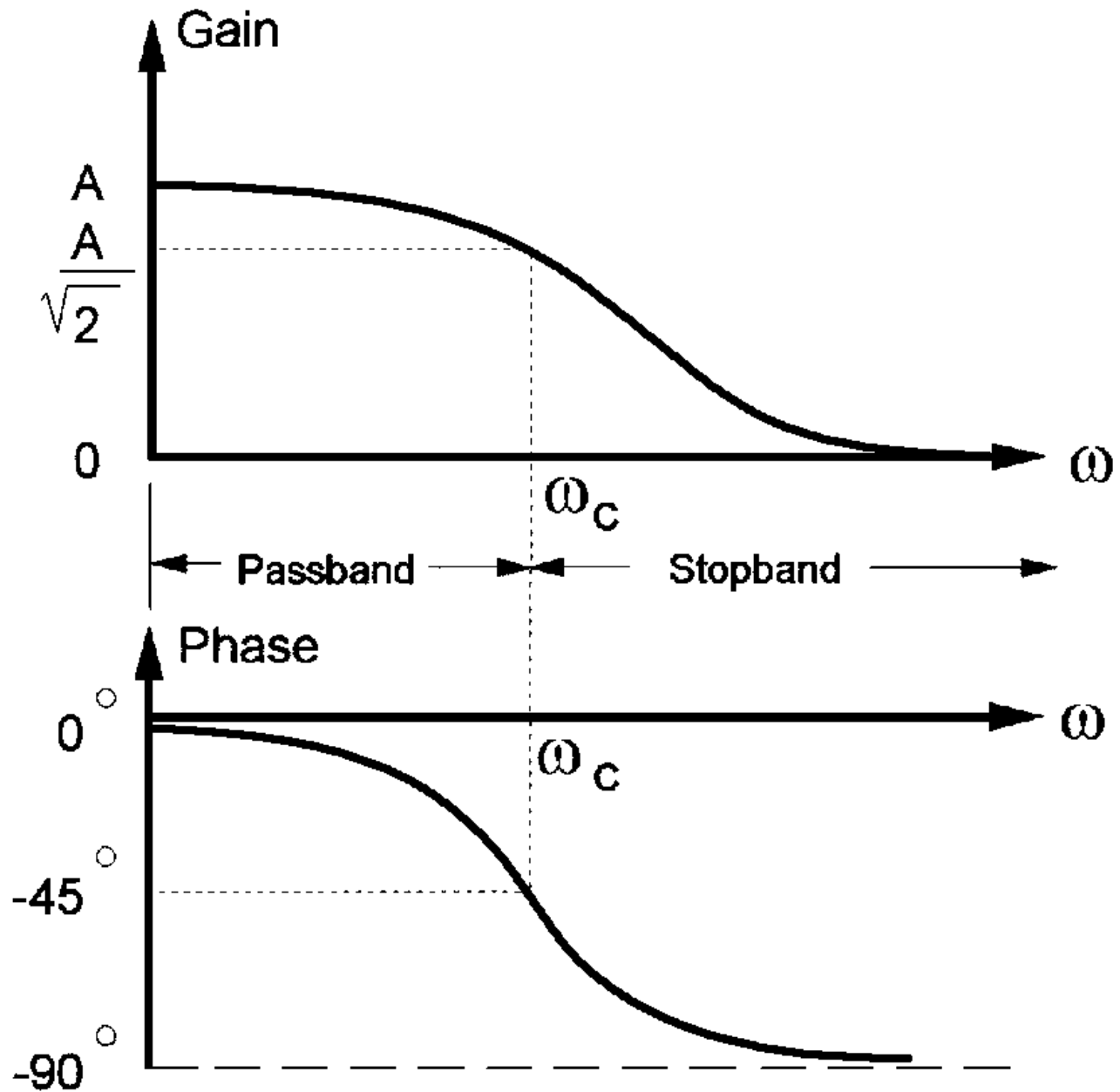
A **filter** is a frequency-sensitive two-port circuit that transmits signals in a band of frequencies and rejects (or attenuates) signals in other bands. The electric filter was invented during the First World War by two engineers working independently of each other—the American engineer G. A. Campbell and the German engineer K. W. Wagner. These devices were developed to serve the growing areas of telephone and radio communication. Today, filters are found in all types of electrical and electronic applications from power to communications. Filters can be both active and passive. In this section we will confine our discussion to those filters that employ no active devices for their operation. The main advantage of passive filters over active ones is that they require no power (other than the signal) to operate.

---

## 107.1 Fundamentals

The basis for filter analysis involves the determination of a filter circuit's sinusoidal steady state response from its transfer function  $T(j\omega)$ . [Some references use  $H(j\omega)$  for the transfer function.] The filter's transfer function  $T(j\omega)$  is a complex function and can be represented through its gain  $|T(j\omega)|$  and phase  $\angle T(j\omega)$  characteristics. The gain and phase responses show how the filter alters the amplitude and phase of the input signal to produce the output response. The two characteristics of the filter's transfer function can be used to describe its frequency response. The terminology used to describe the gain and phase characteristics shows how the circuit modifies the input amplitude and phase angle to produce the output sinusoid. The two characteristics describe the *frequency response* of the circuit since they depend on the frequency of the input sinusoid. The signal-processing performance of devices, circuits, and systems is often specified in terms of frequency response. The gain and phase functions can be expressed mathematically or graphically as *frequency-response* plots. [Figure 107.1](#) shows examples of gain and phase responses versus frequency,  $\omega$ .

**Figure 107.1** Low-pass filter characteristics showing passband, stopband, and the cutoff frequency,  $\omega_C$  .



The terminology used to describe the frequency response of circuits and systems is based on the form of the gain plot. For example, at high frequencies the gain in [Fig. 107.1](#) falls off so that output signals in this frequency range are reduced in amplitude. The range of frequencies over which the output is significantly attenuated is called the *stopband*. At low frequencies the gain is essentially constant and there is relatively little attenuation. The frequency range over which there is little attenuation is called a *passband*. The frequency associated with the boundary between a passband and an adjacent stopband is called the *cutoff frequency* ( $\omega_C = 2\pi f_C$ ) . In general, the transition from the passband to the stopband, called the *transition band*, is gradual, so the precise location of

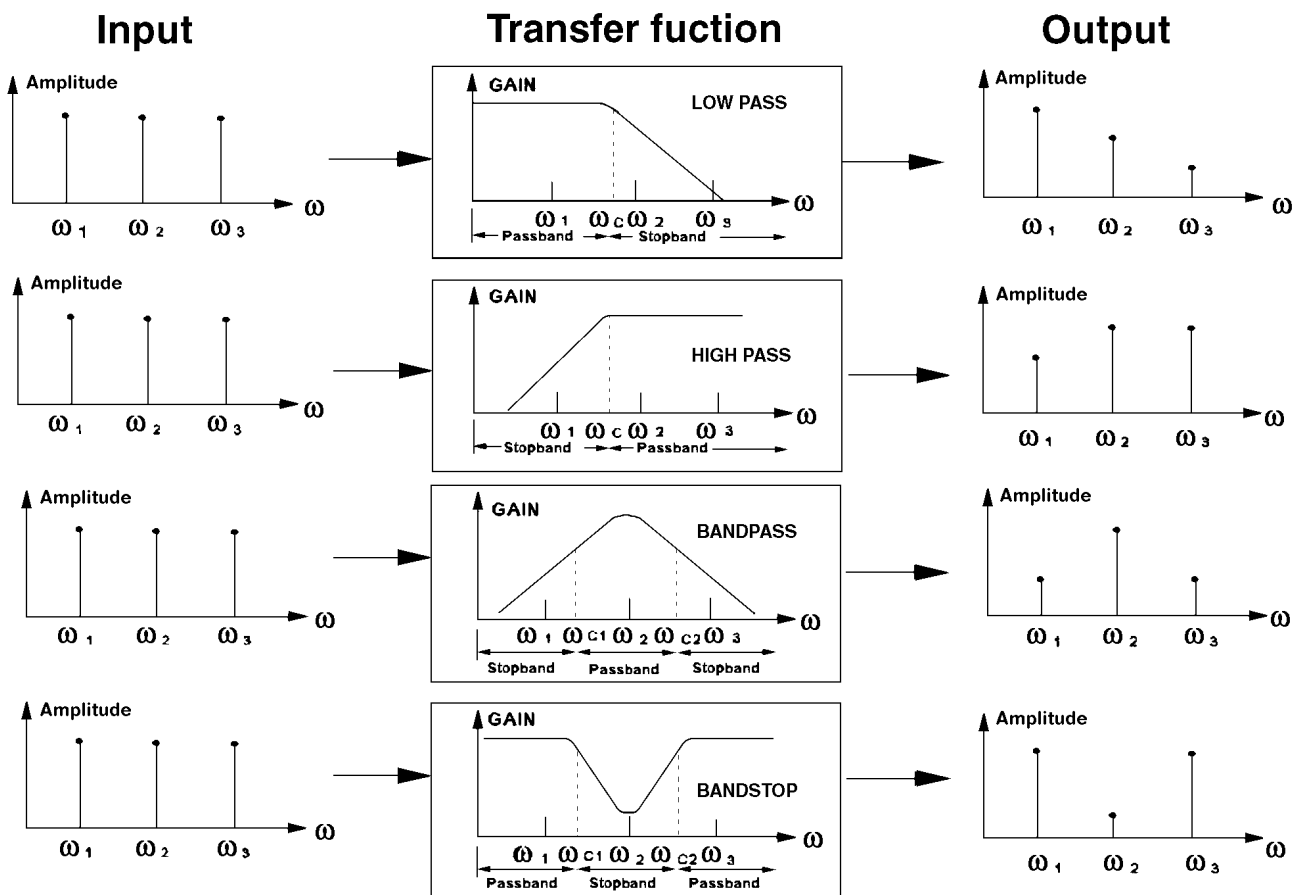


the cutoff frequency is a matter of definition. The most widely used approach defines the cutoff frequency as the frequency at which the gain has decreased by a factor of  $1/\sqrt{2} = 0.707$  from its maximum value in the passband.

This definition is based on the fact that the power delivered to a resistance by a sinusoidal current or voltage waveform is proportional to the square of its amplitude. At a cutoff frequency the gain is reduced by a factor of  $1/\sqrt{2}$  and the square of the output amplitude is reduced by a factor of one half. For this reason the cutoff frequency is also called the *half-power frequency*.

There are four prototypical filters. These are *low pass* (LP), *high pass* (HP), *band pass* (BP), and *bandstop* (BS). Figure 107.2 shows how the amplitude of an input signal consisting of three separate frequencies is altered by each of the four prototypical filter responses. The low-pass filter passes frequencies below its cutoff frequency  $\omega_C$ , called its *passband*, and attenuates the frequencies above the cutoff, called its *stopband*. The high-pass filter passes frequencies above the cutoff frequency  $\omega_C$  and attenuates those below. The band-pass filter passes those frequencies that lie between two cutoff frequencies,  $\omega_{C1}$  and  $\omega_{C2}$ , its *passband*, and attenuates those frequencies that lie outside the passband. Finally, the band-reject filter attenuates those frequencies that lie in its reject or stopband, between  $\omega_{C1}$  and  $\omega_{C2}$ , and passes all others.

**Figure 107.2** Four prototype filters and their effects on an input signal consisting of three frequencies.



The *bandwidth* of a gain characteristic is defined as the frequency range spanned by its passband. For the band-pass case in Fig. 107.2 the bandwidth is the difference in the two cutoff frequencies.

$$BW = \omega_{C2} - \omega_{C1} \quad (107.1)$$

This equation applies to the low-pass response with the lower cutoff frequency  $\omega_{C1}$  set to zero. In other words, the bandwidth of a low-pass circuit is equal to its cutoff frequency ( $BW = \omega_C$ ). The bandwidth of a high-pass characteristic is infinite since the upper cutoff frequency  $\omega_{C1}$  is infinity. For the band-stop case, Eq. (107.1) defines the bandwidth of the stopband rather than the passbands.

Frequency-response plots are usually made using logarithmic scales for the frequency variable because the frequency ranges of interest often span several orders of magnitude. A logarithmic frequency scale compresses the data range and highlights important features in the gain and phase responses. The use of a logarithmic frequency scale involves some special terminology. A frequency range whose end points have a 2:1 ratio is called an *octave* and one with a 10:1 ratio is called a *decade*.

In frequency-response plots the gain  $|T(j\omega)|$  is often expressed in *decibels* (dB), defined as

$$|T(j\omega)|_{dB} = 20 \log_{10} |T(j\omega)| \quad (107.2)$$

Gains expressed in decibels can be either positive, negative, or zero. A gain of zero dB means that  $|T(j\omega)| = 1$  —that is, the input and output amplitudes are equal. A positive dB gain means the output amplitude exceeds the input since  $|T(j\omega)| > 1$ , whereas a negative dB gain means the output amplitude is smaller than the input since  $|T(j\omega)| < 1$ . A cutoff frequency usually occurs when the gain is reduced from its maximum passband value by a factor  $1/\sqrt{2}$  or 3 dB.

Figure 107.3 shows the asymptotic gain characteristics of ideal and real low-pass filters. The gain of the *ideal filter* is unity (0 dB) throughout the passband and zero ( $-\infty$  dB) in the stopband. It also has an infinitely narrow transition band. The asymptotic gain responses of real low-pass filters show that we can only approximate the ideal response. As the order of the filter or number of poles  $n$  increases, the approximation improves since the asymptotic slope in the stopband is  $-20 \times n$  dB/decade. On the other hand, adding poles requires additional stages in a cascade realization, so there is a trade-off between (1) filter complexity and cost and (2) how closely the filter gain approximates the ideal response.

**Figure 107.3** The effect of increasing the order  $n$  of a filter relative to an ideal filter.

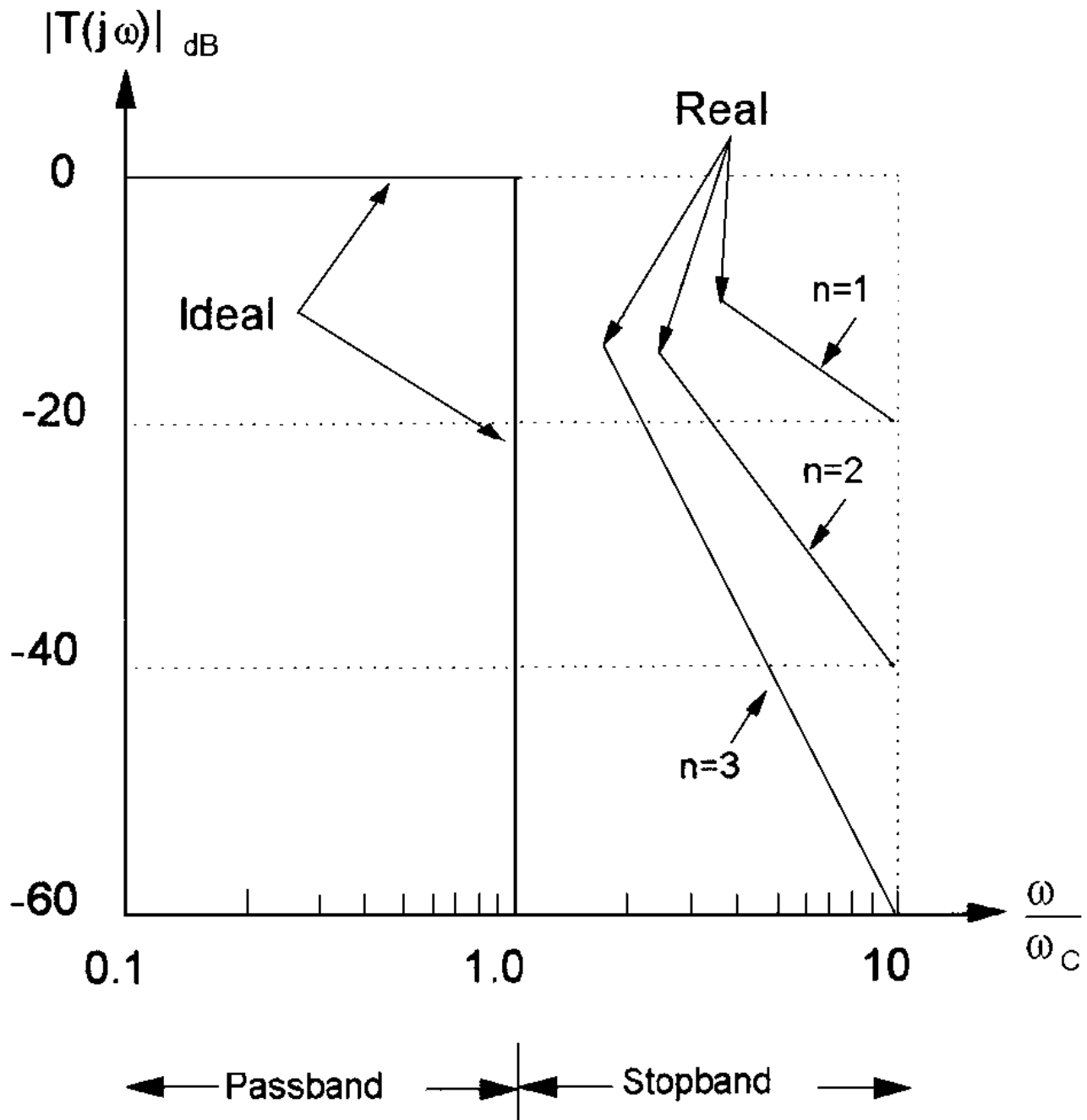
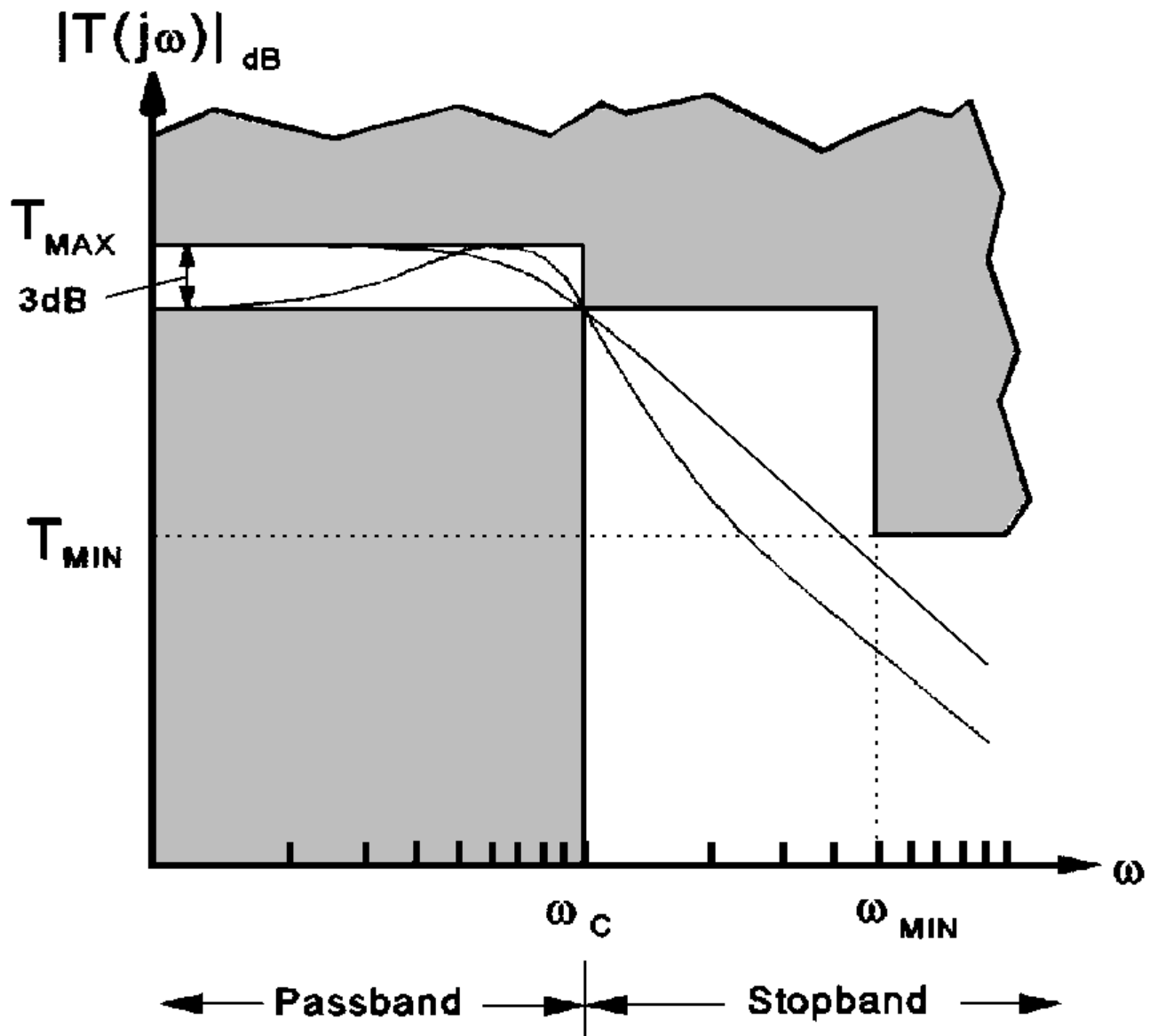


Figure 107.4 shows how low-pass filter requirements are often specified. To meet the specification, the gain response must lie within the unshaded region in the figure, as illustrated by the two responses shown in Fig. 107.4. The parameter  $T_{\max}$  is the *passband gain*. In the passband the gain must be within 3 dB of  $T_{\max}$  and must equal  $T_{\max}/\sqrt{2}$  at the cutoff frequency  $\omega_C$ . In the stopband the gain must decrease and remain below a gain of  $T_{\min}$  for all  $\omega \geq \omega_{\min}$ . A low-pass filter design requirement is usually defined by specifying values for these four parameters. The

parameters  $T_{\max}$  and  $\omega_C$  define the passband response, whereas  $T_{\min}$  and  $\omega_{\min}$  specify how rapidly the stopband response must decrease.

**Figure 107.4** Parameters for specifying low-pass filter requirements.



## 107.2 Applications

### Simple RL and RC Filters

A first-order LP filter has the following transfer function:

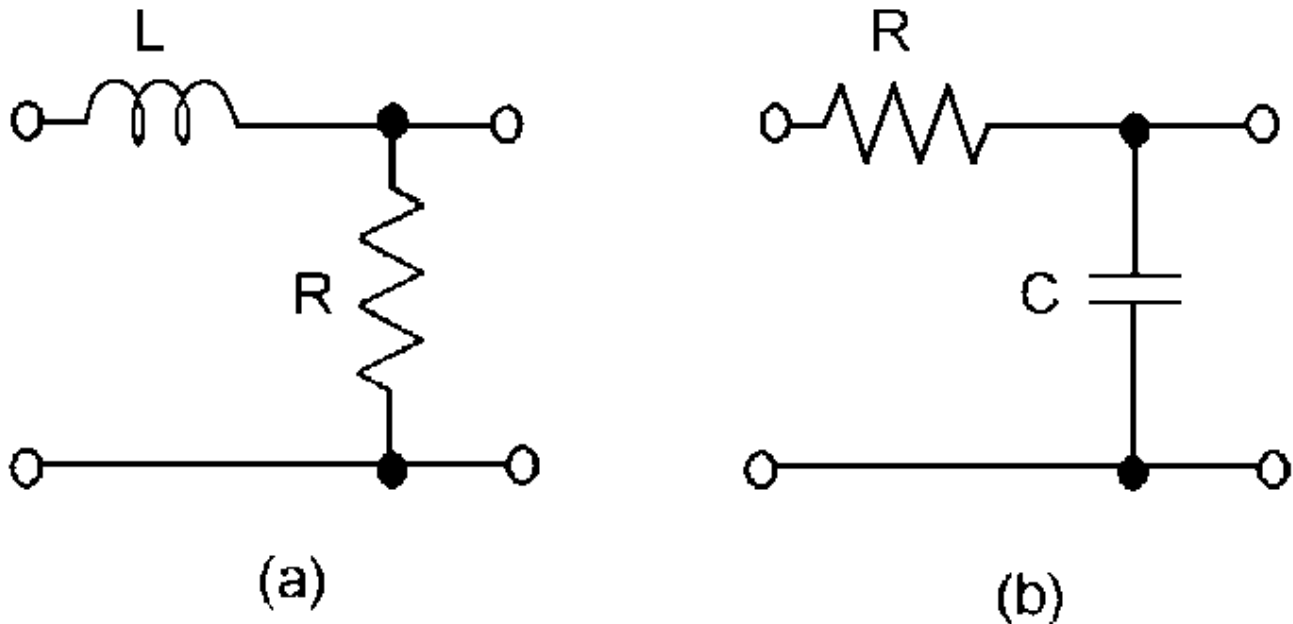
$$T(s) = \frac{K}{s + \alpha} \quad (107.3)$$

where for a passive filter  $K \leq \alpha$  and  $\alpha = \omega_C$ . This transfer function can be realized using either of the two circuits shown in Fig. 107.5. For sinusoidal response the respective transfer functions are

$$T(j\omega)_{RL} = \frac{R/L}{j\omega + (R/L)}; \quad T(j\omega)_{RC} = \frac{1/RC}{j\omega + (1/RC)} \quad (107.4)$$

For these filters the passband gain is one and the cutoff frequency is determined by  $R/L$  for the RL filter and  $1/RC$  for the RC filter. The gain  $|T(j\omega)|$  and phase  $\angle T(j\omega)$  plots of these circuits are shown in Fig. 107.1.

**Figure 107.5** Single-pole LP filter realizations: (a) RL, (b) RC.



A first-order HP filter is given by the following transfer function:

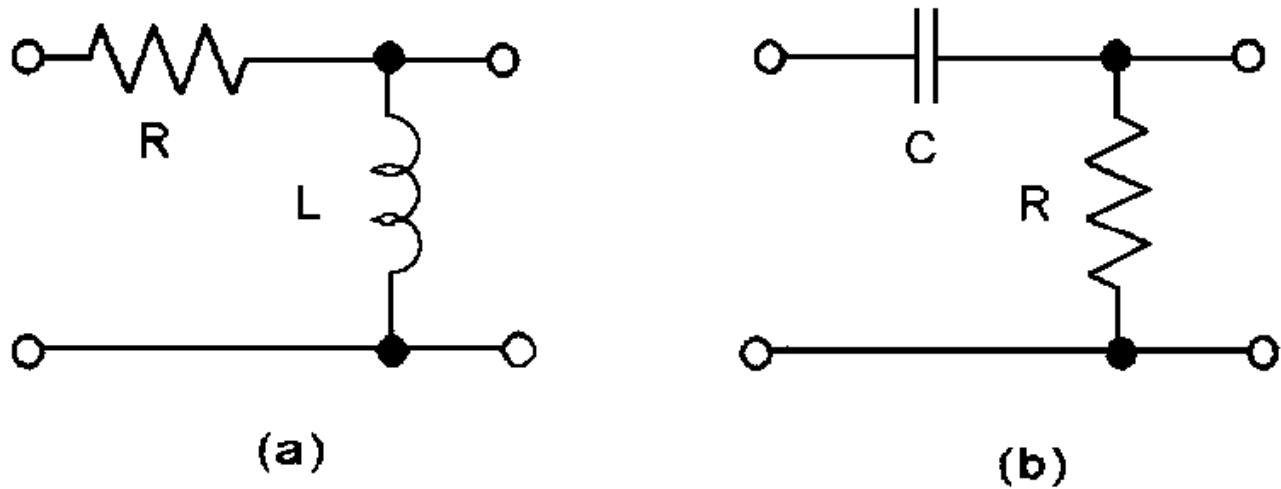
$$T(s) = \frac{Ks}{s + \alpha} \quad (107.5)$$

where, for a passive filter,  $K \leq 1$  and  $\alpha$  is the cutoff frequency. This transfer function can be realized using either of the two circuits shown in Fig. 107.6. For sinusoidal response the respective transfer functions are

$$T(j\omega)_{RL} = \frac{j\omega}{j\omega + (R/L)}; \quad T(j\omega)_{RC} = \frac{j\omega}{j\omega + (1/RC)} \quad (107.6)$$

For the LP filters the passband gain is one and the cutoff frequency is determined by  $R/L$  for the RL filter and  $1/RC$  for the RC filter. The gain  $|T(j\omega)|$  and phase  $\angle T(j\omega)$  plots of these circuits are shown in Fig. 107.7.

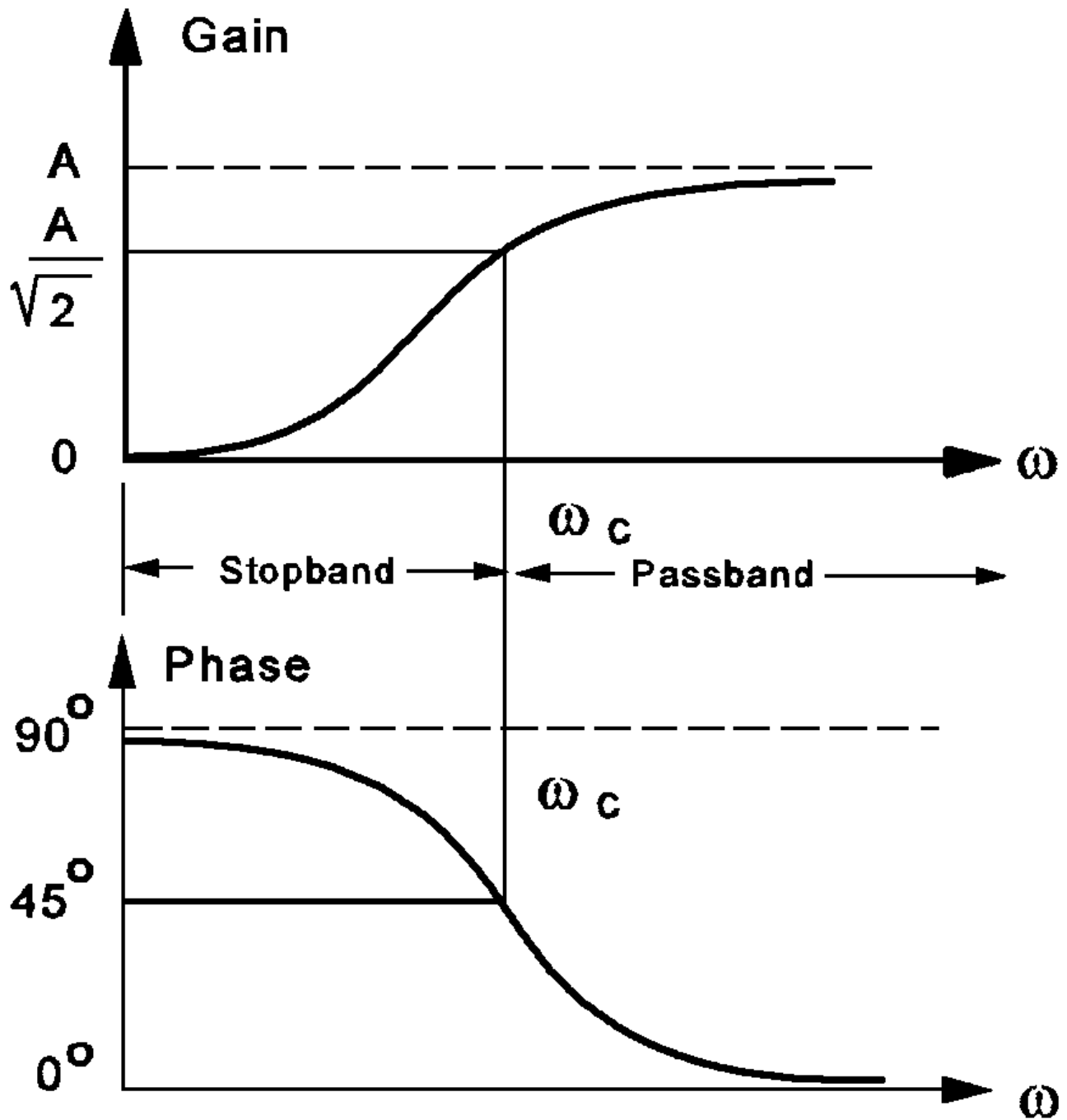
**Figure 107.6** Single-pole HP filter realizations: (a) RL, (b) RC.



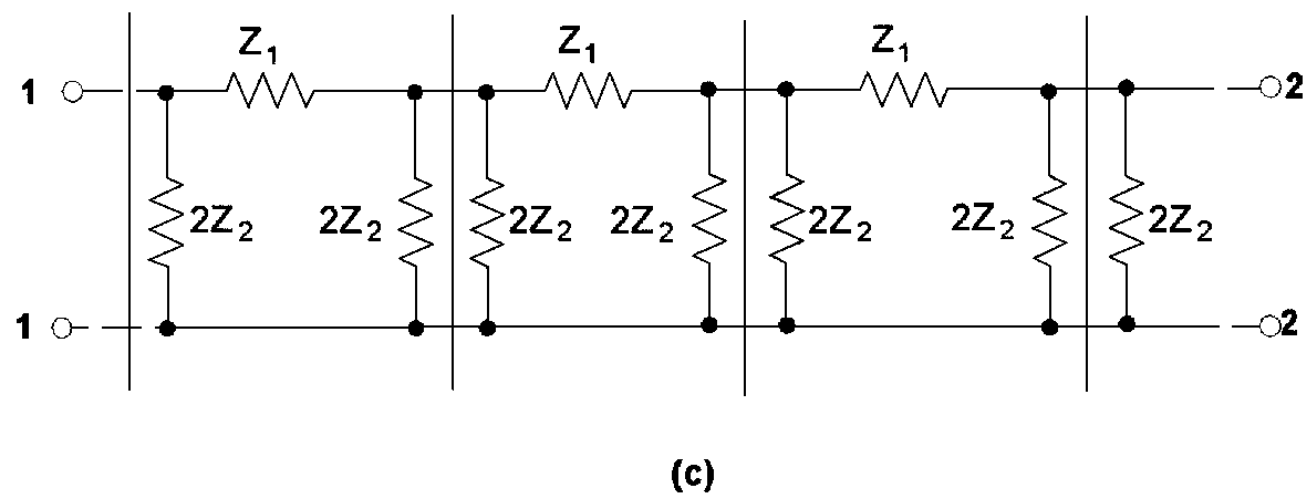
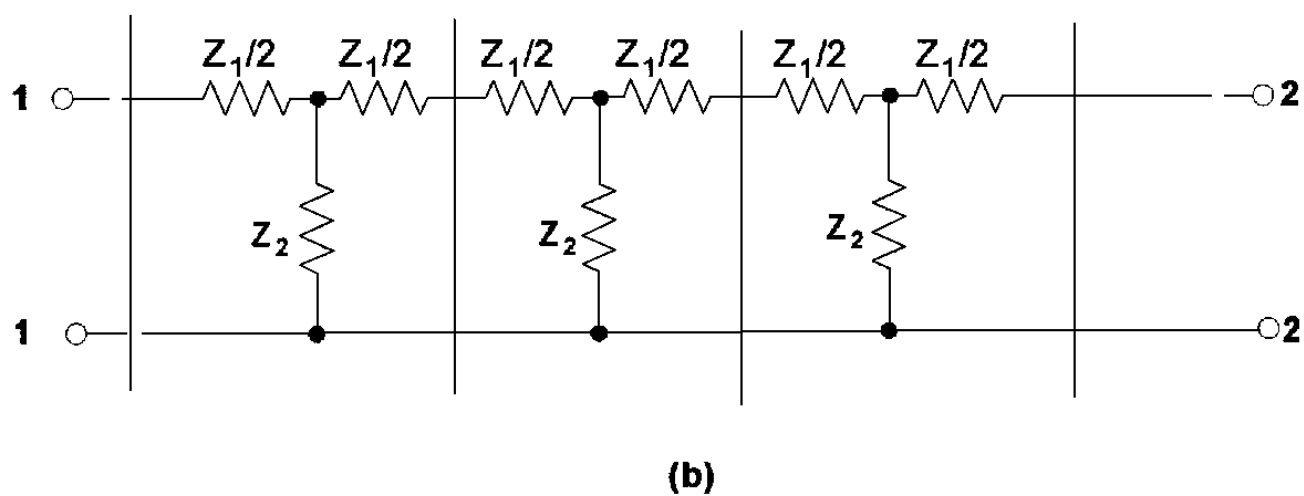
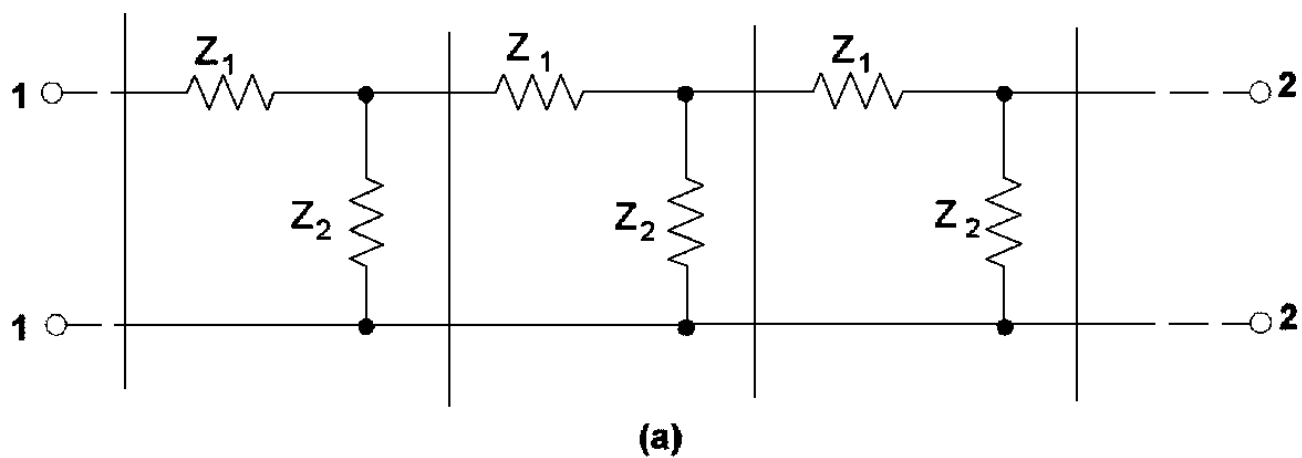
## Compound Filters

Compound filters are higher-order filters obtained by cascading lower-order designs. *Ladder circuits* are an important class of compound filters. Two of the more common passive ladder circuits are the *constant-k* and the *m-derived* filter (either of which can be configured using a *T-section*,  $\pi$ -*section*, *L-section*, or combinations thereof), the **bridge-T network** and **parallel-T network**, and the **Butterworth** and **Chebyshev** realizations. Only the first two will be discussed in this section. Figure 107.8(a) shows a standard ladder network consisting of two impedances,  $Z_1$  and  $Z_2$ , organized as an L-section filter. Figures 107.8(b) and (c) show how the circuit can be redrawn to represent a T-section or  $\Pi$ -section filter, respectively.

**Figure 107.7** High-pass filter characteristics showing passband, stopband, and the cutoff frequency,  $\omega_c$ .



**Figure 107.8** Ladder networks: (a) standard L-section, (b) T-section, (c)  $\Pi$ -section.





T- and  $\Pi$ -section filters are usually designed to be symmetrical so that either can have its input and output reversed without changing its behavior. The L-section is unsymmetrical and orientation is important. Since cascaded sections "load" each other, the choice of termination impedance is important. The *image impedance*,  $Z_i$ , of a symmetrical filter is the impedance with which the filter must be terminated in order to "see" the same impedance at its input terminals. The image impedance of a filter can be found from

$$Z_i = \sqrt{Z_{1O} Z_{1S}} \quad (107.7)$$

where  $Z_{1O}$  is the input impedance of the filter with the output terminals open circuited, and  $Z_{1S}$  is its input impedance with the output terminals short circuited. For symmetrical filters the output and input can be reversed without any change in its image impedance—that is,

$$\begin{aligned} Z_{1i} &= \sqrt{Z_{1O} Z_{1S}} \quad \text{and} \quad Z_{2i} = \sqrt{Z_{2O} Z_{2S}} \\ Z_{1i} &= Z_{2i} = Z_i \end{aligned} \quad (107.8)$$

The concept of matching filter sections and terminations to a common image impedance permits the development of symmetrical filter designs.

The image impedances of T- and  $\Pi$ -section filters are given as

$$\begin{aligned} Z_{iT} &= \sqrt{Z_{1O} Z_{1S}} = \sqrt{\frac{1}{4} Z_1^2 + Z_1 Z_2} \quad \text{and} \\ Z_{i\Pi} &= \sqrt{Z_{1O} Z_{1S}} = \frac{Z_1 Z_2}{\sqrt{(1/4) Z_1^2 + Z_1 Z_2}} \end{aligned} \quad (107.9)$$

The image impedance of an L-section filter, being unsymmetrical, depends on the terminal pair being calculated. For the L-section shown in [Fig. 107.8\(a\)](#) image impedances are

$$\begin{aligned} Z_{1iL} &= \sqrt{\frac{1}{4} Z_1^2 + Z_1 Z_2} = Z_{iT} \quad \text{and} \\ Z_{2iL} &= \frac{Z_1 Z_2}{\sqrt{(1/4) Z_1^2 + Z_1 Z_2}} = Z_{i\Pi} \end{aligned} \quad (107.10)$$

These equations show that the image impedance of an L-section at its input is equal to the image impedance of a T-section, whereas the image impedance of an L-section at its output is equal to the image impedance of a  $\Pi$ -section. This relationship is important in achieving an optimum termination when cascading L-sections with T- and/or  $\Pi$ -sections to form a composite filter.

Since  $Z_1$  and  $Z_2$  vary significantly with frequency, the image impedances of T- and  $\Pi$ -sections will also change. This condition does not present any particular problem in combining any number of equivalent filter sections together, since their impedances vary equally at all frequencies.

To develop the theory of constant- $k$  and  $m$ -derived filters, consider the circuit of [Fig. 107.9](#). The

current transfer function in the sinusoidal steady state is given by

$$T(j\omega) = |T(j\omega)| \angle T(j\omega) = I_2/I_1 \quad :$$

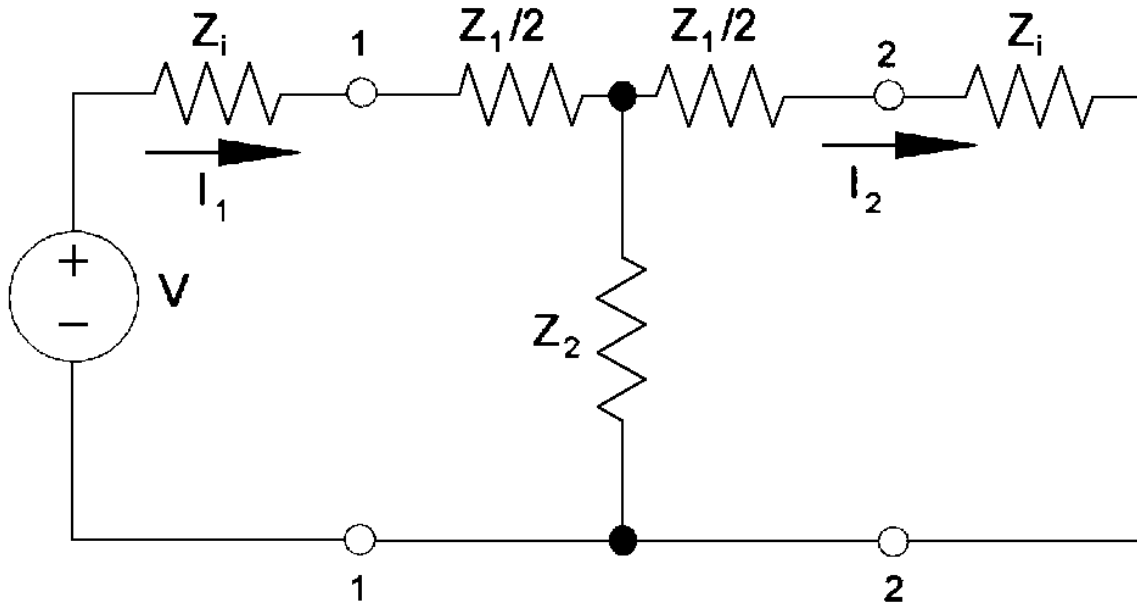
$$\begin{aligned} T(j\omega) &= |T(j\omega)| \angle T(j\omega) \\ &= \frac{I_2(j\omega)}{I_1(j\omega)} = e^{-\alpha} e^{-j\beta} = e^{-\gamma} \end{aligned} \quad (107.11)$$

where  $\alpha$  is the attenuation in dB,  $\beta$  is the phase shift in radians, and  $\gamma$  is the image transfer function. For the circuit shown in Fig. 107.9 the following relationship can be derived:

$$\tanh \gamma = \sqrt{Z_{1S}/Z_{1O}} \quad (107.12)$$

This relationship and those in Eq. (107.8) will be used to develop the constant- $k$  and  $m$ -derived filters.

**Figure 107.9** Circuit for determining the transfer function of a T-section filter.



## Constant- $k$ Filters

O. Zobel developed an important class of symmetrical filters called *constant- $k$  filters* with the conditions that  $Z_1$  and  $Z_2$  are purely reactive—that is,  $\pm X(j\omega)$  and

$$Z_1 Z_2 = k^2 = R^2 \quad (107.13)$$

Note that the units of  $k$  are ohms; in modern references the  $k$  is replaced by an  $R$ . The advantage of this type of filter is that the image impedance in the passband is a pure resistance, whereas in the stopband it is purely reactive. Hence if the termination is a pure resistance and equal to the image impedance, all the power will be transferred to the load since the filter itself is purely reactive. Unfortunately, the value of the image impedance varies significantly with frequency, and any termination used will result in a mismatch except at one frequency.

In LC constant- $k$  filters,  $Z_1$  and  $Z_2$  have opposite signs, so that  $Z_1 Z_2 = \pm jX_1 \mp jX_2 = +X_1 X_2 = R^2$ . The image impedances become

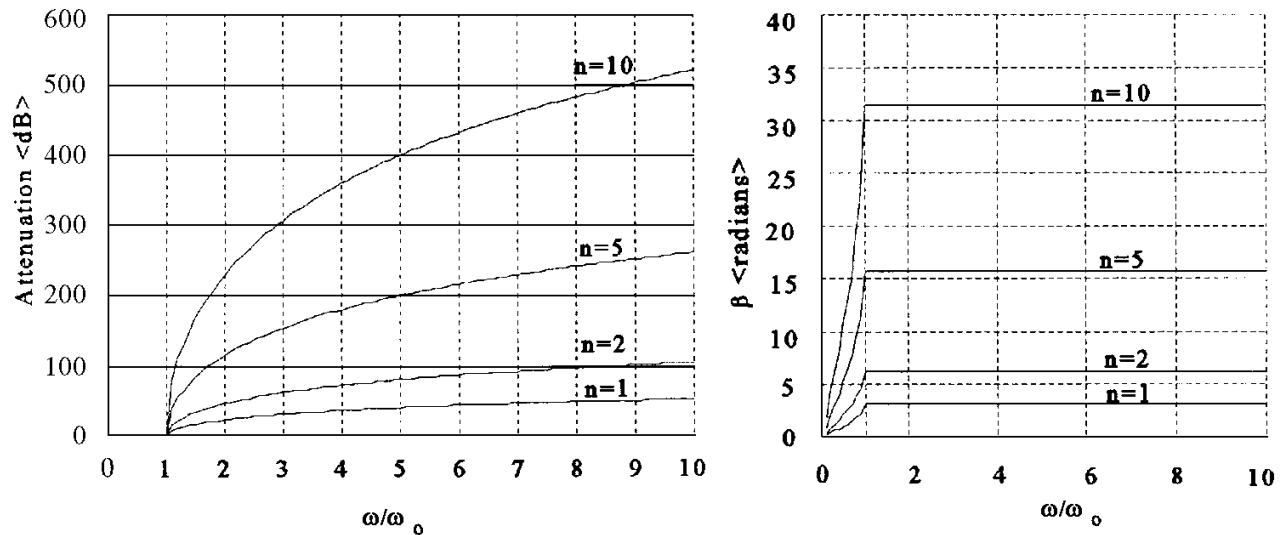
$$Z_{iT} = R\sqrt{1 - (-Z_1/4Z_2)} \quad \text{and} \quad Z_{i\Pi} = \frac{R}{\sqrt{1 - (-Z_1/4Z_2)}} \quad (107.14)$$

Therefore, in the stopband and passband, we have the following relations for standard T- or  $\Pi$ -sections, where  $n$  represents the number of identical sections:

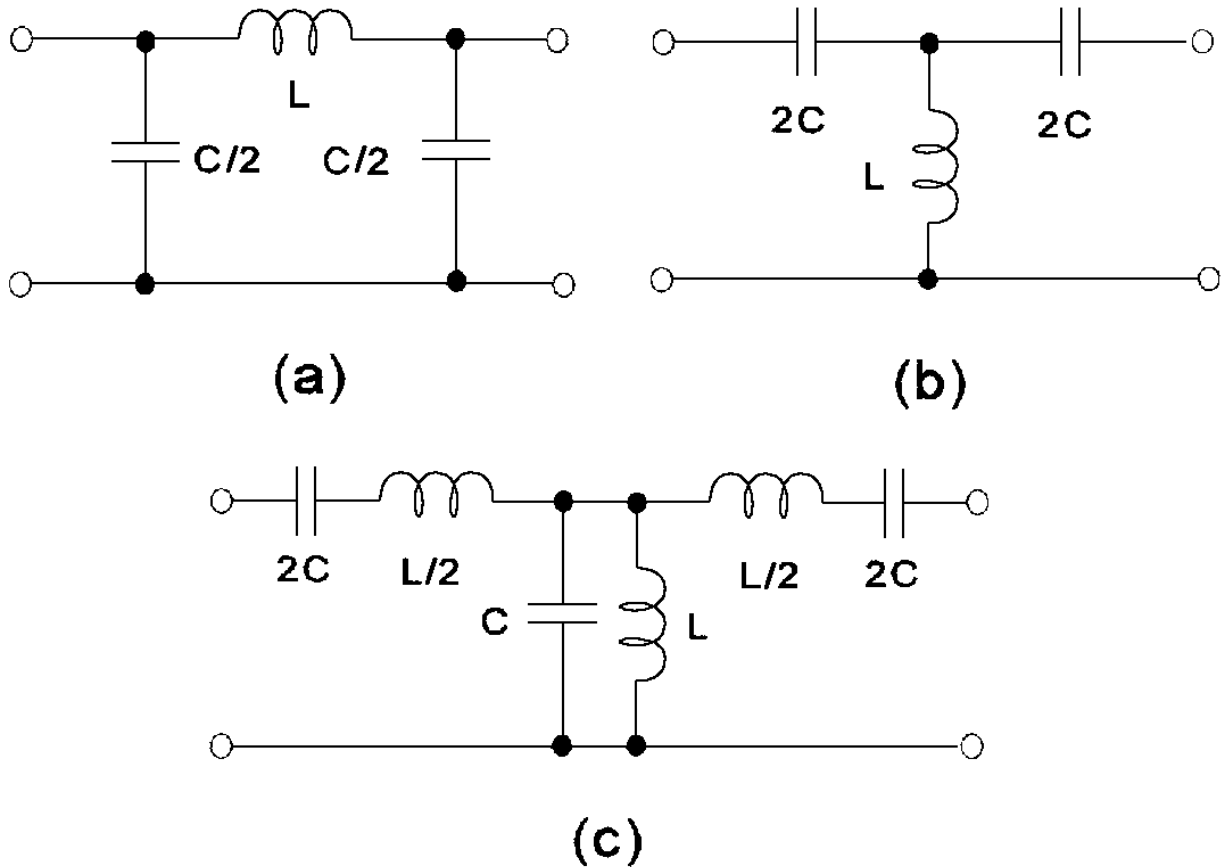
$$\begin{aligned} &\text{Stopband} \\ &\alpha = 2n \cosh^{-1} \sqrt{-Z_1/4Z_2} \\ &\beta = \pm n\pi, \pm 3n\pi, \dots \\ &\text{Passband} \\ &\alpha = 0 \\ &\beta = 2n \sin^{-1} \sqrt{-Z_1/4Z_2} \end{aligned} \quad (107.15)$$

Figure 107.10 shows normalized plots of  $\alpha$  and  $\beta$  versus  $\sqrt{-Z_1/4Z_2}$ . These curves are generalized and apply to low-pass, high-pass, band-pass, or band-reject filters. Figure 107.11 shows examples of a typical LP  $\Pi$ -section, an HP T-section, and a BP T-section.

**Figure 107.10** Normalized plots of attenuation and phase angle for various numbers of sections  $n$ .



**Figure 107.11** Typical sections: (a) LP  $\Pi$ -section, (b) HP T-section, (c) BP T-section.



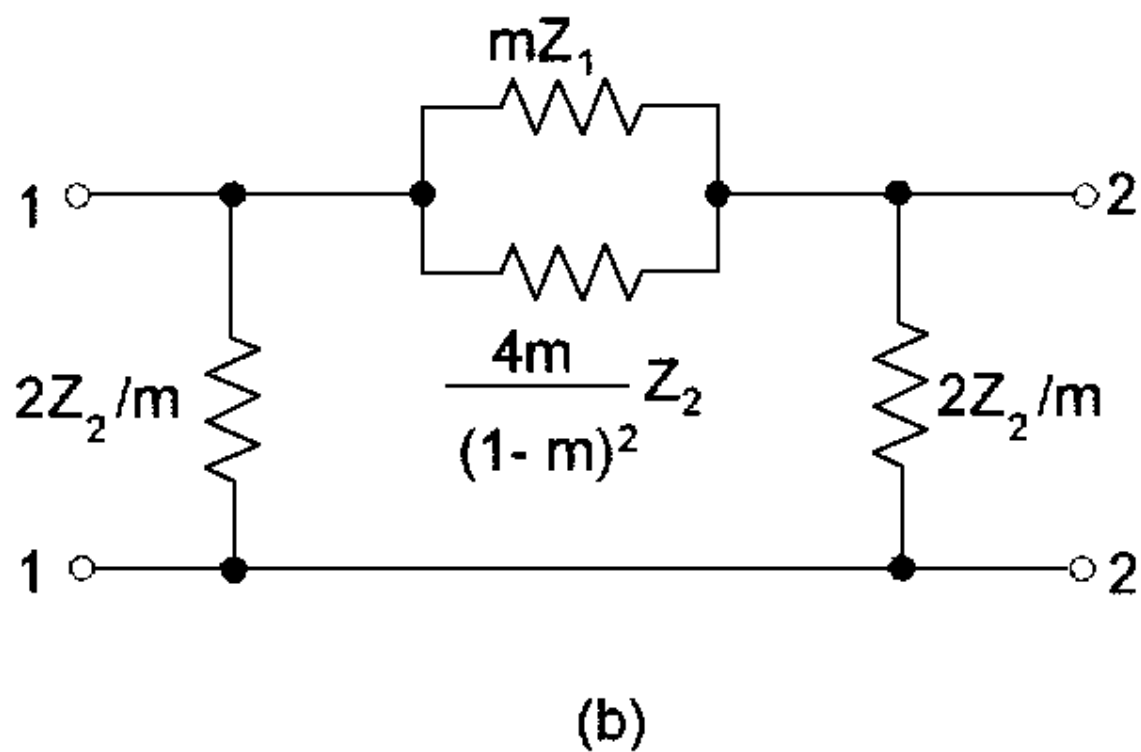
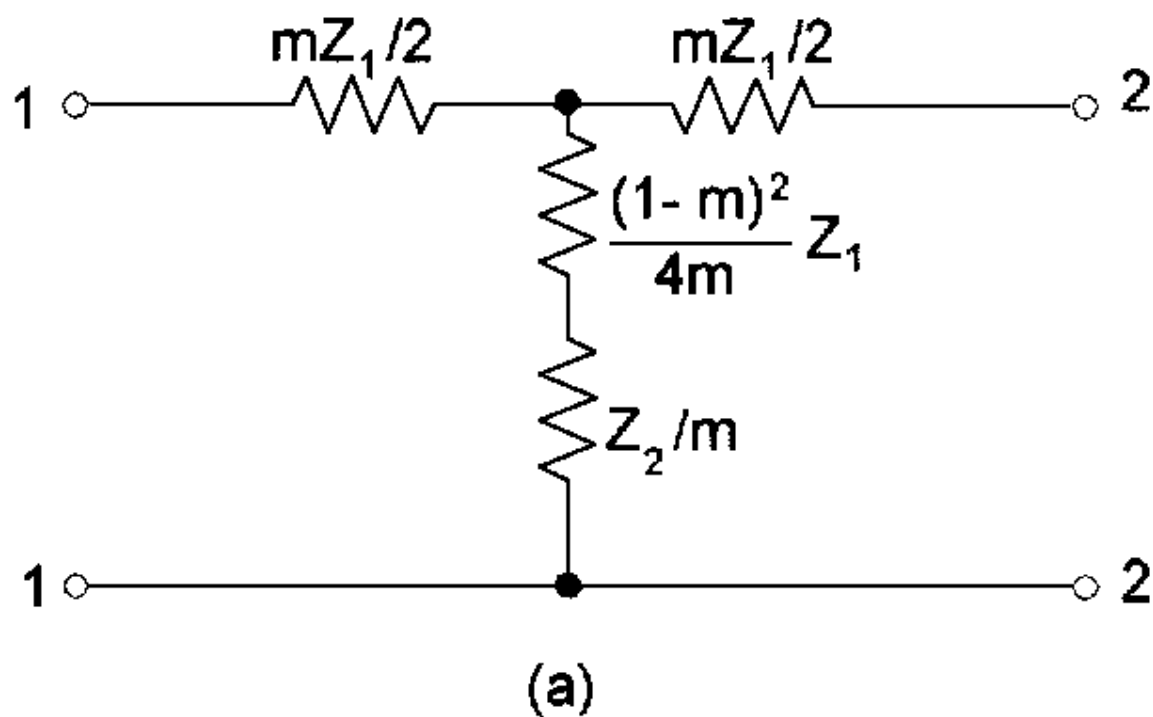
## ***m*-Derived Filters**

The need to develop a filter section that could provide high attenuation in the stopband near the cutoff frequency prompted the development of the *m*-derived filter. O. Zobel developed a class of filters that had the same image impedance as the constant-*k* but had a higher attenuation near the cutoff frequency. The impedances in the *m*-derived filter were related to those in the constant-*k* as

$$Z_{1m} = mZ_{1k} \quad \text{and} \quad Z_{2m} = \left( \frac{1 - m^2}{4m} \right) Z_{1k} + \frac{1}{m} Z_{2k} \quad (107.16)$$

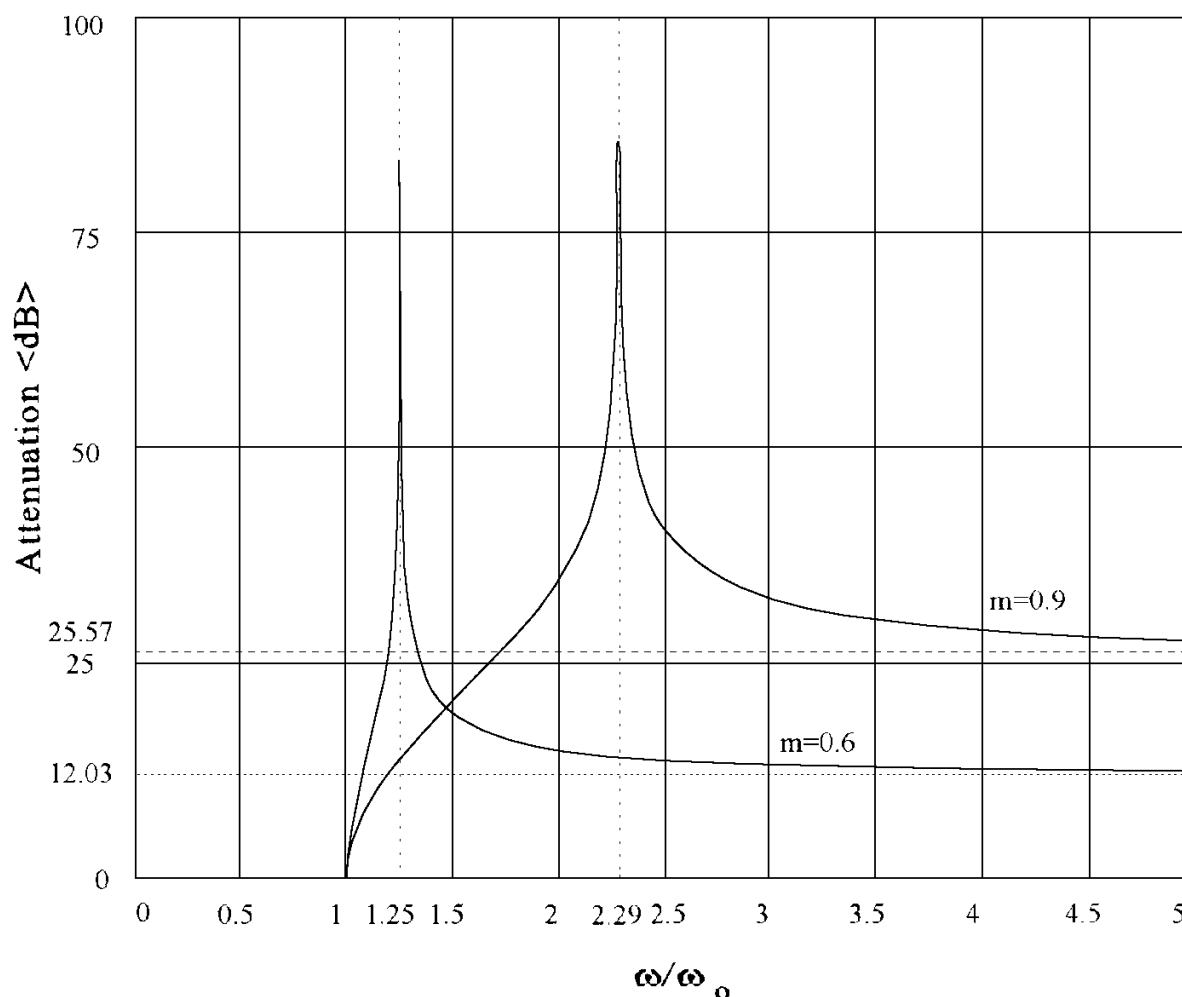
where *m* is a positive constant  $\leq 1$ . If *m* = 1 then the impedances reduce to those of the constant-*k*. [Figure 107.12](#) shows generalized *m*-derived T- and  $\Pi$ -sections.

**Figure 107.12**  $m$ -derived filters:(a) T-section, (b)  $\Pi$ -section.



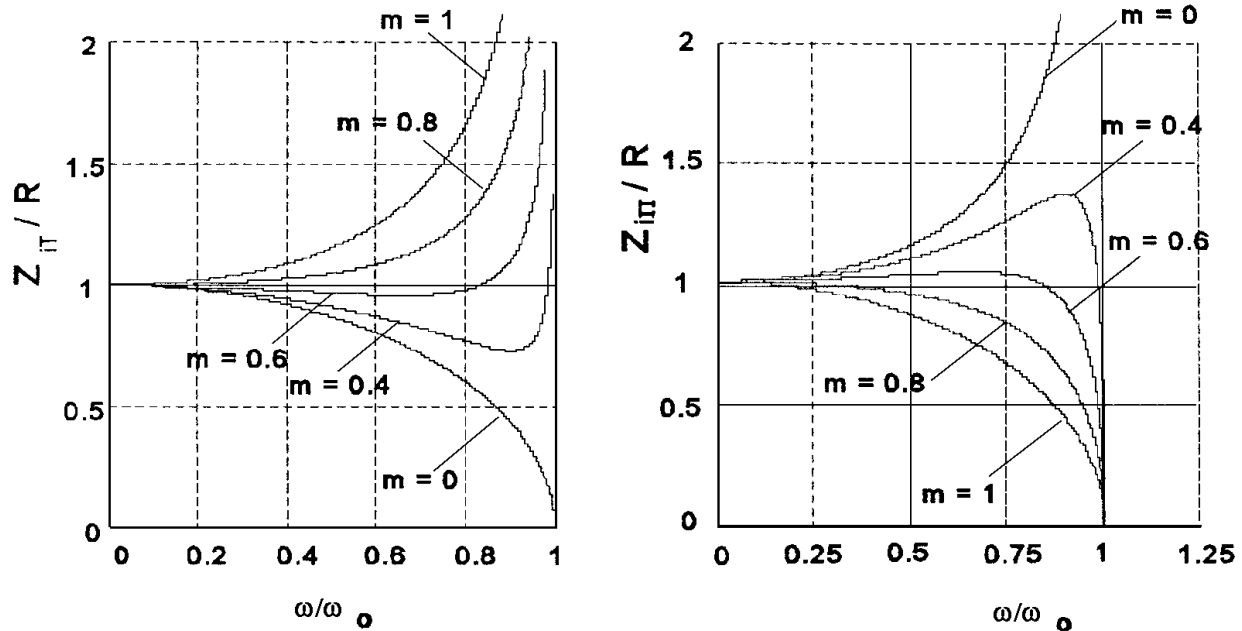
The advantage of the  $m$ -derived filter is that it gives rise to infinite attenuation at a selectable frequency just beyond cutoff. This singularity gives rise to a more rapid attenuation in the stopband than can be obtained using constant- $k$  filters. Figure 107.13 shows the attenuation curve for a single  $m$ -derived LP stage for two values of  $m$ . The smaller  $m$  becomes, the steeper the attenuation near the cutoff, but also the lesser the attenuation at higher frequencies.

**Figure 107.13** Attenuation curves for a single-stage filter with  $m = 0.6$  and  $m = 0.9$ .



Constant- $k$  filters have an image impedance in the passband that is always real but that varies with frequency, making the choice of an optimum termination difficult. The impedance of an  $m$ -derived filter also varies, but how it varies depends on  $m$ . Figure 107.14 shows how the impedance varies with frequency (both normalized) and  $m$ . In most applications,  $m$  is chosen to be 0.6, keeping the image impedance nearly constant over about 80% of the passband.

**Figure 107.14**  $Z_{iT}/R$  and  $Z_{i\Pi}/R$  versus normalized frequency for various values of  $m$ .



## Defining Terms

**Bridge-T network:** A two-port network that consists of a basic T-section and another element connected so as to "bridge across" the two arms. Such networks find applications as band rejection filters, calibration bridges, and feedback networks.

**Butterworth filters:** Ladder networks that enjoy a unique passband frequency response characteristic that remains very constant until near the cutoff, hence the designation "maximally flat." This filter has its critical frequency remain fixed regardless of the number of stages employed. It obtains this characteristic by realizing a transfer function built around a Butterworth polynomial.

**Chebyshev filters:** A variant of the Butterworth design that achieves a significantly steeper transition band for the same number of poles. Although the Chebyshev filter also maintains the integrity of its critical frequency regarding the number of poles, it trades the steeper roll-off for a fixed ripple—usually specified as 1 dB or 3 dB—in the passband. Chebyshev filters are also called *equal-ripple* or *stagger-tuned* filters. They are designed by realizing a transfer function using a Chebyshev polynomial.

**Parallel-T networks:** A two-port network that consists of two separate T-sections in parallel with only the ends of the arms and the stem connected. Parallel-T networks have applications similar to those of the bridge-T but can produce narrower attenuation bandwidths.

## References

- Herrero, J. L. and Willoner, G. 1966. *Synthesis of Filters*. Prentice Hall, Englewood Cliffs, NJ.
- Van Valkenburg, M. E. 1955. Two-terminal-pair reactive networks (filters). In *Network Analysis*. Prentice Hall, Englewood Cliffs, NJ.
- Weinberg, L. 1962. *Network Analysis and Synthesis*. W. L. Everitt (ed.) McGraw-Hill, New York.
- Zobel, O. J. 1923. Theory and Design of Uniform and Composite Electric Wave Filters. *Bell Telephone Syst. Tech. J.* 2:1.

## Further Information

- Huelsman, L. P. 1993. *Active and Passive Analog Filter Design¾An Introduction*. McGraw-Hill, New York. Good current introductory text covering all aspects of active and passive filter design.
- Sedra, A. S. and Brackett, P. O. 1978. *Filter Theory and Design: Active and Passive*. Matrix, Beaverton, OR. Modern approach to filter theory and design.



Broadwater, R., Sargent, A., Lee, R. E. "Power Distribution"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

108.1 Equipment

108.2 System Divisions and Types

108.3 Electrical Analysis, Planning, and Design

108.4 System Control

108.5 Operations

**Robert Broadwater**

*Virginia Polytechnic Institute & State University*

**Albert Sargent**

*Arkansas Power & Light*

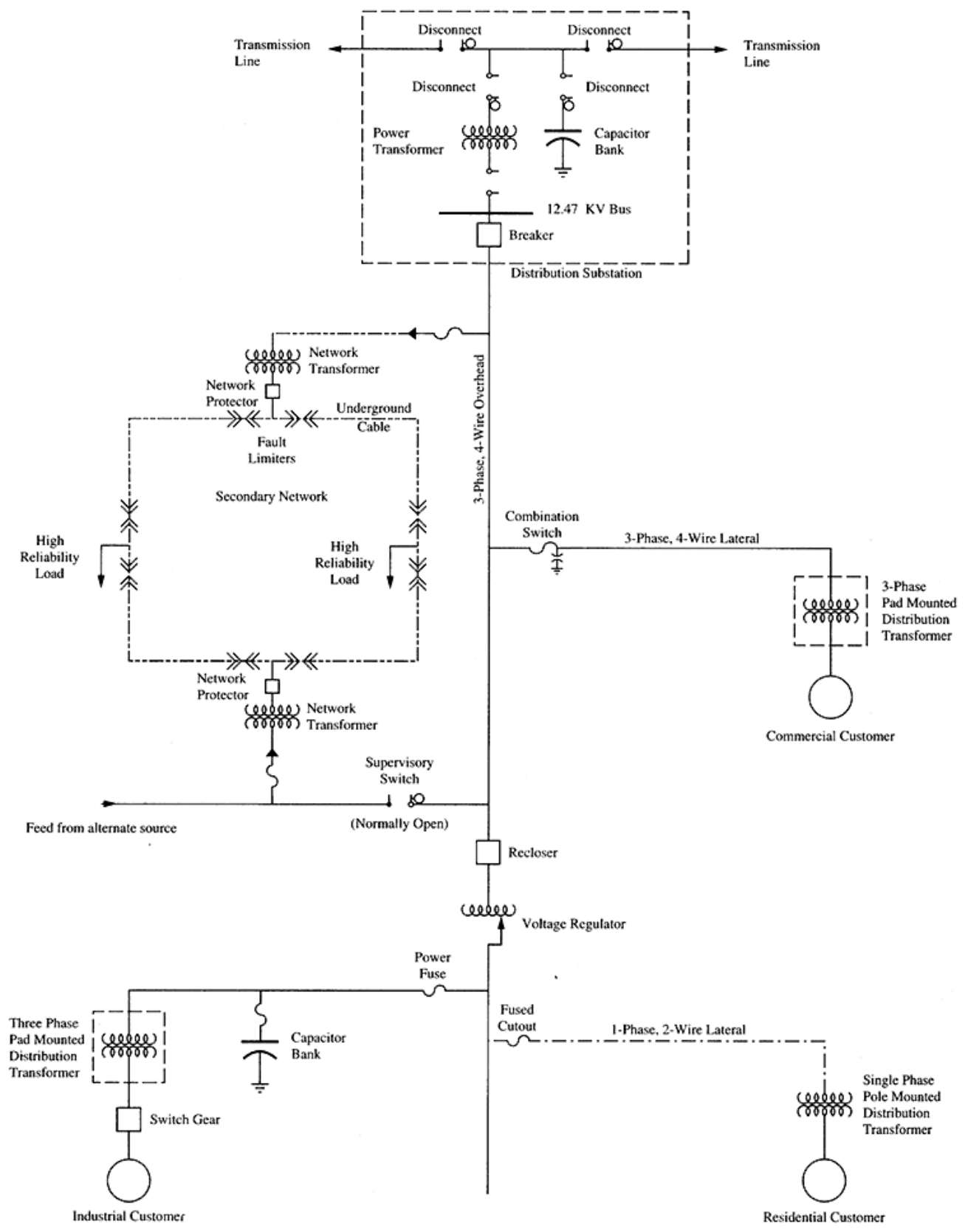
**Robert E. Lee**

*Pennsylvania Power & Light*

The function of power distribution is to deliver to consumers economic, reliable, and safe electrical energy in a manner that conforms to regulatory standards. Power distribution systems receive electric energy from high-voltage transmission systems and deliver energy to consumer service-entrance equipment. Systems typically supply alternating current at voltage levels ranging from 120 V to 46 kV.

Figure 108.1 illustrates aspects of a distribution system. Energy is delivered to the distribution substation (shown inside the dashed line) by three-phase transmission lines. A transformer in the substation steps the voltage down to the distribution primary system voltage—in this case, 12.47 kV. Primary distribution lines leave the substation carrying energy to the consumers. The substation contains a breaker that may be opened to disconnect the substation from the primary distribution lines. If the breaker is opened, outside the substation there is a normally open supervisory switch that may be closed in order to provide an alternate source of power for the customers normally served by the substation. The substation also contains a capacitor bank used for either voltage or power factor control.

**Figure 108.1** Distribution system schematic.



Four types of customers, along with representative distribution equipment, are shown in [Fig. 108.1](#). A set of loads requiring high reliability of service is shown being fed from an underground three-phase secondary network cable grid. A single fault does not result in an interruption to this set of loads. A residential customer is shown being supplied from a two-wire, one-phase overhead lateral. Commercial and industrial customers are shown being supplied from the three-phase, four-wire, overhead primary feeder. At the industrial site a capacitor bank is used to control power factor. Except for the industrial customer, all customers shown have 240/120 V service. The industrial customer has 480Y/277 V service.

For typical electric utilities in the U.S., investment in distribution ranges from 35 to 60% of total capital investment.

## 108.1 Equipment

---

[Figure 108.1](#) illustrates a typical arrangement of some of the most common equipment. Equipment may be placed into the general categories of transmission, protection, and control.

Arresters protect distribution system equipment from transient overvoltages due to lightning or switching operations. In overvoltage situations the arrester provides a low-resistance path to ground for currents to follow.

Capacitor banks are energy storage devices primarily used to control voltage and power factor. System losses are reduced by the application of capacitors.

Conductors are used to transmit energy and may be either bare or insulated. Bare conductors have better thermal properties and are generally used in overhead construction where contact is unlikely. Insulated cables are used in underground/conduit construction. Concentric neutral and tape-shielded cables provide both a phase conductor and a return path conductor in one unit.

Distribution lines are made up of conductors and are classified according to primary voltage, the number of **phases**, number of conductors, and return path. The three-phase, four-wire, multigrounded system is the most common primary system, where one conductor is installed for each of the three phases and the fourth conductor is a neutral that provides a **return current path**. *Multigrounded* means that the neutral is grounded at many points, so that the earth provides a parallel path to the neutral for return current. Three-phase, three-wire primary systems, or *delta-connected systems*, are rarely used because faults therein are more difficult to detect. A lateral is a branch of the system that is shorter in length, more lightly loaded, or has a smaller conductor size than the primary feeder.

Distribution transformers step the voltage down from the primary circuit value to the customer utilization level, thus controlling voltage magnitude. Sizes range from 10 to 2500 kVA. Distribution transformers are installed on poles, ground-level pads, or in underground vaults. A specification of 12470Y/7200 V for the high-voltage winding of a single-phase transformer means the transformer may be connected in a line-to-neutral "wye" connection for a system with a line-to-line voltage of 12470 V. A specification of 240/120 V for the low-voltage winding means the transformer may be used for a three-wire connection with 120 V midtap voltage and 240 V full-winding voltage. A specification of 480Y/277 V for the low voltage winding means the winding may be wye-connected for a three-phase, four-wire service to deliver 480 V line-to-line

and 277 V line-to-neutral.

Distribution substations consist of one or more step-down power transformers configured with switch gear, protective devices, and voltage regulation equipment for the purpose of supplying, controlling, switching, and protecting the primary feeder circuits. The voltage is stepped down for safety and flexibility of handling in congested consumer areas. Overcurrent protective devices open and interrupt current flow in order to protect people and equipment from **fault** current. Switches are used for control to interrupt or redirect power flow. Switches may be operated either manually or remotely with supervisory control. Switches are usually rated to interrupt load current and may be either pad or pole mounted.

Power transformers are used to control and change voltage level. Power transformers equipped with **tap-changing mechanisms** can control secondary voltage over a typical range of plus or minus 10%.

Voltage regulators are autotransformers with tap-changing mechanisms that may be used throughout the system for voltage control. If the voltage at a remote point is to be controlled, then the regulator can be equipped with a line drop compensator that may be set to regulate the voltage at the remote point based upon local voltage and current measurements.

## 108.2 System Divisions and Types

---

Distribution transformers separate the primary system from the secondary. Primary circuits transmit energy from the distribution substation to customer distribution transformers. Three-phase distribution lines that originate at the substation are referred to as *primary feeders* or *primary circuits*. Primary feeders are illustrated in [Fig. 108.1](#). Secondary circuits transmit energy from the distribution transformer to the customer's service entrance. Line-to-line voltages range from 208 to 600 V.

Radial distribution systems provide a single path of power flow from the substation to each individual customer. This is the least costly system to build and operate, and thus the most widely used.

Primary networks contain at least one loop that generally may receive power from two distinct sources. This design results in better continuity of service. A primary network is more expensive than the radial system design because more protective devices, switches, and conductors are required.

Secondary networks are normally underground cable grids providing multiple paths of power flow to each customer. A secondary network generally covers a number of blocks in a downtown area. Power is supplied to the network at a number of points via network units, consisting of a network transformer in series with a network protector. A network protector is a circuit breaker connected between the secondary winding of the network transformer and the secondary network itself. When the network is operating properly, energy flows into the network. The network protector opens when reverse energy flow is detected, such as may be caused by a fault in the primary system.

## 108.3 Electrical Analysis, Planning, and Design

---

The distribution system is planned, designed, constructed, and operated based on the results of electrical analysis. Generally, computer-aided analysis is used.

Line impedances are needed by most analysis applications. Distribution lines are electrically unbalanced due to loads, unequal distances between phases, dissimilar phase conductors, and single-phase or two-phase laterals. Currents flow in return paths due to the imbalance in the system. Three-phase, four-wire, multigrounded lines have two return paths—the neutral conductor and earth. Three-phase, multigrounded concentric neutral cable systems have four return paths. The most accurate modeling of distribution system impedance is based upon Carson's equations. With this approach a  $5 \times 5$  impedance matrix is derived for a system with two return paths, and a  $7 \times 7$  impedance matrix is derived for a system with four return paths. For analysis, these matrices are reduced to  $3 \times 3$  matrices that relate phase voltage drops (i.e.,  $\Delta V_A, \Delta V_B, \Delta V_C$ ) to phase currents (i.e.,  $I_A, I_B, I_C$ ), as indicated by

$$\begin{bmatrix} \Delta V_A \\ \Delta V_B \\ \Delta V_C \end{bmatrix} = \begin{bmatrix} Z_{AA} & Z_{AB} & Z_{AC} \\ Z_{BA} & Z_{BB} & Z_{BC} \\ Z_{CA} & Z_{CB} & Z_{CC} \end{bmatrix} \begin{bmatrix} I_A \\ I_B \\ I_C \end{bmatrix}$$

Load analysis forms the foundation of system analysis. Load characteristics are time varying and depend on many parameters, including connected consumer types and weather conditions. The load demand for a given customer or group of customers is the load averaged over an interval of time, say 15 minutes. The peak demand is the largest of all demands. The peak demand is of particular interest since it represents the load that the system must be designed to serve. Diversity relates to multiple loads having different time patterns of energy use. Due to diversity, the peak demand of a group of loads is less than the sum of the peak demands of the individual loads. For a group of loads,

$$\text{Diversity factor} = \frac{\text{Sum of individual load peaks}}{\text{Group peak}}$$

Loads may be modeled as either lumped parameter or distributed. Lumped parameter load models include constant power, constant impedance, constant current, voltage-dependent, and combinations thereof. Generally, equivalent lumped parameter load models are used to model distributed loads. Consider the line section of length  $L$  shown in [Fig. 108.2\(a\)](#), with a uniformly distributed load current that varies along the length of the line as given by

$$i(x) = \frac{I_2 - I_1}{L}x + I_1$$

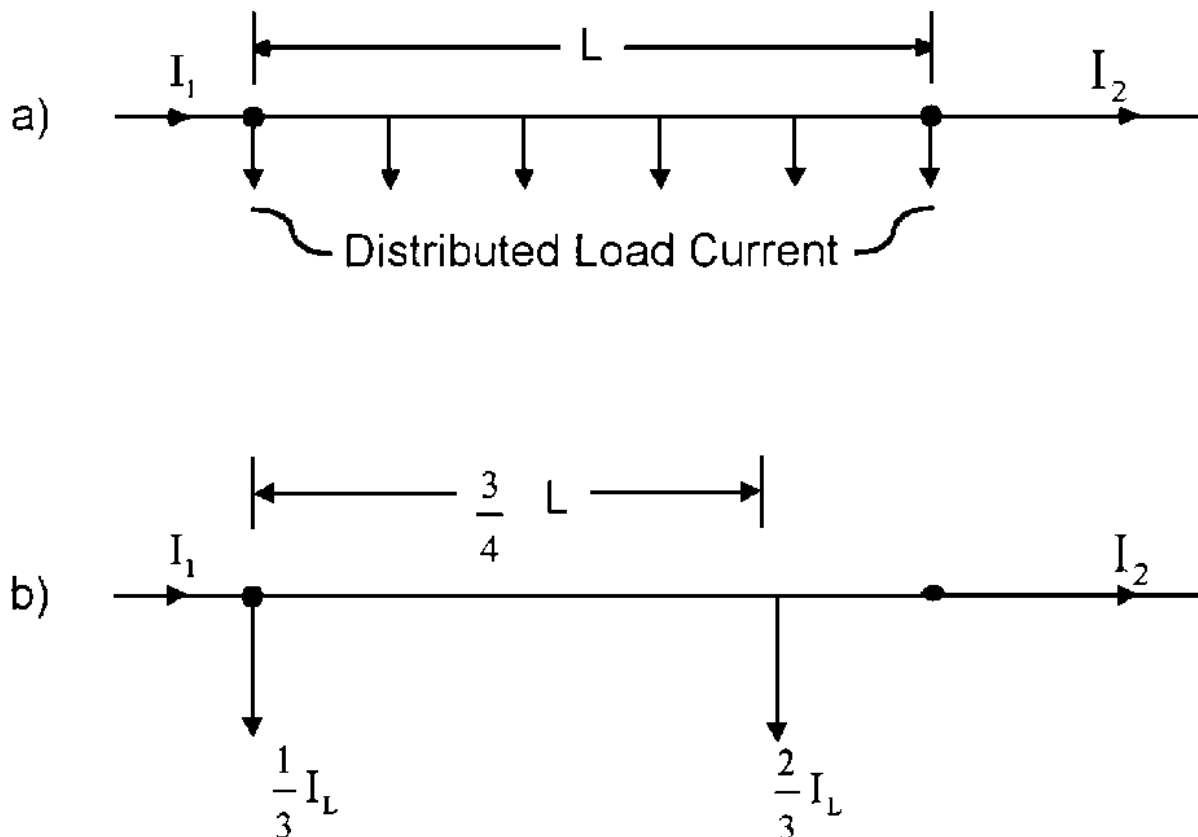
The total load current drawn by the line section is thus

$$I_L = I_2 - I_1$$

An equivalent lumped parameter model for the uniformly distributed current load is shown in [Fig.](#)

108.2(b).

**Figure 108.2** (a) Line section model with distributed load current; (b) lumped parameter equivalent model.



Load forecasting is concerned with determining load magnitudes during future years from customer growth projections. Short-range forecasts generally have time horizons of approximately five years, whereas long-range forecasts project to around twenty years.

Power flow analysis determines system voltages, currents, and power flows. Power flow results are checked to ensure that voltages fall within allowable limits, that equipment overloads do not exist, and that phase imbalances are within acceptable limits. For primary and secondary networks, power flow methods used in transmission system analysis are applied. For radially operated systems, the ladder method is used. The actual implementation of the ladder method may vary with the type of load models used. All ladder load flow methods assume the substation bus voltage is known. An algorithm for the ladder method is as follows:

*Step 1.* Assume a value for all node voltages throughout the circuit. Generally, assumed voltages are set equal to the substation voltage.

*Step 2.* At each load in the circuit, calculate the current from the known load value and assumed voltage.

*Step 3.* Starting at the ending nodes, sum load currents to obtain line section current estimates, performing summation until the substation is reached.

*Step 4.* Having estimates of all line section currents, start at the substation and calculate line section voltage drops and new estimates of node voltages.

*Step 5.* Compare new node voltages with estimates of previous iteration values. The algorithm has converged if the change in voltage is sufficiently small. If the algorithm has not converged, return to step 2.

Dynamic load analysis includes such studies as motor-starting studies. Rapid changes in large loads can result in large currents, with a resultant drop in system voltage. If the dip in voltage is too large or too frequent, then other loads are adversely affected, such as in an annoying flicker of lights. This study generally employs a power flow calculation that is run at a number of points along the dynamic characteristic of the load.

Fault analysis provides the basis for protection system design. Generally, superposition is used to add load currents obtained from power flow analysis to fault currents. Thus, in the model used to calculate fault currents, load currents are neglected. Sources of fault current are the substation bus, cogenerators, and large synchronous motors on the feeder or neighboring feeders. A variety of fault conditions are considered at each line section, including three-phase-to-ground, single-phase-to-ground, and separate phases contacting one another. In performing the calculations, both bolted (i.e., zero-impedance) faults and faults with an impedance in the fault path are considered. Of interest are the maximum and minimum phase and return path fault currents, as well as the fault types that result in these currents.

Reliability analysis involves determining indices that relate to continuity of service to the customer. Reliability is a function of tree conditions, lightning incidence, equipment failure rates, equipment repair times, and circuit design. The reliability of a circuit generally varies from point to point due to protection system design, placement of switches, and availability of alternative feeds. There are many indices used in evaluating system reliability. Common ones include system average interruption frequency index (SAIFI), system average interruption duration index (SAIDI), and customer average interruption frequency index (CAIFI), as defined by

$$\text{SAIFI} = \frac{\text{Total number of customer interruptions}}{\text{Total number of customers served}}$$

$$\text{SAIDI} = \frac{\text{Sum of customer interruption durations}}{\text{Total number of customers}}$$

$$\text{CAIFI} = \frac{\text{Total number of customer interruptions}}{\text{Total number of customers affected}}$$

Phase balancing is used to balance the current or power flows on the different phases of a line section. This results in improved efficiency and primary voltage level balance. The average current in the three phases is defined as



$$I_{\text{avg}} = \frac{I_A + I_B + I_C}{3}$$

The maximum deviation from  $I_{\text{avg}}$  is given by

$$\Delta I_{\text{dev}} = \text{maximum of } \{|I_{\text{avg}} - I_A|, |I_{\text{avg}} - I_B|, |I_{\text{avg}} - I_C|\}$$

Phase imbalance is defined as

$$\text{Phase imbalance} = \frac{\Delta I_{\text{dev}}}{I_{\text{avg}}}$$

Planning involves using load forecasting and other analysis calculations to evaluate voltage level, substation locations, feeder routes, transformer/conductor sizes, voltage/power factor control, and restoration operations. Decisions are based upon considerations of efficiency, reliability, peak demand, and life cycle cost.

Overcurrent protection is the most common protection applied to the distribution system. With overcurrent protection, the protective device trips when a large current is detected. The time to trip is a function of the magnitude of the fault current. The larger the fault current is, the quicker the operation. Various types of equipment are used. A circuit breaker is a switch designed to interrupt fault current, the operation of which is controlled by relays. An overcurrent relay, upon detecting fault current, sends a signal to the breaker to open. A recloser is a switch that opens and then recloses a number of times before finally locking open. A fuse is a device with a fusible member, referred to as a *fuse link*, which in the presence of an overcurrent melts, thus opening up the circuit.

Breakers may be connected to reclosing relays, which may be programmed for a number of opening and reclosing cycles. With a recloser or a reclosing breaker, if the fault is momentary, then the power interruption is also momentary. If the fault is permanent, then after a specified number of attempts at reclosing the device locks open. Breakers are generally more expensive than comparable reclosers. Breakers are used to provide more sophisticated protection, which is available via choice of relays. Fuses are generally used in the protection of laterals.

Protective equipment sizing and other characteristics are determined from the results of fault analysis. Moving away from the substation in a radial circuit, both load current and available fault current decrease. Protective devices are selected based on this current grading. Protective devices are also selected to have different trip delay times for the same fault current. With this time grading, protective devices are coordinated to work together such that the device closest to a permanent fault clears the fault. Thus reclosers can be coordinated to protect load-side fuses from damage due to momentary faults.

## 108.4 System Control

---

Voltage control is required for proper operation of customer equipment. For instance, in the U.S., "voltage range A" for single-phase residential users specifies that the voltage may vary at the service entrance from 114/228 V to 126/252 V. Regulators, tap-changing under load transformers,

and switched capacitor banks are used in voltage control.

Power factor control is used to improve system efficiency. Due to the typical load being inductive, power factor control is generally achieved with fixed and/or switched capacitor banks.

Power flow control is achieved with switching operations. Such switching operations are referred to as *system reconfiguration*. Reconfiguration may be used to balance a load among interconnected distribution substations. Such switching operations reduce losses while maintaining proper system voltage.

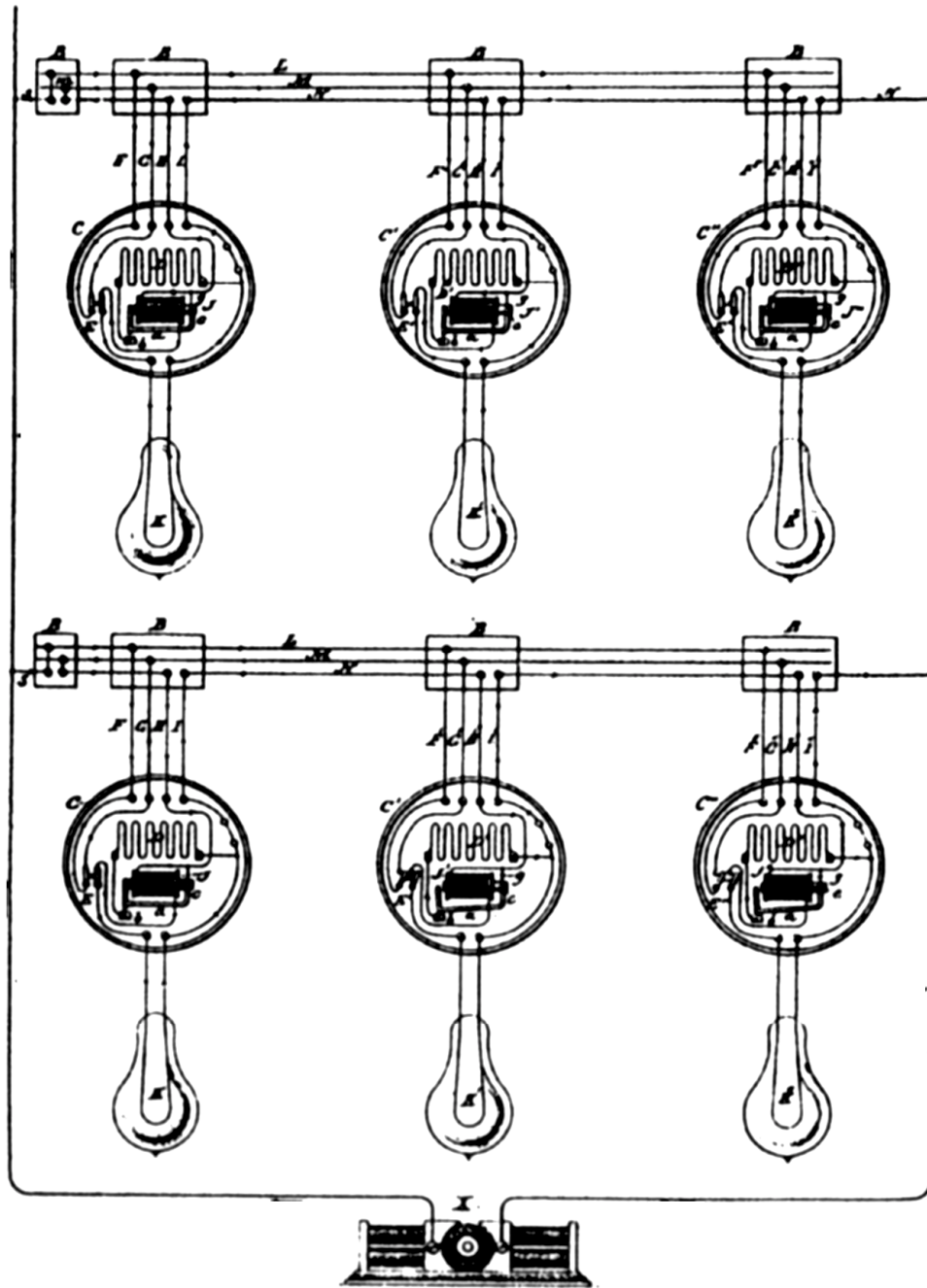
Load control may be achieved with voltage control and also by remotely operated switches that disconnect load from the system. Generally, load characteristics are such that if the voltage magnitude is reduced, then the power drawn by the load will decrease for some period of time. Load control with remotely operated switches is also referred to as *load management*.

## 108.5 Operations

---

The operations function includes system maintenance, construction, and service restoration. Maintenance, such as trimming trees to prevent contact with overhead lines, is important to ensure a safe and reliable system. Interruptions may be classified as *momentary* or *permanent*. A momentary interruption is one that disappears very quickly—for instance, a recloser operation due to a fault from a tree limb briefly touching an overhead conductor. Power restoration operations are required to repair damage caused by permanent interruptions.

While damaged equipment is being repaired, power restoration operations often involve reconfiguration in order to restore power to interrupted areas. With reconfiguration, power flow calculations may be required to ensure that equipment overloads are not created from the switching operations.



# SYSTEM OF ELECTRIC LIGHTING

*William Stanley, Jr.*

*Patented January 5, 1886*

*#333,564*

This invention described a system for wiring lamps in series/parallel combinations to maintain a constant load in the cross circuits even when individual lights were switched in and out.

An excerpt:

My invention consists in arranging and connecting the wires constituting each cross-circuit in the manner hereinafter shown, and connecting with the wires and lamps switches and resistances in such a way that the resistance in each cross-circuit shall remain constant so long as any lamp therein is lighted, and so that each cross-circuit shall be interrupted and no current pass through it when no lamp in it is lighted.

Stanley obtained 130 patents in his lifetime and was instrumental in the development of AC electric power distribution for Westinghouse and others. He developed practical transformers, a self-starting motor, a two-phase AC generator and an electric meter with magnetic suspension instead of jeweled bearings. (©1994, DewRay Products, Inc. Used with permission.)

## Defining Terms

**Current return path:** The path that current follows from the load back to the distribution substation. This path may consist of either a conductor (referred to as the *neutral*) or earth, or the parallel combination of a neutral conductor and the earth.

**Fault:** A conductor or equipment failure or unintended contact between conductors or between conductors and grounded objects. If not interrupted quickly, fault current can severely damage conductors and equipment.

**Phase:** Relates to the relative angular displacement of the three sinusoidally varying voltages produced by the three windings of a generator. For instance, if phase A voltage is  $120\angle 0^\circ$  V, phase B voltage  $120\angle -120^\circ$  V, and phase C voltage  $120\angle 120^\circ$  V, the phase rotation is referred to as ABC. Sections of the system corresponding to the phase rotation of the voltage carried are commonly referred to as phase A, B, or C.

**Tap-changing mechanism:** A control device that varies the voltage transformation ratio between the primary and secondary sides of a transformer. The taps may only be changed by discrete amounts, say 0.625%.

## References

- Broadwater, R. P., Shaalan, H. E., Oka, A., and Lee, R. E. 1993. Distribution system reliability and restoration analysis. *Electric Power Sys. Res. J.* 29(2):203–211.
- Carson, J. R. 1926. Wave propagation in overhead wires with ground return. *Bell System Tech. J.* 5: 40–47.
- Engel, M. V., Greene, E. R., and Willis, H. L. (Eds.) *IEEE Tutorial Course: Power Distribution Planning*. 1992. Course Text 92 EHO 361-6-PWR IEEE Service Center, Piscataway, NJ.
- Kersting, W. H. and Mendive, D. L. 1976. *An Application of Ladder Network Theory to the Solution of Three-Phase Radial Load Flow Problems*. IEEE Winter Meeting, New York.

## Further Information

- Burke, J. J. 1994. *Power Distribution Engineering*. Marcel Dekker, New York.
- Redmon, J. R. 1988. IEEE Tutorial Course on Distribution Automation. Course Text 88 EH0 280-8-PWR IEEE Service Center, Piscataway, NJ.
- Electric Utility Engineers, Westinghouse Electric Corporation. 1950. *Electrical Transmission and Distribution Reference Book*. Westinghouse Electric Corporation, Pittsburgh, PA.
- Gönen, T. 1986. *Electric Power Distribution System Engineering*. John Wiley & Sons, New York.
- Lakervi, E. and Holmes, E. J. 1989. *Electricity Distribution Network Design*. Peter Peregrinus, London.
- Pansini, A. J. 1992. *Electrical Distribution Engineering*. Fairmont Press, Liburn, GA.

Duff, W. G. "Grounding and Shielding"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Grounding and Shielding

---

## 109.1 Grounding

Characteristics of Ground Conductors • Ground-Related EMI Coupling • Grounding Configurations • Summary of Grounding Considerations

## 109.2 Shielding

Shielding Theory • Reflection Loss • Absorption Loss • Total Shielding Effectiveness • Shielding Materials • Conductive Coatings • Aperture Leakages • Summary of Shielding Considerations

### William G. Duff

*Computer Sciences Corporation*

Grounding and shielding are two very important factors that must be considered during the design of electronic circuits. Current trends in the electronics industry (such as increases in the number of electronic equipments, reliance on electronic devices in critical applications, higher clock frequencies of computing devices, higher power levels, lower sensitivities, increased packaging densities, use of plastics, etc.) will tend to create more electromagnetic interference (EMI) problems. To avoid problems, EMI control measures must be incorporated into circuit design.

## 109.1 Grounding

---

The material on grounding was adapted from Duff [1989] courtesy of Interference Control Technologies.

Grounding is one of the least understood and most significant factors in many EMI problems. The primary purposes for grounding circuits, cables, equipments, and systems are to prevent a shock hazard; to protect circuits and equipments; and to reduce EMI due to electromagnetic field, common ground impedance, or other forms of interference coupling. The EMI part of the problem is emphasized in this section.

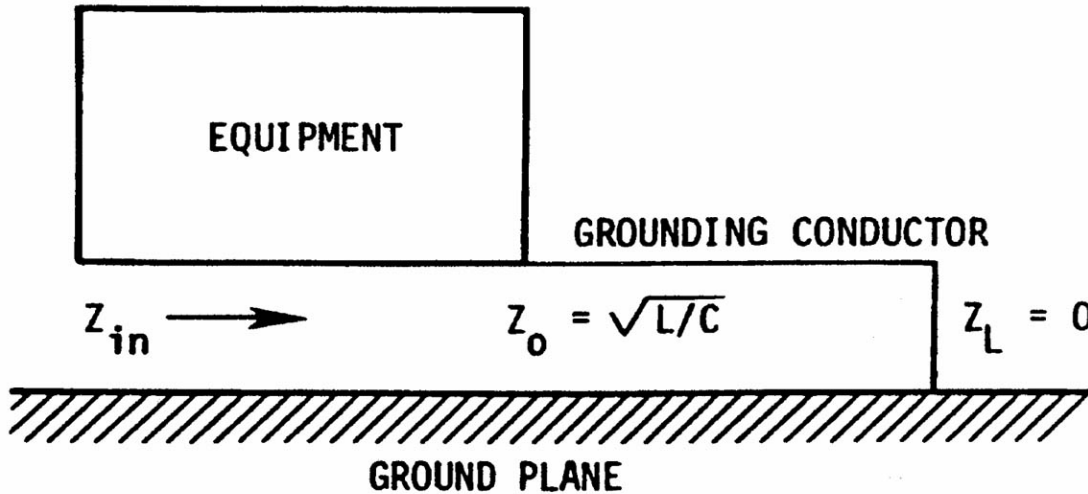
### Characteristics of Ground Conductors

Ideally, a ground conductor should provide a zero-impedance path to all signals for which it serves as a reference. If this were the situation, signal currents from different circuits would return to their respective sources without creating unwanted coupling between circuits. Many interference problems occur because designers treat the ground as ideal and fail to give proper attention to the actual characteristics of the ground conductor.

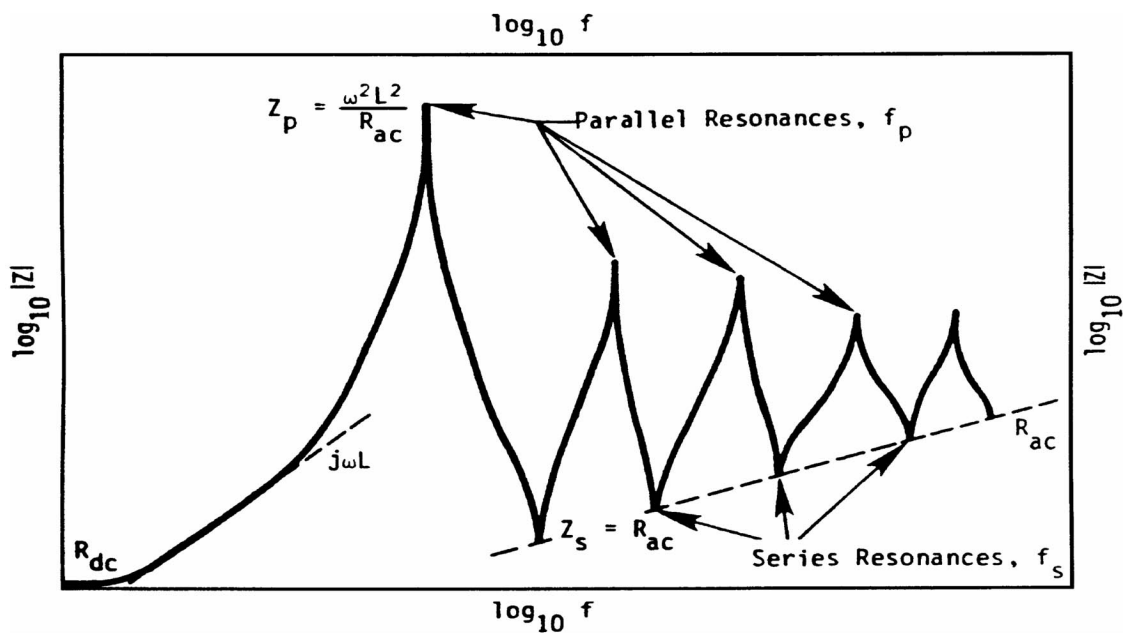
A commonly encountered situation is that of a ground conductor running along in the proximity of a ground plane as illustrated in Fig. 109.1. The ground conductor and ground plane may be

represented as a short-circuited transmission line. At low frequencies the resistance of the ground conductor will predominate. At higher frequencies the series inductance and the shunt capacitance to ground will become significant and the ground conductor will exhibit alternating parallel and series resonances as illustrated in Fig. 109.2. To provide a low impedance to ground, it is necessary to keep the length of the grounding conductor short relative to wavelength (i.e., less than 1/20 of the wavelength).

**Figure 109.1** Idealized equipment grounding.



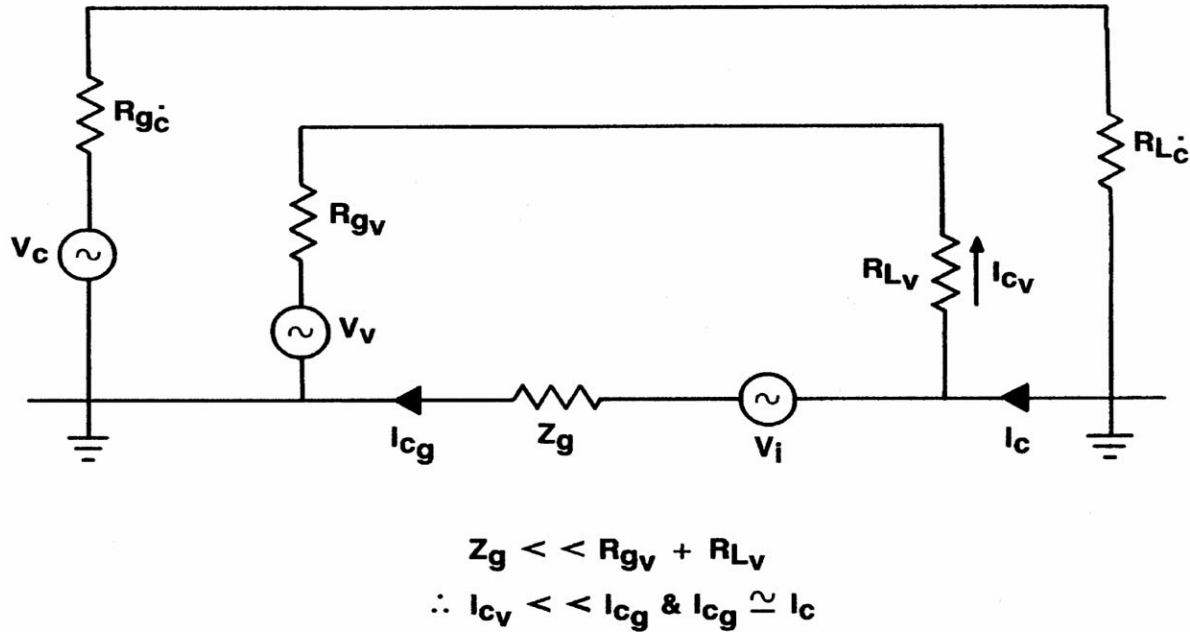
**Figure 109.2** Typical impedance versus frequency behavior of a grounding conductor.



## Ground-Related EMI Coupling

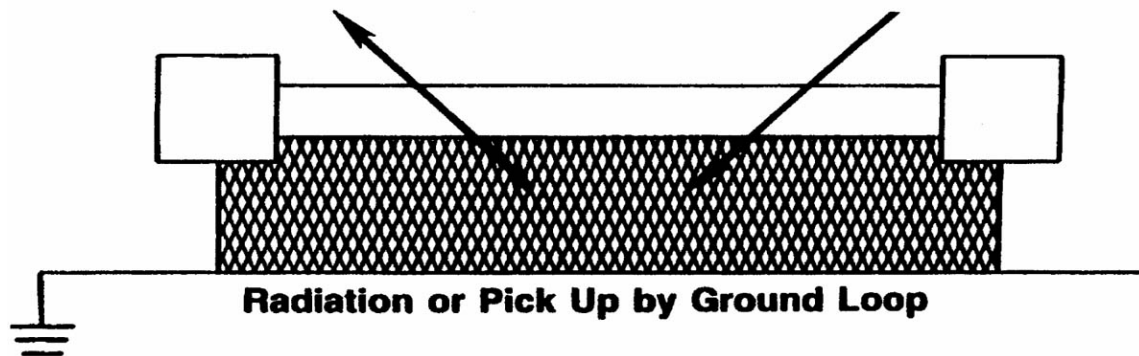
Ground-related EMI involves one of two basic coupling mechanisms. The first mechanism results from circuits sharing the ground with other circuits. [Figure 109.3](#) illustrates EMI coupling between culprit and victim circuits via the common-ground impedance. In this case, the interference current ( $I_{cg}$ ) flowing through the common-ground impedance ( $Z_g$ ) will produce an interfering signal voltage ( $V_i$ ) in the victim circuit. This effect can be reduced by minimizing or eliminating the common-ground impedance.

**Figure 109.3** Common-ground impedance coupling between circuits.



The second EMI coupling mechanism involving ground is a radiated mechanism whereby the ground loop, as shown in [Fig. 109.4](#), acts as a receiving or transmitting antenna. For this EMI coupling mechanism the induced EMI voltage (for the susceptibility case) or the emitted EMI field (for the emission case) is a function of the EMI driving function (field strength, voltage or current), the geometry and dimensions of the ground loop, and the frequency of the EMI signal. Radiated effects can be minimized by routing conductors close to ground and minimizing the ground-loop area.

**Figure 109.4** Common-mode radiation into and from ground loops.



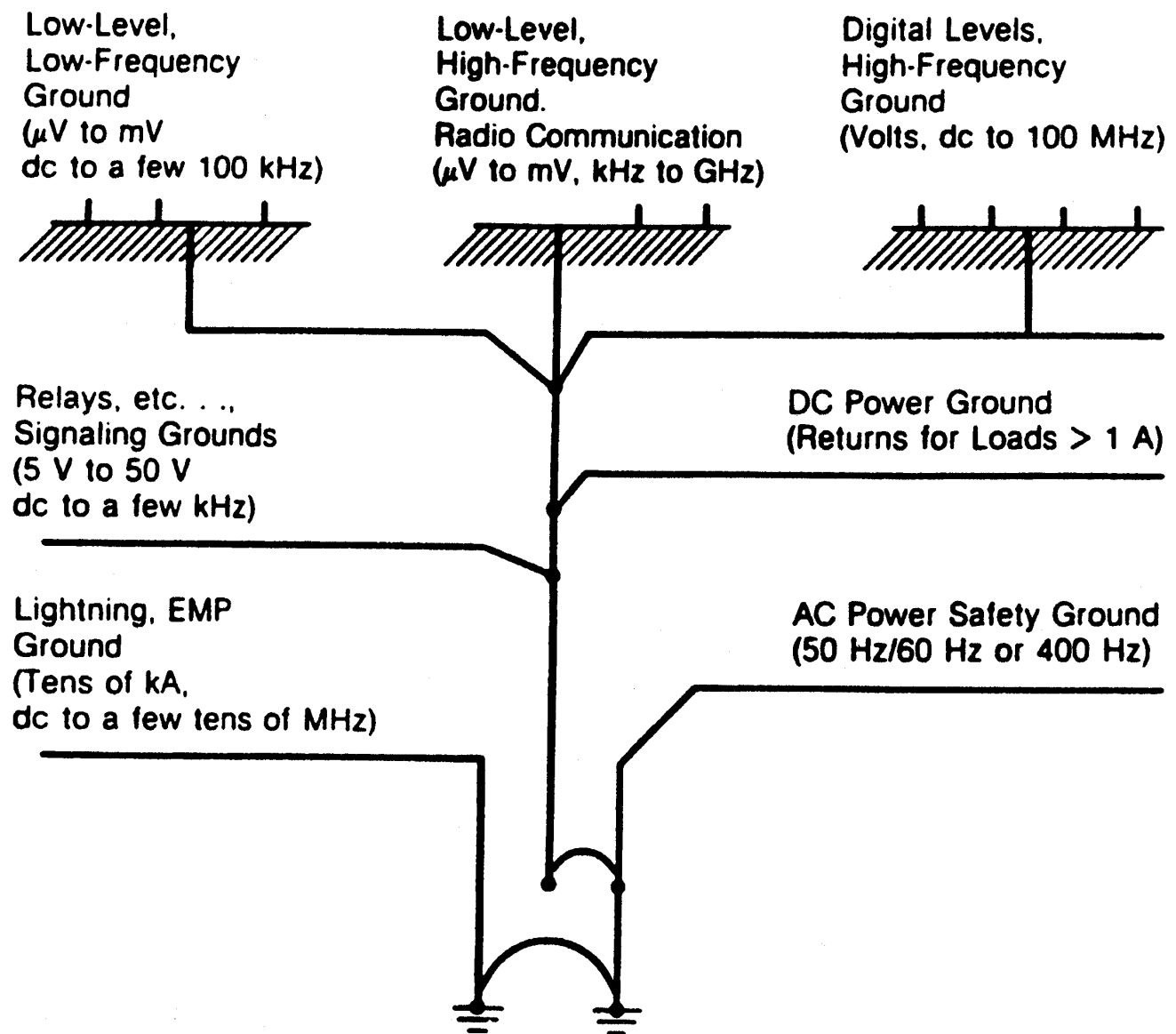


It should be noted that both the conducted and radiated EMI coupling mechanisms identified earlier involve a "ground loop." It is important to recognize that ground loop EMI problems can exist without a physical connection to ground. In particular, at RF frequencies, capacitance-to-ground can create a ground loop condition even though circuits or equipments are floated with respect to ground.

## Grounding Configurations

A typical electronic equipment may have a number of different types of functional signals as shown in [Fig. 109.5](#). To mitigate interference due to common-ground impedance coupling, as many separate grounds as possible should be used.

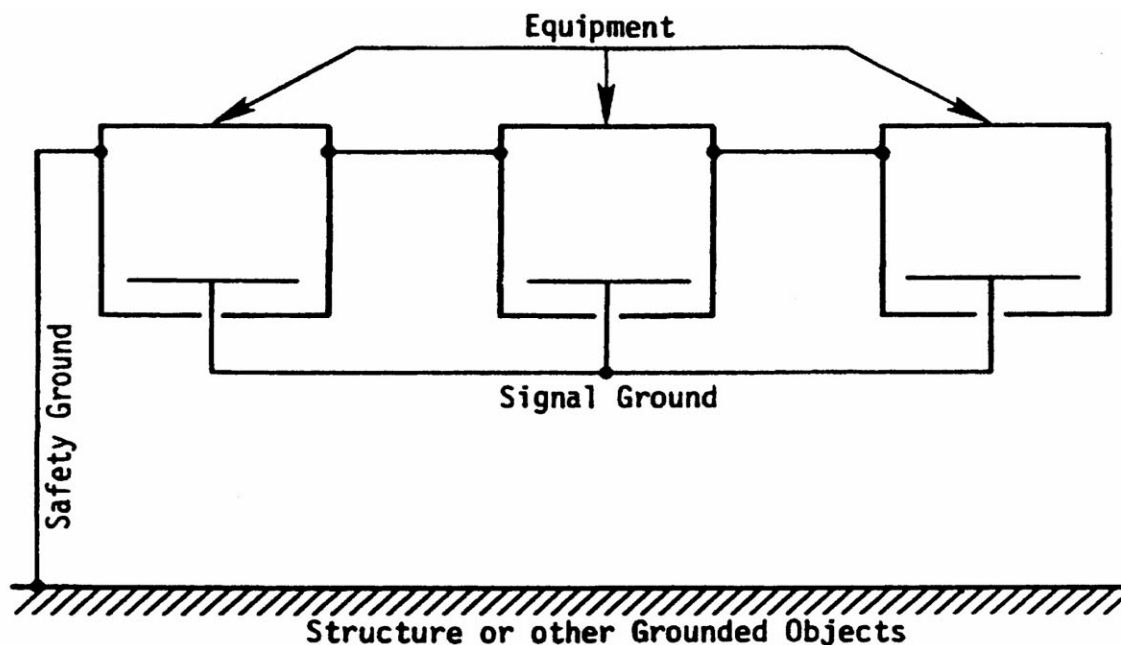
**Figure 109.5** Grounding hierarchy.



The grounding scheme for a collection of circuits within an equipment can assume any one of several configurations. Each of these configurations tends to be optimum under certain conditions and may contribute to EMI problems under other conditions. In general, the ground configurations are a floating ground, a single-point ground, a multiple-point ground, or some hybrid combination.

A floating ground configuration is illustrated in [Fig. 109.6](#). The signal ground is electrically isolated from the equipment ground and other conductive objects. Hence, equipment noise currents present in the equipment and power ground will not be conductively coupled to the signal circuits.

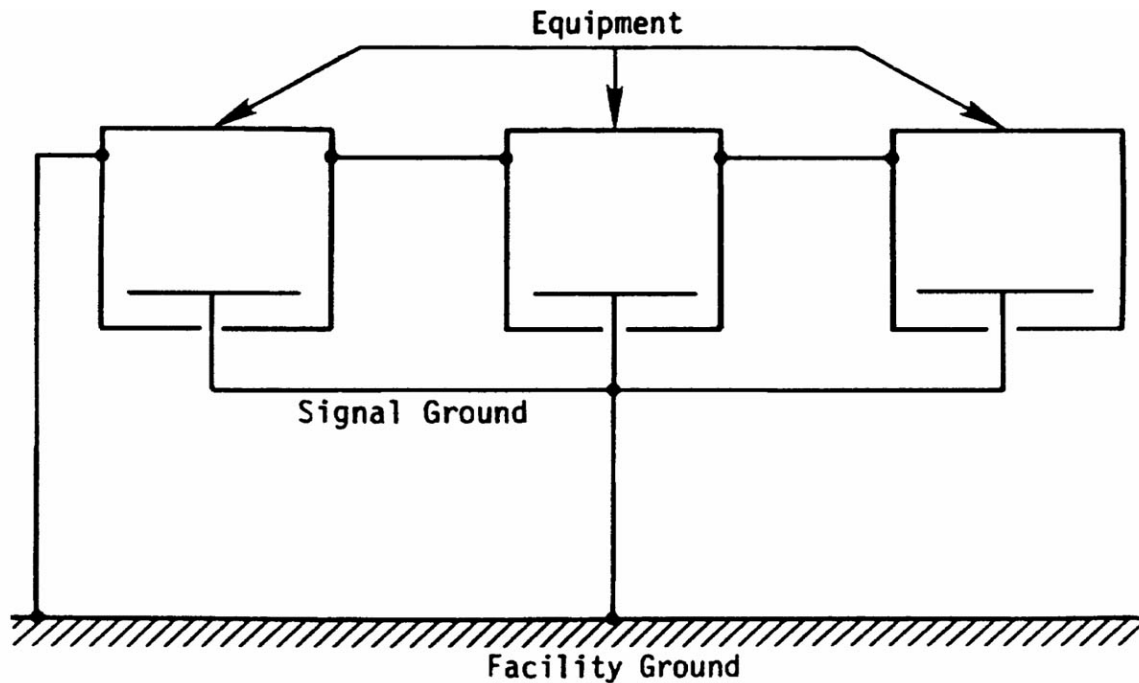
**Figure 109.6** Floating signal ground.



The effectiveness of floating ground configurations depends upon their true isolation from other nearby conductors; that is, to be effective, floating ground systems must really float. It is often difficult to achieve and maintain an effective floating system. A floating ground configuration is most practical if only a few circuits are involved and power is supplied from either batteries or DC-to-DC converters.

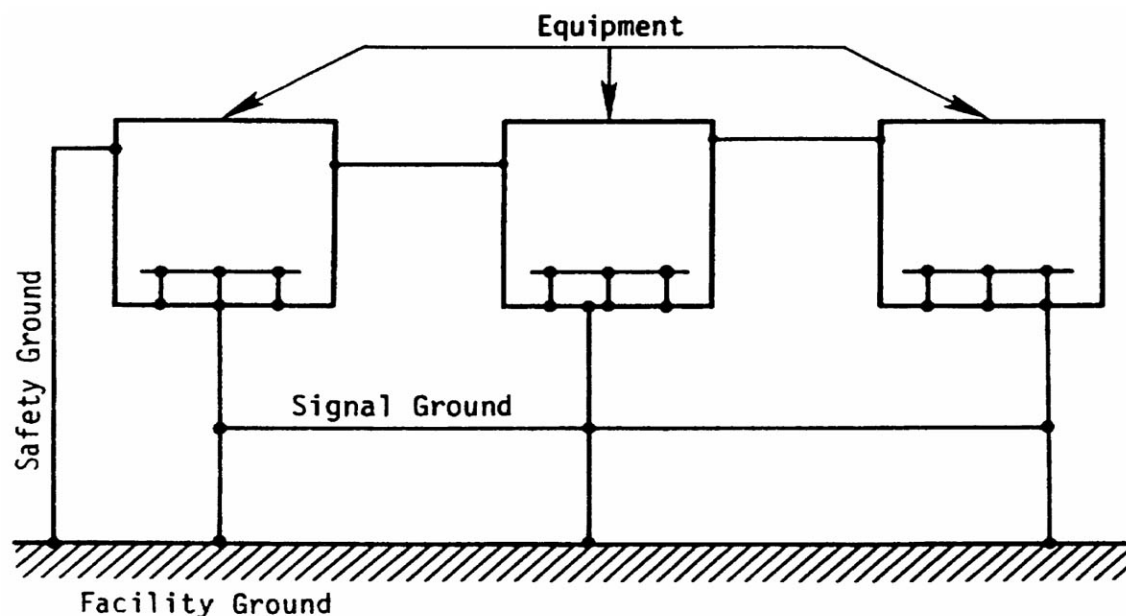
A single-point ground configuration is illustrated in [Fig. 109.7](#). An important advantage of the single-point configuration is that it helps control conductively coupled interference. As illustrated in [Fig. 109.7](#), EMI currents or voltages in the equipment ground are not conductively coupled into the signal circuits via the signal ground. Therefore, the single-point signal ground network minimizes the effects of any EMI currents that may be flowing in the equipment ground.

**Figure 109.7** Single-point signal ground.



The multiple-point ground illustrated in [Fig. 109.8](#) is the third configuration frequently used for signal grounding. With this configuration, circuits have multiple connections to ground. Thus, in an equipment, numerous parallel paths exist between any two points in the multiple-point ground network. Multipoint grounding is more economical and practical for printed circuits and integrated circuits. Interconnection of these components through wafer risers, mother boards, and so forth should use a hybrid grounding approach in which single-point grounding is used to avoid low-frequency ground loops and/or common-ground impedance coupling; multipoint grounding is used otherwise.

**Figure 109.8** Multiple-point ground configuration.



## Summary of Grounding Considerations

A properly designed ground configuration is one of the most important engineering elements in protecting against the effects of EMI. The ground configuration should provide effective isolation between power, digital, high-level analog, and low-level analog signals. In designing the ground it is essential to consider the circuit, signal characteristics, equipment, cost, maintenance, and so forth. In general, either floating or single-point grounding is optimum for low-frequency situations and multiple-point grounding is optimum for high-frequency situations. In many practical applications a hybrid ground approach is employed to achieve the single-point configuration for low frequencies and the multiple-point configuration for high frequencies.

## 109.2 Shielding

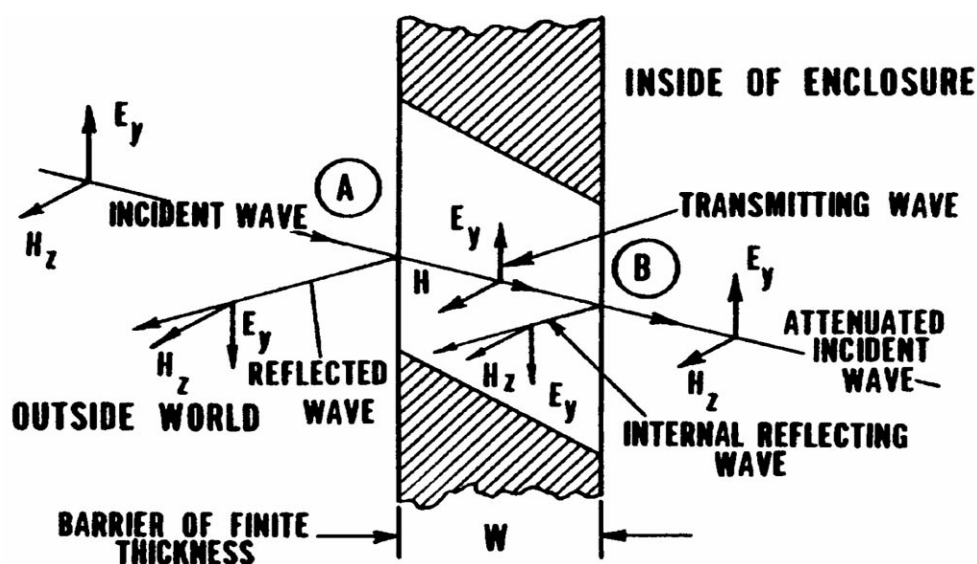
The material on shielding was adapted from Duff [1991] courtesy of Interference Control Technologies.

Shielding is one of the most effective methods for controlling radiated EMI effects at the component, circuit, equipment, subsystem, and system levels. The performance of shields is a function of the characteristics of the incident electromagnetic fields. Therefore, shielding considerations in the near-field region of an EMI source may be significantly different from shielding considerations in the far-field region.

### Shielding Theory

If a metallic barrier is placed in the path of an electromagnetic field as illustrated in Fig. 109.9, only a portion of the electromagnetic field may be transmitted through the barrier. There are several effects that may occur when the incident wave encounters the barrier. First, a portion of the incident wave may be reflected by the barrier. Second, the portion of the incident wave that is not reflected will penetrate the barrier interface and may experience absorption loss while traversing the barrier. Third, additional reflection may occur at the second barrier interface, where the electromagnetic field exits the barrier. Usually this second reflection is insignificant relative to the other effects that occur and may be neglected.

**Figure 109.9** Shielding of plane waves.



The shielding effectiveness of the barrier may be defined in terms of the ratio of the impinging field intensity to the exiting field intensity. For high-impedance electromagnetic fields or plane waves, the shielding effectiveness is given by

$$SE_{dB} = 20 \log \left( \frac{E_1}{E_2} \right) \quad (109.1)$$

where  $E_1$  is the impinging field intensity in volts per meter and  $E_2$  is the exiting field intensity in

The shielding effectiveness of the barrier may be defined in terms of the ratio of the impinging field intensity to the exiting field intensity. For high-impedance electromagnetic fields or plane waves, the shielding effectiveness is given by

$$SE_{dB} = 20 \log \left( \frac{E_1}{E_2} \right) \quad (109.1)$$

where  $E_1$  is the impinging field intensity in volts per meter and  $E_2$  is the exiting field intensity in volts per meter. For low-impedance magnetic fields, the shielding effectiveness is defined in terms of the ratio of the magnetic field strengths.

The total shielding effectiveness of a barrier results from the combined effects of reflection loss and absorption loss. Thus, the shielding effectiveness,  $S$ , in dB is given by

$$S_{dB} = R_{dB} + A_{dB} + B_{dB} \quad (109.2)$$

where  $R_{dB}$  is the reflection loss,  $A_{dB}$  is the absorption loss, and  $B_{dB}$  is the internal reflection loss. Characteristics of the reflection and absorption loss are discussed in the following sections.

## Reflection Loss

When an electromagnetic wave encounters a barrier, a portion of the wave may be reflected. The reflection occurs as a result of a mismatch between the wave impedance and the barrier impedance. The resulting reflection loss,  $R$ , is given by

$$\begin{aligned} R_{dB} &= 20 \log_{10} \frac{(K + 1)^2}{4K}, \quad K = \frac{Z_w}{Z_b} \\ &\simeq 20 \log_{10} \left( \frac{Z_w}{4Z_b} \right), \quad K \geq 10 \end{aligned} \quad (109.3)$$

where  $Z_w$  is the wave impedance  $= E/H$ , and  $Z_b$  is the barrier impedance.

## Absorption Loss

When an electromagnetic wave encounters a barrier, a portion of the wave penetrates the barrier. As the wave traverses the barrier, the wave may be reduced as a result of the absorption loss that occurs in the barrier. This absorption loss,  $A$ , is independent of the wave impedance and may be expressed as follows:

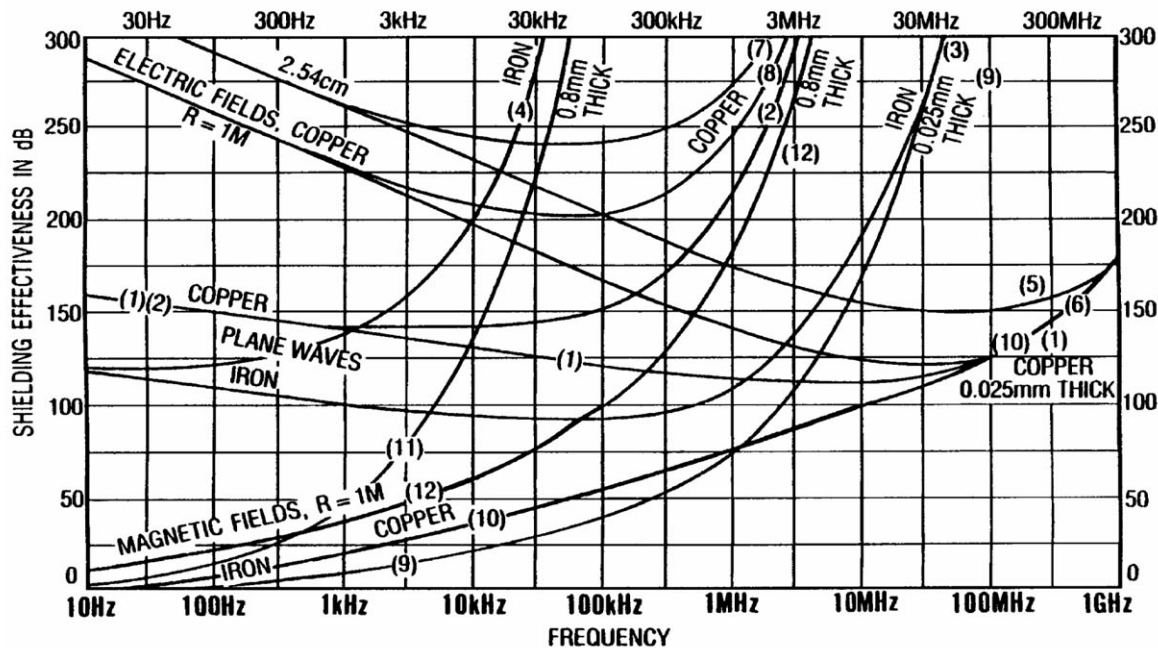
$$A_{dB} = 8.68t/\delta = 131t\sqrt{f_{MHz}\mu_r\sigma_r} \quad (109.4)$$

where  $t$  is the thickness in mm,  $f_{\text{MHz}}$  is the frequency in MHz,  $\mu_r$  is the permeability relative to copper, and  $\sigma_r$  is the conductivity relative to copper.

## Total Shielding Effectiveness

The total shielding effectiveness resulting from the combined effects of reflection and absorption loss are plotted in Fig. 109.10 for copper and iron materials having thicknesses of 0.025 mm and 0.8 mm, having electric and magnetic fields and plane-wave sources, and having source-to-barrier distances of 2.54 cm and 1 meter.

**Figure 109.10** Total shielding effectiveness.



## Shielding Materials

As shown in Fig. 109.10, good shielding efficiency for plane waves or electric (high-impedance) fields is obtained by using materials of high conductivity such as copper and aluminum. However, low-frequency magnetic fields are more difficult to shield because both the reflection and absorption loss of nonmagnetic materials, such as aluminum, may be insignificant. Consequently, to shield against low-frequency magnetic fields, it may be necessary to use magnetic materials.

## Conductive Coatings

Conductive coatings applied to nonconductive materials such as plastics will provide some degree of EMI shielding. The principal techniques for metalizing plastic are the following:

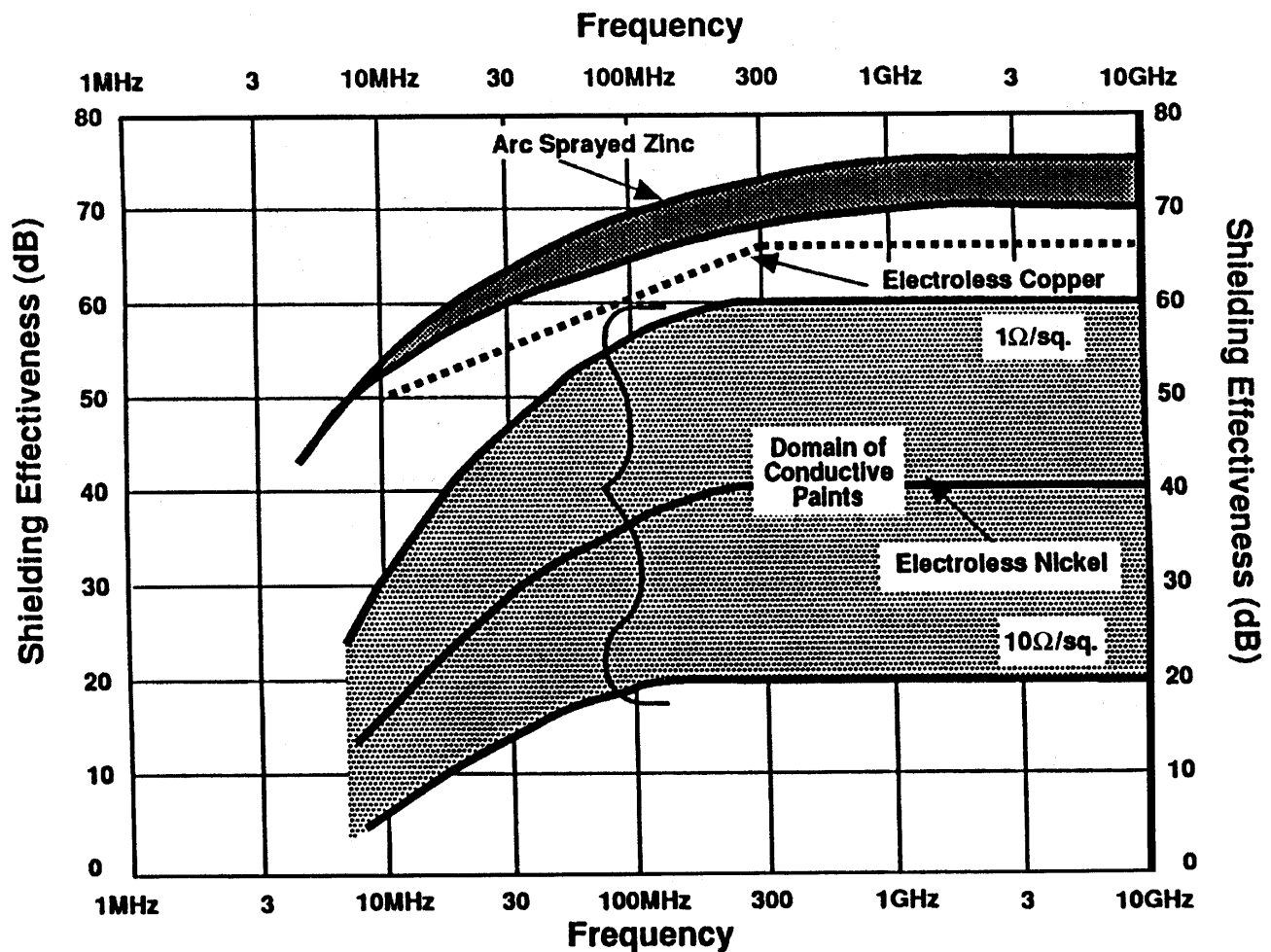
- Conductive paints
- Plating (electrolytic)
- Electroless plating
- Flame spray
- Arc spray
- Ion (plasma torch) spray



- Vacuum deposition

Because the typical conductive coatings provide only a thin film of conductive material, the shielding results from reflection losses that are determined by the ratio of the wave impedance to the conductive barrier impedance. The surface resistance (in ohms per square) will determine shielding effectiveness. Figure 109.11 shows comparative data for shielding effectiveness for various conductive coatings. The most severe situation (i.e., a low-impedance magnetic field source) has been assumed.

**Figure 109.11** Shielding of conductive coatings. (By standard. 30 cm distance test. Near field attenuation is given against H field.) For paints, thickness is typically 2 mil. = .05 mm.



## Aperture Leakages

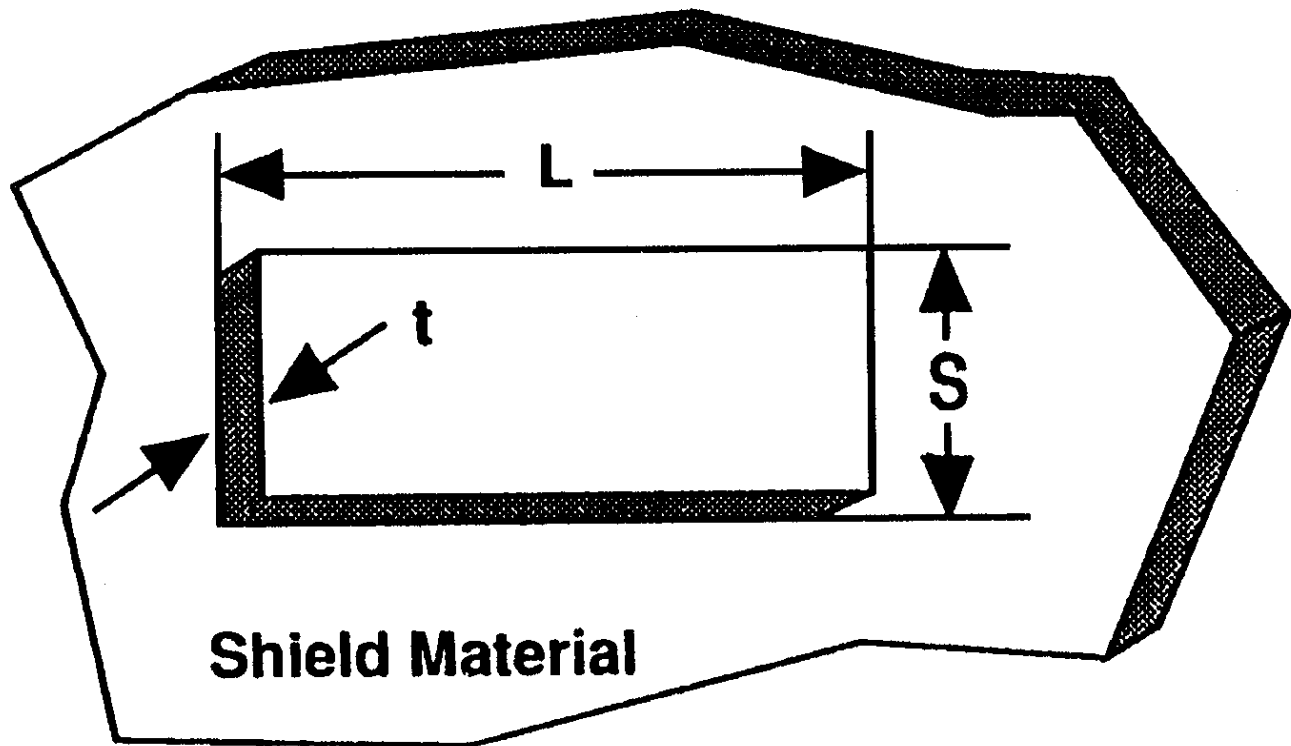
Various shielding materials are capable of providing a high degree of shielding effectiveness under somewhat idealized conditions. However, when these materials are used to construct a shielded housing, the resulting enclosure will typically have holes and seams that may severely compromise

the overall shielding effectiveness.

Figure 109.12 shows a rectangular aperture in a metal (or metalized) panel. A vertically polarized incident electric field will induce currents in the surface of the conductive panel. If the aperture dimensions are much less than a half wavelength, the path around the slot will provide a low impedance to the induced currents and, as a result, the aperture leakage will be small. On the other hand, as the aperture dimensions approach a half wavelength, the path around the slot will provide a high impedance to the induced currents and the aperture leakage will be significant. An aperture with dimensions equal to or greater than a half wavelength will provide almost no shielding (i.e., the incident field will propagate through the aperture with very little loss). In general, the shielding effectiveness of a conductive panel with an aperture may be approximated by the following equation:

$$SE_{dB} \cong 100 - 20 \log L_{mm} \times F_{MHz} + 20 \log \left( 1 + \ln \frac{L}{S} \right) \quad (109.5)$$

**Figure 109.12** Slot and aperture leakage.



To maintain shielding integrity for an equipment enclosure, it may be necessary to provide EMI



protection for the apertures.

## Summary of Shielding Considerations

Shielding can provide an effective means of controlling radiated EMI effects. To ensure that shielding effectiveness requirements are met, it is necessary to

- Select a material that is capable of providing the required shielding
- Minimize the size of openings to control aperture leakages
- Subdivide large openings into a number of smaller ones
- Protect leaky apertures (e.g., cover with wire screen)
- Use EMI gaskets on leaky seams
- Filter conductors at points where they enter or exit a shielded compartment

## Defining Terms

**Ground:** Any reference conductor that is used for a common return.

**Near-field/far-field transition distance:** For electrically small radiators (i.e., dimensions  $\ll$  wavelength ), the near-field/far-field transition occurs at a distance equal to approximately one sixth of a wavelength from the radiating source.

**Plane wave:** Far-field electromagnetic wave with an impedance of 377 ohms in air.

**Reference:** Some object whose potential (often 0 volts with respect to earth or a power supply) is the one to which analog and logic circuits, equipments, and systems can be related or benchmarked.

**Return:** The low (reference) voltage side of a wire pair (e.g., neutral), outer jacket of a coax, or conductor providing a path for intentional current to get back to the source.

**Wavelength:** The distance corresponding to a period for the electromagnetic wave spatial variation. Wavelength (meters) = 300 /frequency (MHz).

## References

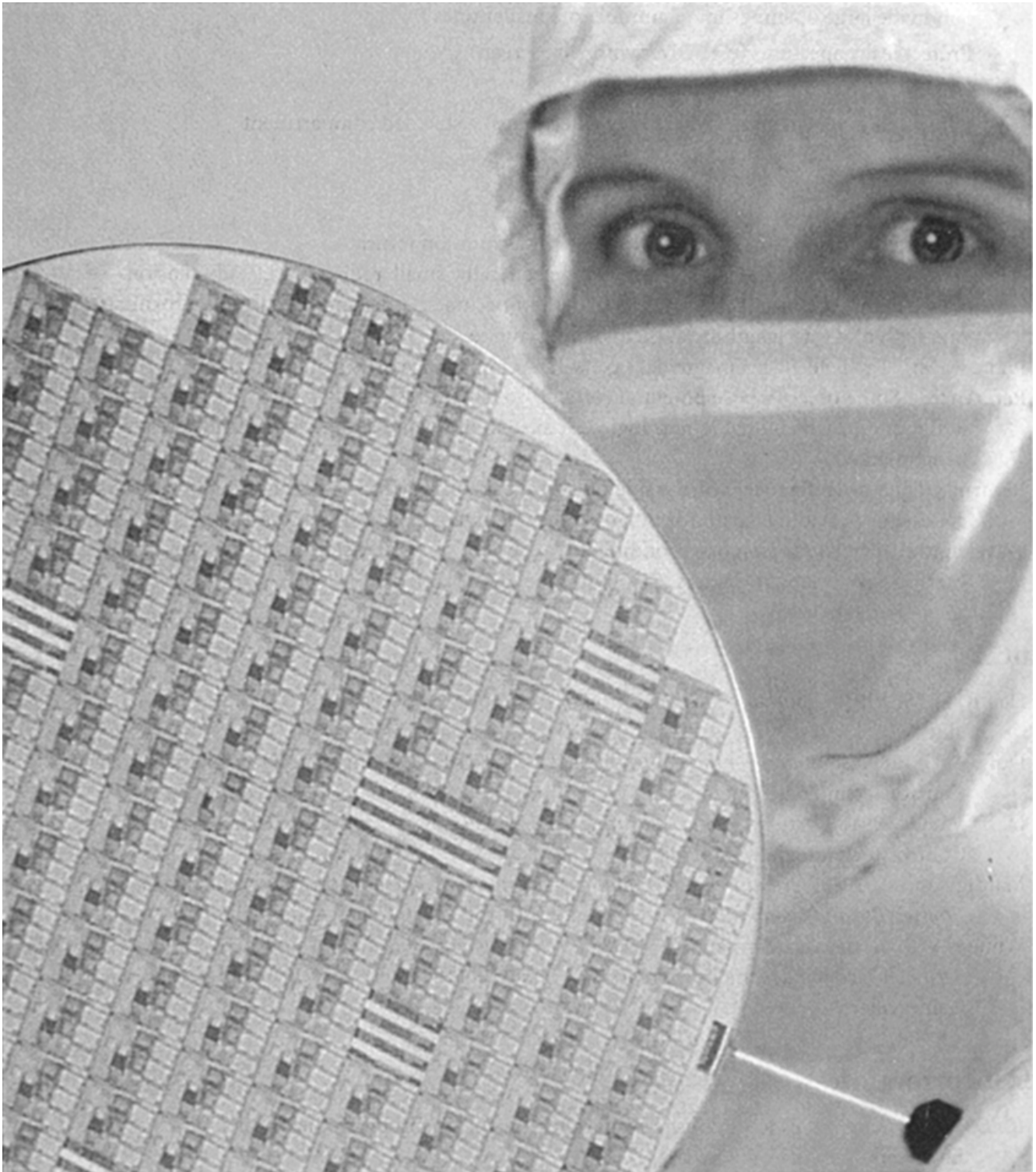
- Denny, H. W. 1983. *Grounding for the Control of EMI*. Interference Control Technologies, Gainesville, VA.
- Duff, W. G. 1989. *Grounding for the Control of EMI*. EMC EXPO 89 Symposium Record. Interference Control Technologies, Gainesville, VA.
- Duff, W. G. 1991. *Electromagnetic Shielding*. EMC EXPO 91 Symposium Record. Interference Control Technologies, Gainesville, VA.
- Mardiguian, M. 1988. *Grounding and Bonding, Volume 2 3/4A Handbook Series on Electromagnetic Interference and Compatibility*. Interference Control Technologies, Gainesville, VA.
- White, D. R. J. and Mardiguian, M. 1988. *Electromagnetic Shielding, Volume 3 3/4A Handbook Series on Electromagnetic Interference and Compatibility*. Interference Control Technologies, Gainesville, VA.

## **Further Information**

*IEEE Transactions on EMC*. Published quarterly by the Institute of Electrical and Electronic Engineers.

*IEEE International EMC Symposium Records*. Published annually by the Institute of Electrical and Electronic Engineers.

Jackson, T. N. "Electronics"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



Increasingly, electronics is dominated by integrated devices and circuits, both analog and digital. The photo above shows a silicon wafer containing advanced application-specific integrated circuits (ASICs). Such circuits can provide analog, digital, or mixed-signal electronic functions from DC to low microwave frequencies. (Photo by Tom Way. Courtesy of IBM)

# XVIII

## Electronics

---

**Thomas N. Jackson**

*Pennsylvania State University*

**110 Operational Amplifiers** *P. J. Hurst*

The Ideal Op Amp • Feedback Circuit Analysis • Input and Output Impedances • Practical Limitations and Considerations

**111 Active RC Filters** *M. A. Soderstrand*

History of Active Filters • Active Filter Design Techniques • Filter Specifications and Approximations • Filter Design

**112 Diodes and Transistors** *S. Soclof*

Semiconductors • Bipolar Junction Transistors • Junction Field-Effect Transistors • Metal-Oxide Silicon Field-Effect Transistors

**113 Analog Integrated Circuits** *S. Soclof*

Operational Amplifiers • Voltage Comparators • Voltage Regulators • Power Amplifiers • Wide-Bandwidth (Video) Amplifiers • Modulators, Demodulators, and Phase Detectors • Voltage-Controlled Oscillators • Waveform Generators • Phase-Locked Loops • Digital-to-Analog and Analog-to-Digital Converters • Radio-Frequency Amplifiers • Integrated Circuit Transducers

**114 Optoelectronic Devices** *P. Bhattacharya*

Light-Emitting Diodes • Lasers • Photodetectors • Conclusion

**115 Power Electronics** *K. S. Rajashekara*

Power Semiconductor Devices • Power Conversion

**116 A/D and D/A Converters** *J. C. Hamann*

The Fundamentals of D/A Converters • The Fundamentals of A/D Converters

**117 Superconductivity** *K. A. Delin and T. P. Orlando*

Introduction • General Electromagnetic Properties • Superconducting Electronics • Types of Superconductors

THE TERM *ELECTRONICS* IS USED in the field of electrical engineering to distinguish devices and phenomena that depend on the motion of charge carriers in a semiconductor, vacuum, or gas, from those that depend on the motion of charge carriers in a metal. The latter is where electrical engineering began, with motors, generators, light bulbs, telegraphy, and so on, but today electronics is by far the dominant area of interest and activity. Perhaps more than any other area of technology, it is advances in electronics that have provided the foundation on which our modern world is built. From personal computers to computer-controlled automobiles to advanced medical technology to fly-by-wire aircraft to space technology to high-bandwidth communications to a vast array of consumer products, it is the devices and fundamental circuits of electronics that form the essential foundation on which the technology is built.

This section will consider the most important building blocks of electronics. Although historically these building blocks have been discrete electronic devices (first vacuum tubes and later semiconductor diodes and transistors of various types), today the building blocks may be more complex and be themselves built up from several simpler electronic devices. One of the most important examples is the operational amplifier, or op amp, and this section begins with this device and also introduces the important approach of treating building blocks with considerable internal complexity as simple electronic devices. The section continues by considering an important op amp application—active filters—which may themselves be considered as individual electronic building blocks for application in more complex systems.

It is this hierarchical approach of assembling electronic devices into building blocks, which then may themselves be considered as the "devices" for constructing more complicated blocks, that makes it possible to manage the incredible complexity of modern electronic systems. This theme is continued by a chapter that discusses diodes and transistors (the most fundamental electronic device building blocks), followed by a chapter that describes how these devices are used in general to form analog integrated circuits (of which the op amp is just one example). Next, less familiar, but no less important, opto-electronic and power electronic devices are discussed, followed by a chapter that describes how electronics is used to make the link between the analog and digital worlds. Finally, a chapter on superconductivity is included to give insight into another important aspect of the continually broadening field of electronics.

Hurst, P. J. "Operational Amplifiers"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

- 110.1 The Ideal Op Amp
- 110.2 Feedback Circuit Analysis
- 110.3 Input and Output Impedances
- 110.4 Practical Limitations and Considerations

**Paul J. Hurst**

*University of California, Davis*

The operational amplifier (op amp) is one of the most versatile building blocks for analog circuit design. The earliest op amps were constructed of discrete devices—first vacuum tubes and later transistors. Today, an op amp is constructed of numerous transistors along with a few resistors and capacitors, all fabricated on a single piece of silicon. Parts are sold with one, two, or four op amps in a small integrated circuit package.

Op amps are popular because they are small, versatile, easy to use, and inexpensive. Despite their low cost, modern op amps offer superb performance. An op amp provides high voltage gain, high input impedance, and low output impedance. Many op amps are commercially available, and they differ in specifications such as input noise, bandwidth, gain, offset voltage, output swing, and supply voltage.

Op amps are used as high-gain amplifiers in negative feedback circuits. An advantage of such circuits is that design and analysis is relatively simple, with the overall gain depending only on external passive components that provide the feedback around the op amp. In the following sections, analysis and design of op-amp feedback circuits are covered.

---

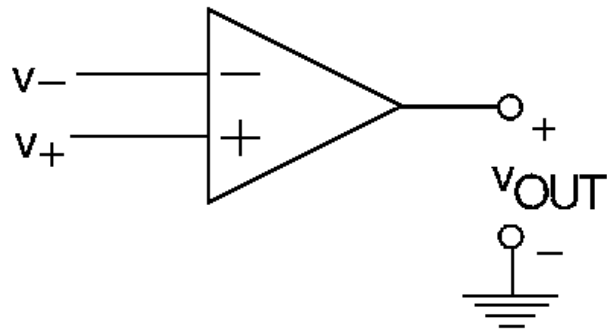
## 110.1 The Ideal Op Amp

---

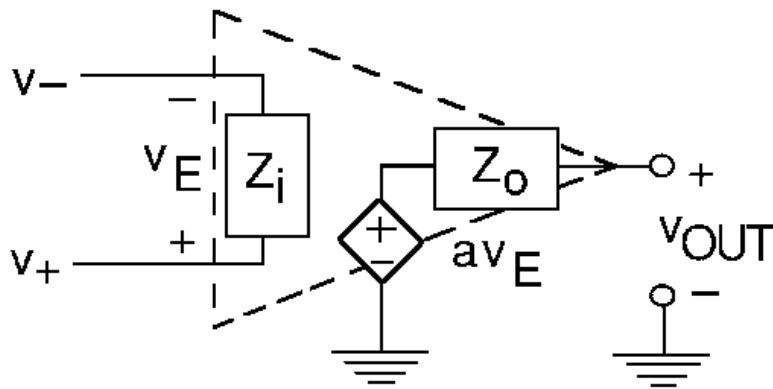
The schematic symbol for an op amp is shown in [Fig. 110.1\(a\)](#). The op amp has two inputs,  $v_+$  and  $v_-$ , and a single output. The voltages  $v_+$ ,  $v_-$ , and  $v_{\text{OUT}}$  are measured with respect to ground. The op amp amplifies the voltage difference  $v_+ - v_-$  to produce the output voltage  $v_{\text{OUT}}$ . A simple model for the op amp that is valid when the op amp is operating in its linear high-gain region is shown in [Fig. 110.1\(b\)](#). The model consists of an input impedance  $Z_i$ , output impedance  $Z_o$ , and voltage gain  $a$ . Here, the voltage difference  $v_+ - v_-$  is called  $v_E$ . If  $Z_i$ ,  $Z_o$ , and  $a$  are known, this model can be used to analyze an op-amp circuit if the op amp is biased in the linear region, which is usually the case when the op amp is in a negative feedback loop. A typical op amp has large  $Z_i$ , small  $Z_o$ , and large  $a$ . As a result, the op-amp model is often further simplified by setting  $Z_i = \infty$  and  $Z_o = 0$ , as shown in [Fig. 110.1\(c\)](#).



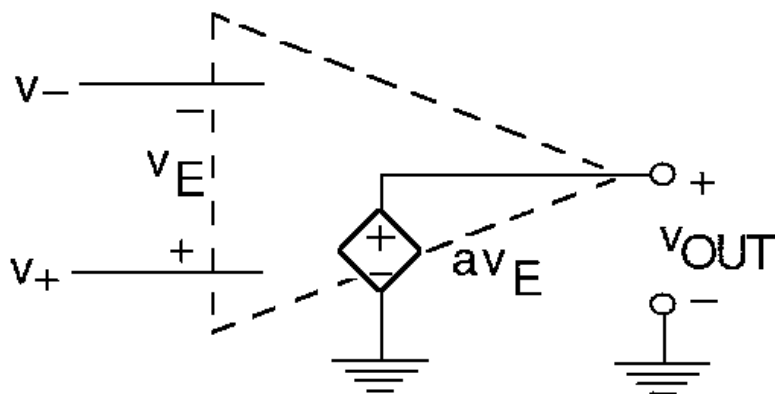
**Figure 110.1** (a) The schematic symbol for an op amp; (b) a simple three-element model for the op amp; (c) the model in (b) simplified further by setting  $Z_i = \infty$  and  $Z_o = 0$ .



(a)



(b)



(c)

For an example of an op-amp feedback circuit, consider the inverting gain amplifier shown in Fig. 110.2. Using the model of Fig. 110.1(c) in Fig. 110.2, the gain can be found by summing currents at the  $v_-$  input of the op amp. Since the current into the op amp is zero (due to  $Z_i = \infty$ ),  $i_1 = i_2$ ; therefore,

$$\frac{v_{\text{IN}} - (-v_E)}{R_1} = i_1 = i_2 = \frac{-v_E - v_{\text{OUT}}}{R_2} \quad (110.1)$$

Using the relationship  $v_E = v_{\text{OUT}}/a$ , Eq. (110.1) yields

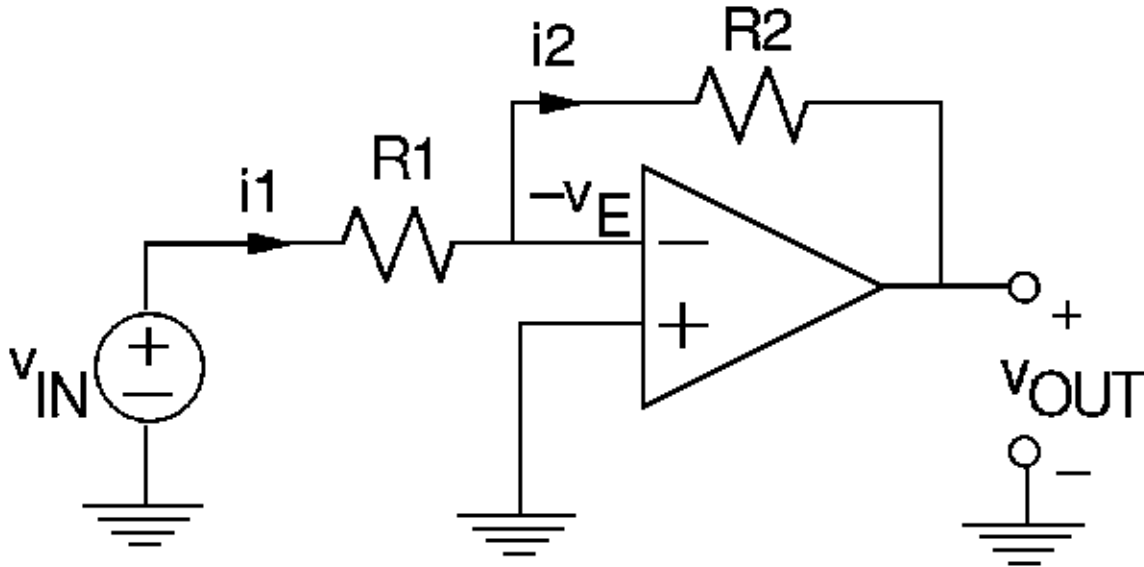
$$\frac{v_{\text{OUT}}}{v_{\text{IN}}} = -\frac{R_2}{R_1} \frac{1}{1 + (R_1 + R_2)/aR_1} \quad (110.2)$$

Assuming that the op-amp gain is very large ( $a \rightarrow \infty$ ), Eq. (110.2) becomes

$$\frac{v_{\text{OUT}}}{v_{\text{IN}}} = -\frac{R_2}{R_1} \quad (110.3)$$

The gain depends on the ratio of resistors and is independent of the op-amp parameters.

**Figure 110.2** An inverting gain amplifier.



The above analysis can be made easier by assuming that the op amp is ideal ( $Z_i = \infty$ ,  $Z_o = 0$ , and  $a = \infty$ ) before beginning the analysis. This **ideal op-amp model** greatly simplifies analysis and gives results that are surprisingly accurate. The assumption  $Z_i = \infty$  ensures that the currents flowing into the op-amp input terminals are zero. With  $Z_o = 0$ , the controlled source directly controls the output, giving  $v_{\text{OUT}} = av_E$ . The final assumption  $a = \infty$  leads to  $v_E = v_{\text{OUT}}/a = 0$  if  $v_{\text{OUT}}$  is bounded, which is typically true in negative feedback circuits. The condition  $v_E = 0$  is referred to as a *virtual short circuit* because the op-amp input voltages are equal ( $v_+ = v_-$ ), even though these inputs are not actually connected together. Negative feedback forces  $v_E$  to be equal to zero.

To demonstrate the advantage of the ideal op-amp model, consider the circuit in [Fig. 110.3](#). If the op amp is ideal,  $v_+ = v_-$ , and therefore  $v_- = v_{\text{IN}}$  as a result of the virtual short circuit. Since no current flows into the op-amp input, the currents  $i_3$  flowing through resistor  $R_3$  and  $i_4$  through  $R_4$  must be equal:

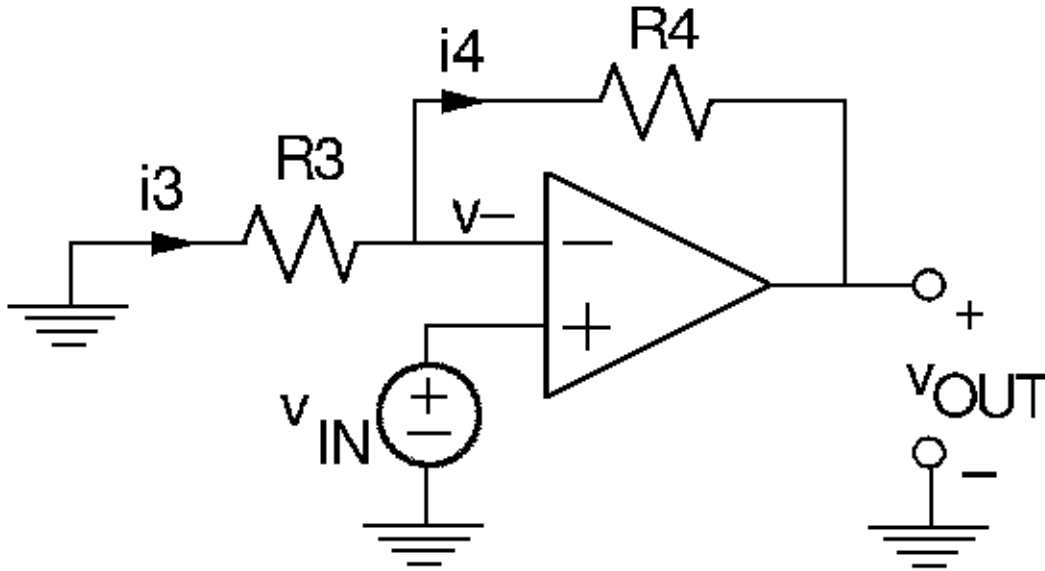
$$\frac{0 - v_{\text{IN}}}{R_3} = i_3 = i_4 = \frac{v_{\text{IN}} - v_{\text{OUT}}}{R_4} \quad (110.4)$$

The resulting gain is positive, or noninverting, and is given by

$$\frac{v_{\text{OUT}}}{v_{\text{IN}}} = 1 + \frac{R_4}{R_3} \quad (110.5)$$

This example demonstrates how the ideal op-amp model can be used to quickly analyze op-amp feedback circuits.

**Figure 110.3** A noninverting gain amplifier.

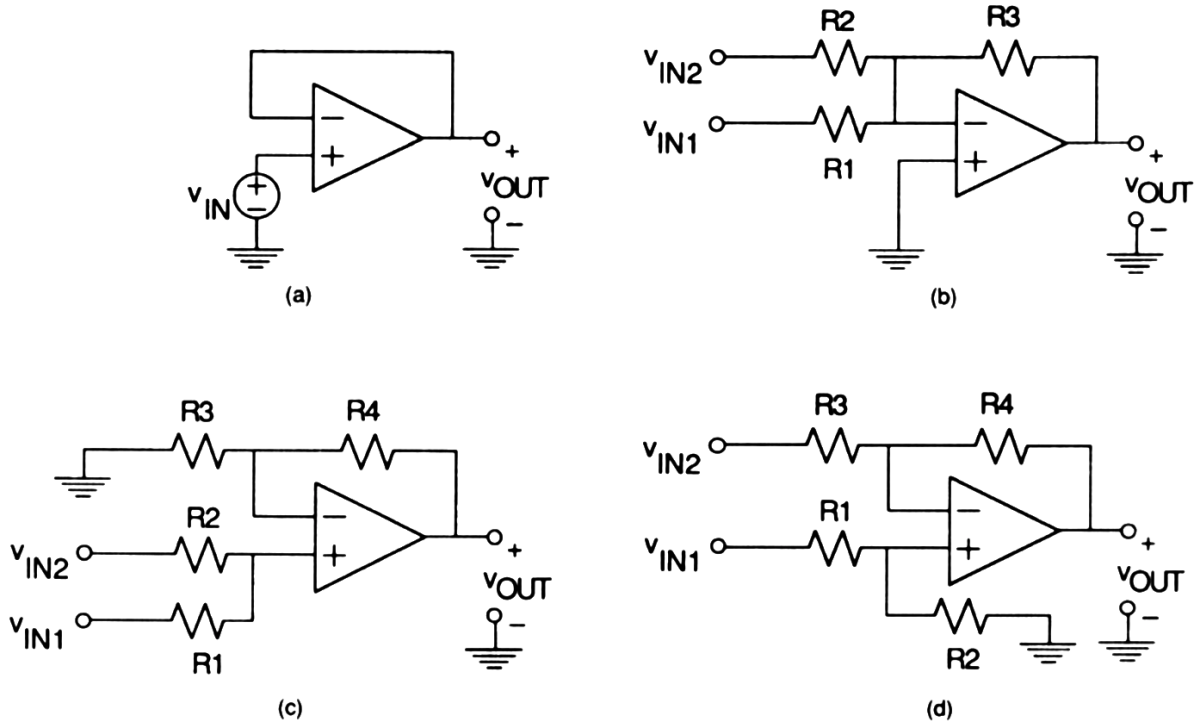


A number of interesting op-amp feedback circuits are shown in Fig. 110.4. The corresponding expressions for  $v_{OUT}$  are given in the caption for Fig. 110.4, assuming an ideal op amp. The simplest circuit is the voltage buffer of Fig. 110.4(a), which has a voltage gain of one. The circuits in Fig. 110.4(b–d) have multiple inputs. Figure 110.4(b) is an inverting, summing amplifier if  $R_1 = R_2$ . Figure 110.4(c) is a noninverting, summing amplifier if  $R_1 = R_2$ . If  $R_1/R_2 = R_3/R_4$  in Fig. 110.4(d), this circuit is a differencing amplifier with

$$v_{OUT} = \frac{R_4}{R_3}(v_{IN1} - v_{IN2}) \quad (110.6)$$

**Figure 110.4** (a) A voltage buffer,  $v_{OUT} = v_{IN}$ ; (b) a two-input inverting gain amplifier,  $v_{OUT} = -(R_3/R_1)v_{IN1} - (R_3/R_2)v_{IN2}$ ; (c) a two-input noninverting stage,  $v_{OUT} = \{[R_2/(R_1 + R_2)]v_{IN1} + [R_1/(R_1 + R_2)]v_{IN2}\}(R_3 + R_4)/R_3$ ; (d) a differencing amplifier,  $v_{OUT} = -(R_4/R_3)v_{IN2} + [R_2/(R_1 + R_2)][(R_3 + R_4)/(R_3)]v_{IN1}$ .

**Figure 110.4**



Op amps can be used to construct filters for signal conditioning. If  $R_1$  in Fig. 110.2 is replaced by a capacitor  $C_1$ , the circuit becomes a differentiator and its input and output are related by either of the following:

$$v_{OUT}(t) = -R_2 C_1 \frac{dv_{IN}(t)}{dt} \quad (110.7a)$$

$$V_{OUT}(s) = -s R_2 C_1 V_{IN}(s) \quad (110.7b)$$

where  $s$  is the Laplace operator and  $V(s)$  is the Laplace transform of  $v(t)$ .

Again, starting with Fig. 110.2, if  $R_2$  is replaced by a capacitor  $C_2$ , the circuit becomes an integrator. Its input and output are related by either of the following:

$$v_{OUT}(t) = -\frac{1}{R_1 C_2} \int_0^t v_{IN}(\tau) d\tau \quad (110.8a)$$

$$V_{OUT}(s) = -\frac{V_{IN}(s)}{s R_1 C_2} \quad (110.8b)$$

Here,  $v_{OUT}(t = 0) = 0$  is assumed. Since there is no DC feedback from the op-amp output to its input through capacitor  $C_2$ , an integrator will only work properly when used in an application that

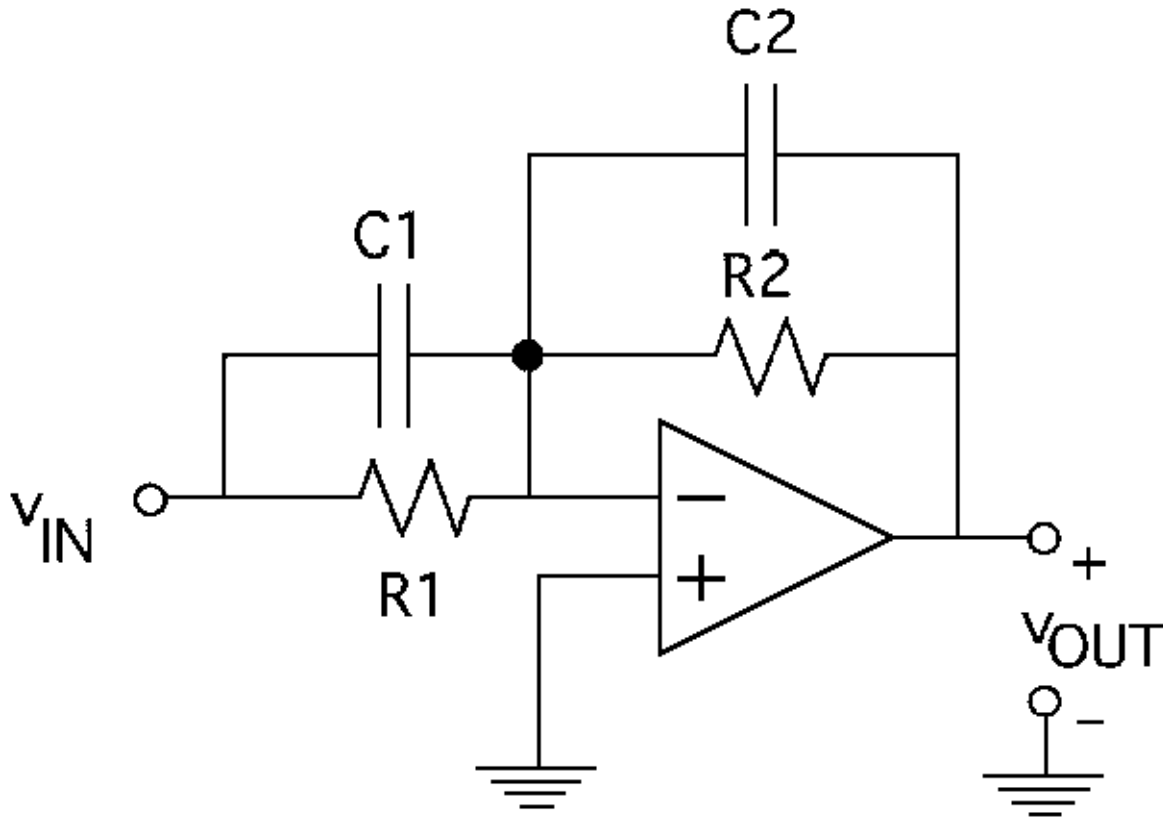
provides DC feedback around the integrator for biasing. Integrators are often used inside feedback loops to construct filters or control loops.

Figure 110.5 is a one-pole, one-zero filter with transfer function

$$\frac{V_{\text{OUT}}(s)}{V_{\text{IN}}(s)} = -\frac{R_2}{R_1} \frac{1 + sR_1C_1}{1 + sR_2C_2} \quad (110.9)$$

Higher-order filters can be constructed using op amps [Jung, 1986].

**Figure 110.5** A one-pole, one-zero filter.



## 110.2 Feedback Circuit Analysis

The goal of every op-amp feedback circuit is an input/output relationship that is independent of the op amp itself. An exact formula for the closed-loop gain  $A$  of an op-amp feedback circuit (e.g., Fig. 110.2) is given by Rosenstark [1986]:

$$A = \frac{v_{\text{OUT}}}{v_{\text{IN}}} = A_{\infty} \frac{\text{RR}}{1 + \text{RR}} + \frac{d}{1 + \text{RR}} \quad (110.10)$$

Here,  $A_{\infty}$  is the gain when the op-amp gain  $a = \infty$ . The term  $d = v_{\text{OUT}}/v_{\text{IN}}|_{a=0}$  accounts for feed-forward directly from input to output; typically,  $d$  is zero or close to zero. RR is the return ratio for the controlled source  $a$ . From Eq. (110.10), it can be seen that gain  $A$  approaches the ideal value  $A_{\infty}$  as RR approaches infinity. The return ratio for the controlled source  $a$  can be found by (1) setting all independent sources to zero, (2) breaking the connection between the controlled source  $a$  and the rest of the circuit, (3) driving the circuit at the break with a voltage source with value  $v_t$ , and (4) finding the resulting voltage  $v_r$  across the dependent source. Then  $\text{RR} = -v_r/v_t$ . For negative feedback, RR is positive. Ideally, RR is large so that the gain  $A$  is close to  $A_{\infty}$ .

For example, consider Fig. 110.2. Using the op-amp model of Fig. 110.1(c), the return ratio for the dependent source  $a$  can be found using Fig. 110.6, and the result is

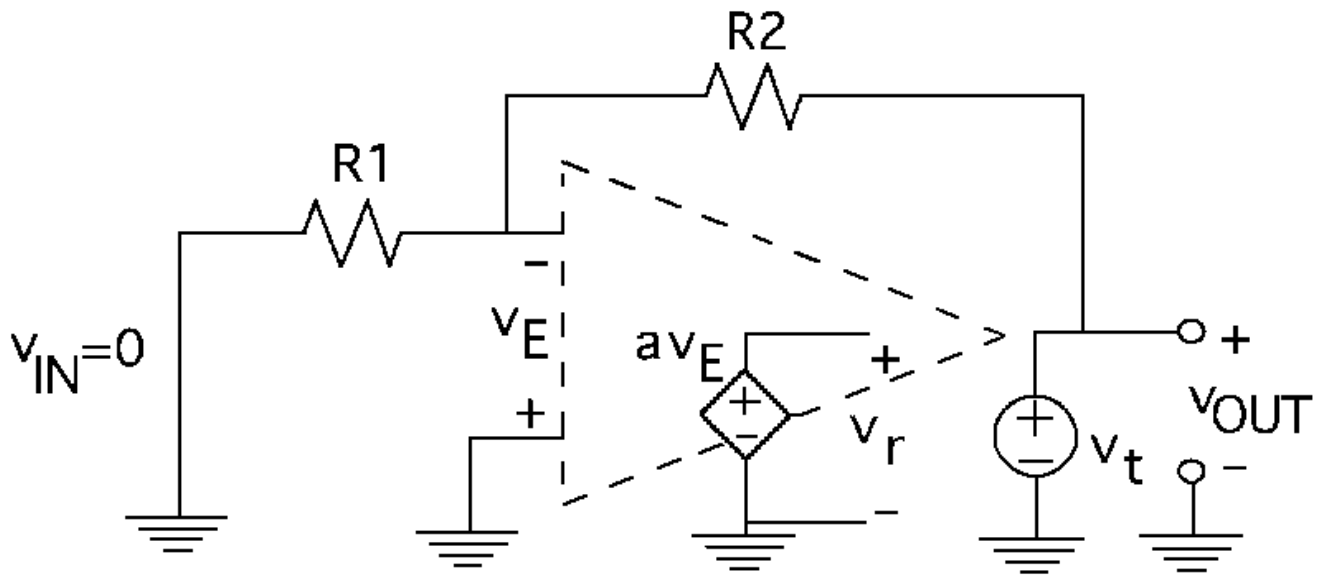
$$\text{RR} = -\frac{v_r}{v_t} = a \frac{R_1}{R_1 + R_2} \quad (110.11)$$

With this model,  $d$  is zero, since setting  $a$  equal to 0 forces  $v_{\text{OUT}}$  equal to 0. Therefore Eq. (110.10) simplifies to

$$A = \frac{v_{\text{OUT}}}{v_{\text{IN}}} = A_{\infty} \frac{\text{RR}}{1 + \text{RR}} \quad (110.12)$$

The RR in Eq. (110.11) can be used in Eq. (110.10) to determine  $A$  at DC or as a function of frequency if  $a(s)$  is known. Information on  $a(s = j\omega)$  is usually provided graphically on a data sheet.

**Figure 110.6** Figure 110.2 modified to allow calculation of the return ratio.



Since op-amp circuits use feedback, stability is a concern. In circuits with passive feedback and load elements, the terms  $A_\infty$ ,  $RR$ , and  $d$  in Eq. (110.10) are stable (all poles are in the left half of the  $s$  plane). Therefore, stability of the closed-loop transfer function  $A(s)$  is determined by the location of the zeroes of  $[1 + RR(s)]$ , which are poles of  $A(s)$ . The location of the zeroes of  $[1 + RR(s)]$  can be determined by examining the phase and gain margins of  $RR$  [Rosenstark, 1986]. The **phase margin** is  $180^\circ - [\text{the phase of } RR(s = j\omega_U)]$ , where  $\omega_U$  is the frequency where the magnitude of  $RR$  is unity. The **gain margin** is  $-20 \log_{10} |RR(s = j\omega_{180})|$ , where  $\omega_{180}$  is the frequency where the phase of  $RR$  is  $-180^\circ$ . The phase and gain margins are positive for a stable circuit. Using an accurate frequency-dependent model for the op amp, or using the frequency response plot of  $a(s = j\omega)$  on the data sheet, the gain and phase margins can be found. Roughly,  $A(s = j\omega) = A_\infty$  for frequencies below  $\omega_U$ , and  $A(s = j\omega)$  will deviate from  $A_\infty$  at frequencies near and above  $\omega_U$ .

Most op amps are designed to be unity-gain stable—that is, they are stable when connected as a buffer [see Fig. 110.4(a)]. In that configuration,  $RR(s) = a(s)Z_i/(Z_i + Z_o) \approx a(s)$  (the approximation follows from  $|Z_o| \ll |Z_i|$ ).

## 110.3 Input and Output Impedances

The input impedance of an op-amp circuit can be found by applying a test voltage source across the input port and measuring the current that flows into the port. The applied voltage divided by the resulting current is the input impedance. (When computing the output impedance, the input source(s) must be set to zero and then the same procedure is carried out on the output port.) Using the model in Fig. 110.1(b), the input impedance for the voltage buffer in Fig. 110.4(a) can be found by following these steps, and the result is

$$Z_{in}(\text{with feedback}) = (Z_i + Z_o) \cdot \left(1 + a \frac{Z_i}{Z_o + Z_i}\right) \approx aZ_i \quad (110.13)$$



where the approximation is valid since  $|Z_o| \ll |Z_i|$  and  $a \gg 1$ . The negative feedback increases the input impedance from  $Z_i$  to  $aZ_i$ . For the OP-07 op amp (see Table 110.1),  $Z_i = 60 \text{ M}\Omega$  and  $a = 500\,000$  at DC, giving a remarkably large  $Z_{in}(\text{with feedback}) = 3 \cdot 10^{13} \Omega$  at DC. Calculation of the output impedance in Fig. 110.4(a) gives

$$Z_{out}(\text{with feedback}) = \frac{Z_i \parallel Z_o}{1 + a(Z_i)/(Z_o + Z_i)} \approx \frac{Z_o}{a} \quad (110.14)$$

where  $x \parallel y = xy/(x + y)$ , and  $Z_i \parallel Z_o \approx Z_o$  because  $|Z_o| \ll |Z_i|$ . The feedback reduces the output impedance. Again, using values for the OP-07,  $Z_{out}(\text{with feedback}) = 0.1 \text{ m}\Omega$  at DC.

**Table 110.1** Typical Op-Amp Specifications

Part #	$a$ at DC	$R_i = Z_i$ (DC)	$R_o = Z_o$ (DC)	Output Swing	$ V_{os} $	$I_B$	Bandwidth for $A_{\infty} = 1$	Slew Rate
OP-07	$5 \cdot 10^5$	$60 \text{ M}\Omega$	$60 \Omega$	$\pm 14 \text{ V}$	$30 \mu\text{V}$	$1.8 \text{ nA}$	$0.6 \text{ MHz}$	$0.3 \text{ V}/\mu\text{s}$
LF411	$2 \cdot 10^5$	$10^{12} \Omega$	$40 \Omega$	$\pm 13.5 \text{ V}$	$0.8 \text{ mV}$	$50 \text{ pA}$	$4 \text{ MHz}$	$15 \text{ V}/\mu\text{s}$
OP-177F	$1.2 \cdot 10^7$	$45 \text{ M}\Omega$	$60 \Omega$	$\pm 12.5 \text{ V}$	$10 \mu\text{V}$	$1.2 \text{ nA}$	$0.6 \text{ MHz}$	$0.3 \text{ V}/\mu\text{s}$

Supply voltage =  $\pm 15 \text{ V}$ ; temperature =  $25^\circ \text{C}$

In some cases, the ideal op-amp model can be used to quickly determine the input impedance. For example, the input resistance in Fig. 110.2 is approximately  $R_1$  because feedback forces  $v_-$  to be close to 0. The output impedance of an ideal op-amp circuit is zero because  $Z_o = 0$ . The actual value can be found by using the model in Fig. 110.1(b) and carrying out the computation described earlier.

## 110.4 Practical Limitations and Considerations

The ideal op-amp model is easy to use, but practical limitations are not included in this simple model. Specifications for a few popular op amps are given in Table 110.1 [[Analog Devices, 1992](#); [National Semiconductor, 1993](#)]. The LF411 uses field-effect input transistors to give extremely high-input resistance and low-input bias current. The input bias current  $I_B$  is the DC current that flows into the op-amp inputs; this is bias current required by the input transistors in the op amp. (This current is not included in the simple models of Fig. 110.1, and it cannot be computed using  $Z_i$ .)

The output voltage swing over which the linear model is valid is limited by the power supply voltage(s) and the op-amp architecture. The peak output current is limited by internal bias currents and/or protection circuitry. The output slew rate, which is the maximum slope of the output signal before distortion occurs, is limited by op-amp internal currents and capacitances. The common-mode input voltage range is the range of the common-mode (or average) input voltage, which is  $(v_+ + v_-)/2$ , for which the linear op-amp model is valid. In Fig. 110.2, it is important

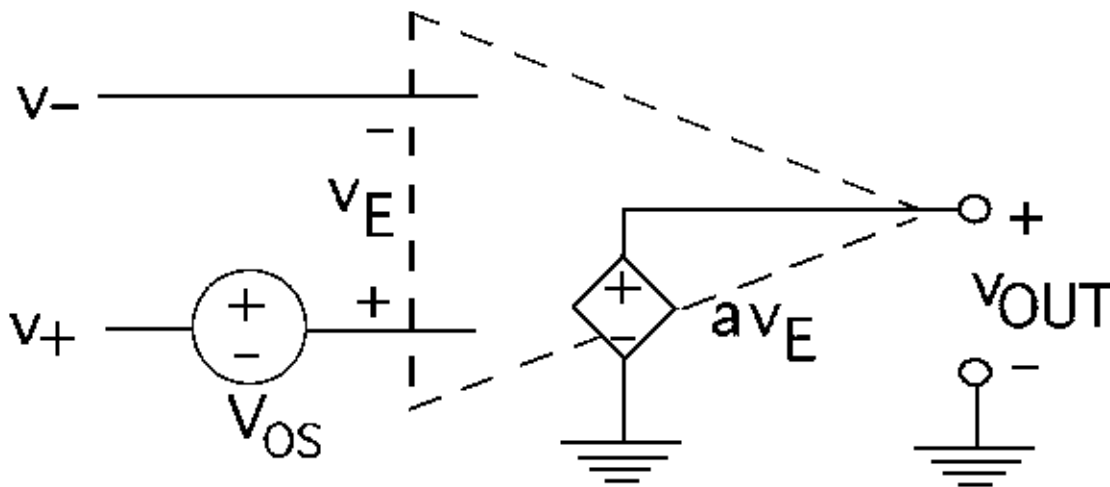
that ground be within this range since  $(v_+ + v_-)/2 \approx 0$  in this circuit. In Fig. 110.3,  $v_{IN}$  must stay in the common-mode input range since  $(v_+ + v_-)/2 \approx v_{IN}$ .

The **input offset voltage**  $V_{OS}$  is the DC op-amp input voltage that causes  $V_{OUT}$  to equal 0 V DC. The offset voltage varies from part to part; it is dependent on imbalances within the op amp. The input offset voltage can be included in the op-amp model, as shown in Fig. 110.7. When this model is used in Fig. 110.2, for example, the output voltage depends on the input voltage and the offset voltage:

$$v_{OUT} = -\frac{R_2}{R_1}v_{IN} + \left(1 + \frac{R_2}{R_1}\right)V_{OS} \quad (110.15)$$

For high gain ( $R_2 \gg R_1$ ), the amplification of  $V_{OS}$  can cause problems, in which case a capacitor can be added in series with  $R_1$ . This eliminates the amplification of  $V_{OS}$ , but it also causes the DC gain from  $v_{IN}$  to  $v_{OUT}$  to be zero.

**Figure 110.7** The op-amp model of Figure 110.1(c) with the addition of the input offset voltage  $V_{OS}$ .



These parameters and many more are specified on the data sheet and are incorporated into an op-amp macro model, which is an interconnection of circuit elements that model the linear and nonlinear behavior of the op amp. Most companies provide macro models for their op amps that can be used in a circuit simulation program such as **SPICE**.

## Defining Terms

**Ideal op-amp model:** Has  $Z_i = \infty$ ,  $Z_o = 0$ , and  $a = \infty$ , which leads to  $v_E = v_+ - v_- = 0$  (the virtual-short-circuit condition).

**Input offset voltage:** The op-amp DC input voltage,  $V_+ - V_-$ , that causes  $V_{OUT}$  to equal 0 V DC.

**Phase margin:** Is  $180^\circ - [\text{the phase of } RR(s = j\omega_U)]$ , where  $\omega_U$  is the frequency where the magnitude of  $RR$  is unity (i.e.,  $|RR(s = j\omega_U)| = 1$ ). Phase margin is measured in degrees.

**Gain margin:** Is  $-20 \log_{10} |\text{RR}(s = j\omega_{180})|$ , where  $\omega_{180}$  is the frequency where the phase of RR is  $-180^\circ$  (i.e., phase of  $\text{RR}(s = j\omega_{180}) = -180^\circ$ ). Gain margin is measured in decibels. Since RR is some factor times  $a(s)$  [e.g., see Eq. (110.11)], plots of the magnitude and phase of RR versus frequency can be generated from plots of  $a(s = j\omega)$ . The gain and phase margins can be found from magnitude and phase plots of  $\text{RR}(s = j\omega)$ .

**SPICE:** A computer program that can simulate circuits with linear and nonlinear elements. Many versions of SPICE are available.

## References

Analog Devices Inc., 1992. *Amplifier Reference Manual*. ADI, Norwood, MA.  
Frederiksen, T. M. 1984. *Intuitive IC Op Amps*. C.M.C., Milpitas, CA.  
Jung, W. G. 1986. *IC Op-Amp Cookbook*. Howard W. Sams, Indianapolis, IN.  
National Semiconductor Corp. 1993. *Operational Amplifiers Databook*. Santa Clara, CA.  
Rosenstark, S. 1986. *Feedback Amplifier Principles*. Macmillan, New York.

## Further Information

Two books that include many useful op-amp feedback circuits and valuable practical information are *IC Op-Amp Cookbook* by Walter G. Jung and *Intuitive IC Op Amps* by Thomas M. Frederiksen. Other good sources of practical information are the op-amp data and application books that are available from op-amp manufacturers. Stability and return ratio are covered in *Feedback Amplifier Principles* by Sol Rosenstark and in *Theory of Linear Active Networks* by E. S. Kuh and R. A. Rohrer.

Soderstrand, M. A. "Active RC Filters"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## 111.1 History of Active Filters

## 111.2 Active Filter Design Techniques

Cascaded Biquads • Cascade of Sallen and Key Filters • Simulation of LC Ladders • Coupled Biquads

## 111.3 Filter Specifications and Approximations

Bessel-Thompson Filters • Butterworth Filters • Chebyshev Filters • Inverse Chebyshev Filters •  
Elliptical-Cauer Filters

## 111.4 Filter Design

Low-Pass All-Pole Filter Design Example • Low-Pass Finite-Zero Filter Design Example

### Michael A. Soderstrand

*University of California, Davis*

**Active RC filters** arose out of the need for filters that are compatible with modern integrated circuit (IC) technology. Attempts to implement **passive RLC filters** in IC technology have largely failed due to the difficulties in implementing inductors using IC technology. Circuit theorists, however, have shown that adding an active device (e.g., an operational amplifier) to resistors and capacitors makes it possible to implement any filter function that can be implemented using passive components (i.e., passive R, L, C filters). Since operational amplifiers (op-amps), resistors, and capacitors are all compatible with modern IC technology, active RC filters are very attractive.

While the primary motivation for active RC filters comes from the desire to implement **filter circuits** in IC technology, there are additional advantages over passive RLC filters:

1. Op-amps, resistors, and capacitors perform much more closely to their ideal characteristics than inductors, which typically exhibit large parasitic resistances and significant nonlinearities.
2. Passive RLC filters are not able to take advantage of the component tracking and very accurate matching of component ratios available in IC technologies. (Note: Filter characteristics depend on ratios of element values rather than the absolute value of the components.)
3. Passive RLC filters cannot realize power gain, only attenuation.

In their infancy, active filters often suffered from a tendency to oscillate. Early work in active filters eliminated these problems but led to filters with high sensitivity to component values. Finally, in the 1970s, active filters emerged that clearly outperformed passive filters in cost, power consumption, ease of tuning, sensitivity, and flexibility. While IC technology allows for active filters that take full advantage of the new active RC filter design techniques, even active RC filters

designed with discrete components will perform as well as, if not better than, their passive counterparts.

## 111.1 History of Active Filters

---

The first practical circuits for active RC filters emerged during World War II and were documented much later in a classic paper by Sallen and Key [1955]. With the advent of IC technology in the early 1960s, active filter research flourished with the emphasis on obtaining stable filters [Mitra, 1971]. During the early 1970s, research focused on reducing the sensitivity of active filters to their component values [Schaumann *et al.*, 1976]. By the end of the 1970s, active filters had been developed that were better than passive filters in virtually every aspect [Bowron and Stephenson, 1979; Ghausi and Laker, 1981; Schaumann *et al.*, 1990; Tsividis and Voorman, 1993].

Unfortunately, most active filter handbooks and books that concentrate on the practical design of active filters have used the old Sallen and Key approach to active filter design. In this chapter we will briefly discuss op-amp-based active RC filter design techniques. (Note: In VLSI applications transconductance-based active filters offer an excellent alternative, which we will not discuss [Tsividis and Voorman, 1993].) Then we will introduce a table-based design procedure for one of the modern op-amp techniques that yields active RC filters as good as or better than passive RLC filters.

## 111.2 Active Filter Design Techniques

---

In this section we define three categories of active RC filter design using very different approaches. Within each category, we will briefly describe several of the specific methods available.

### Cascaded Biquads

The simplest and most popular active RC filter design techniques are based on a **cascade** of second-order (biquadratic or **biquad**) sections. The numerator and denominator of the filter function  $H(s)$  (expressed in terms of the complex frequency variable  $s$ ) are factored into second-order factors (plus one first-order, in the case of an odd-order filter). Second-order transfer functions are then formed by combining appropriate numerator terms with each second-order denominator term such that the product of the second-order transfer functions is equal to the original  $H(s)$ . (Note: In the case of odd-order filters, there will also be one first-order section.) Active RC biquad circuits are then used to implement each second-order section. In the case of odd-order filters, a separate first-order RC section may be added or the real pole can be combined with one of the second-order sections to form a third-order section. The different cascade methods are defined based on which biquad circuits are used and whether the ordering of the biquads is considered as part of the design procedure.

The advantages of the cascade design lie in the ease of design and the fact that the biquads are isolated and thus can be separately tuned during the manufacturing process. The disadvantage of the cascade design also derives from the isolated sections, as it allows only for feedback around

individual second-order stages without overall feedback around the entire filter structure. Thus the cascade of second-order sections is essentially operated open-loop.

## Cascade of Sallen and Key Filters

This approach is the basis for most active RC filter handbooks and many practical texts on active RC filter design. The resulting filters have the advantage of being canonical (i.e., having the fewest possible components—two resistors, two capacitors, and one op-amp per second-order stage). The performance of these filters is generally acceptable but is not as good as passive RLC filters. The primary disadvantage of this approach lies in the fact that it is not possible to use matched op-amps to compensate for the nonideal properties of the operational amplifier, since the sections only use one op-amp. A complete description of the design of a cascade of Sallen and Key filters can be found in the text by Arthur Williams [1975]. Williams also describes a slightly improved design using state-variable-based second-order sections. However, one should use optimum biquads rather than state-variable biquads since the optimum biquads have the same number or fewer components and outperform the state-variable filters.

### Cascade of Optimum Biquads

Optimum biquads require at least two active elements in order to make use of matched op-amps to reduce the effects of the nonideal properties of the active elements. Optimum two-op-amp biquad circuits are based on the gyrator circuits introduced by Antoniou [1967] and modified by Hamilton and Sedra [1971]. The design of these circuits will be covered later in this chapter.

### Ordered Cascade of Optimum Biquads

In order to obtain the best performance possible using the cascade technique, it is not sufficient to use optimum biquads. The second-order sections must also be ordered in an optimum fashion. In this chapter we will show how to design using the optimum biquads, but we will not go into the details of ordering the biquads. For those who need the ultimate in performance, an excellent discussion of ordering is provided in chapter 10 of Sedra and Brackett [1978].

## Simulation of LC Ladders

This approach was originally designed simply to take advantage of the vast reservoir of knowledge on passive **LC ladder filter circuit** design in the design of active filters. The simplest approach uses a gyrator [Antoniou, 1967] and a capacitor to simulate an inductor and replaces the inductors in the passive RLC ladder with this simulation (chapter 11 of Sedra and Brackett [1978]). More sophisticated techniques simulate the voltage and current equations of the passive RLC ladder (chapter 12 of Sedra and Brackett [1978]) or the scattering parameters [Haritantis *et al.*, 1976]. The latter techniques have resulted in the best filter circuits ever constructed.

At first it may seem that simulating passive RLC filters would at best yield filters equal to, but not better than, passive filters. However, the poor quality of passive inductors coupled with the ability of active RC filters to be constructed in IC technology, where very accurate ratios of components can be obtained, gives a huge advantage to active RC filters compared with passive

RLC filters. Even for filters realized with discrete components (i.e., not using IC technology), the advantage of not using an inductor and the use of dual matched op-amps yields circuits superior to those available with passive RLC filters.

### Component Simulation of LC Ladders

The primary attraction of this technique is the ease of design. You simply replace the inductors in a passive RLC filter obtained from any passive RLC design handbook [Hansell, 1969; Zverev, 1967] with active RC simulated inductors. Unfortunately, the floating inductors (inductors which do not have one terminal grounded) required for most passive RLC filters are one of the poorer-performing active RC components. However, modifications of this approach using *frequency-dependent negative resistors* (FDNRs) proposed by Bruton *et al.* [1972] have largely eliminated these problems. For details on the component simulation of LC ladders, see chapter 11 of Sedra and Brackett [1978].

### Operational Simulation of LC Ladders

In this approach active RC components are used as an *analog computer* to simulate the differential equations of Kirchhoff's current and voltage laws at each node in a passive RLC circuit. This approach has resulted in some of the best filter circuits ever designed. For details, see chapter 12 of Sedra and Brackett [1978] (see also [Ghausi and Laker, 1981; Schaumann *et al.*, 1990]).

### Wave Analog Filters

This approach is similar in concept to the operational simulation of RLC ladders except that scattering matrices of the passive RLC filter are simulated rather than Kirchhoff's current and voltage laws. This approach has become very popular in Europe, yielding filters every bit as good as those designed using the operational simulation of RLC ladders. For details see either Schaumann *et al.* [1990] or Haritantis *et al.* [1976].

## Coupled Biquads

The *coupled biquad* approach uses the same optimum biquads that we will use in the cascade approach later in this chapter, but provides feedback around the biquad sections. The advantage of this approach is that this additional feedback makes it possible to obtain filter performance equivalent to the filters realized by simulating RLC ladders, but using the modular, easily tuned biquad sections. However, the coupling of the biquads makes it much more difficult to tune the structure than in the case of cascaded biquads. For an excellent discussion of the coupled biquad design approach, see chapter 5 of Ghausi and Laker [1981] (see also chapter 10 of Sedra and Brackett [1978] and chapter 5 of Schaumann *et al.* [1990]).

## 111.3 Filter Specifications and Approximations

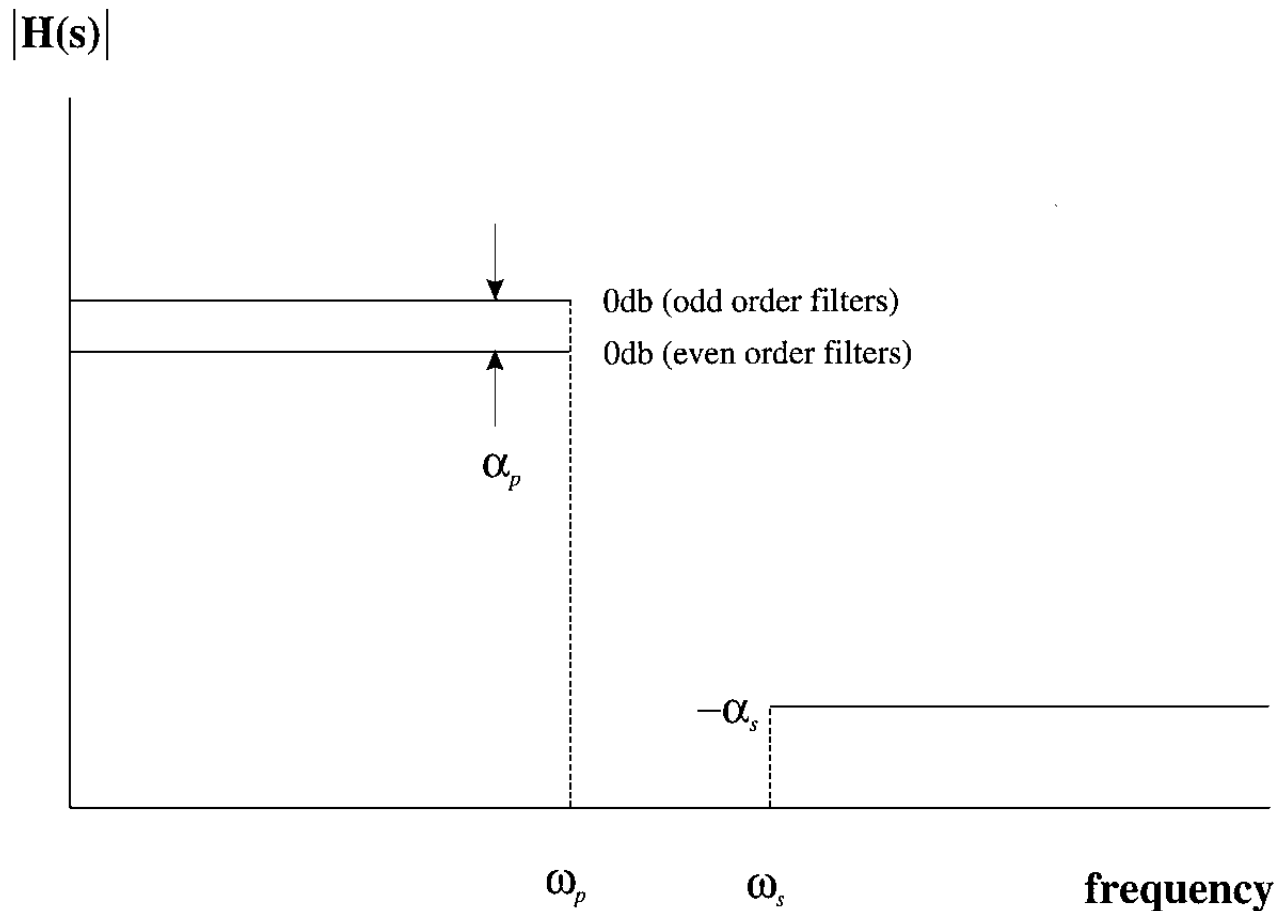
---

Low-pass filters are specified by four parameters: (1)  $\alpha_p$ , the **passband** ripple; (2)  $\omega_p$ , the passband frequency; (3)  $\alpha_s$ , the **stopband** attenuation; and (4)  $\omega_s$ , the stopband frequency. Figure 111.1 illustrates these parameters. A filter meeting these specifications must have  $|H(s)|$  lie between 0



dB and  $\alpha_p$  from DC to frequency  $\omega_p$ , and below  $\alpha_s$  for frequencies above  $\omega_s$ . For passive filters and odd-order active filters,  $\alpha_p$  is always negative. For active even-order filters,  $\alpha_p$  is positive and these filters exhibit power gain. Note that the gain for all active filter designs is normalized to 0 dB at DC (normalization is different for even-order passive filters). In practice, active RC filters are capable of providing an overall gain  $K$ , which multiplies the system function  $H(s)$  with the effect of scaling the vertical axis in Fig. 111.1.

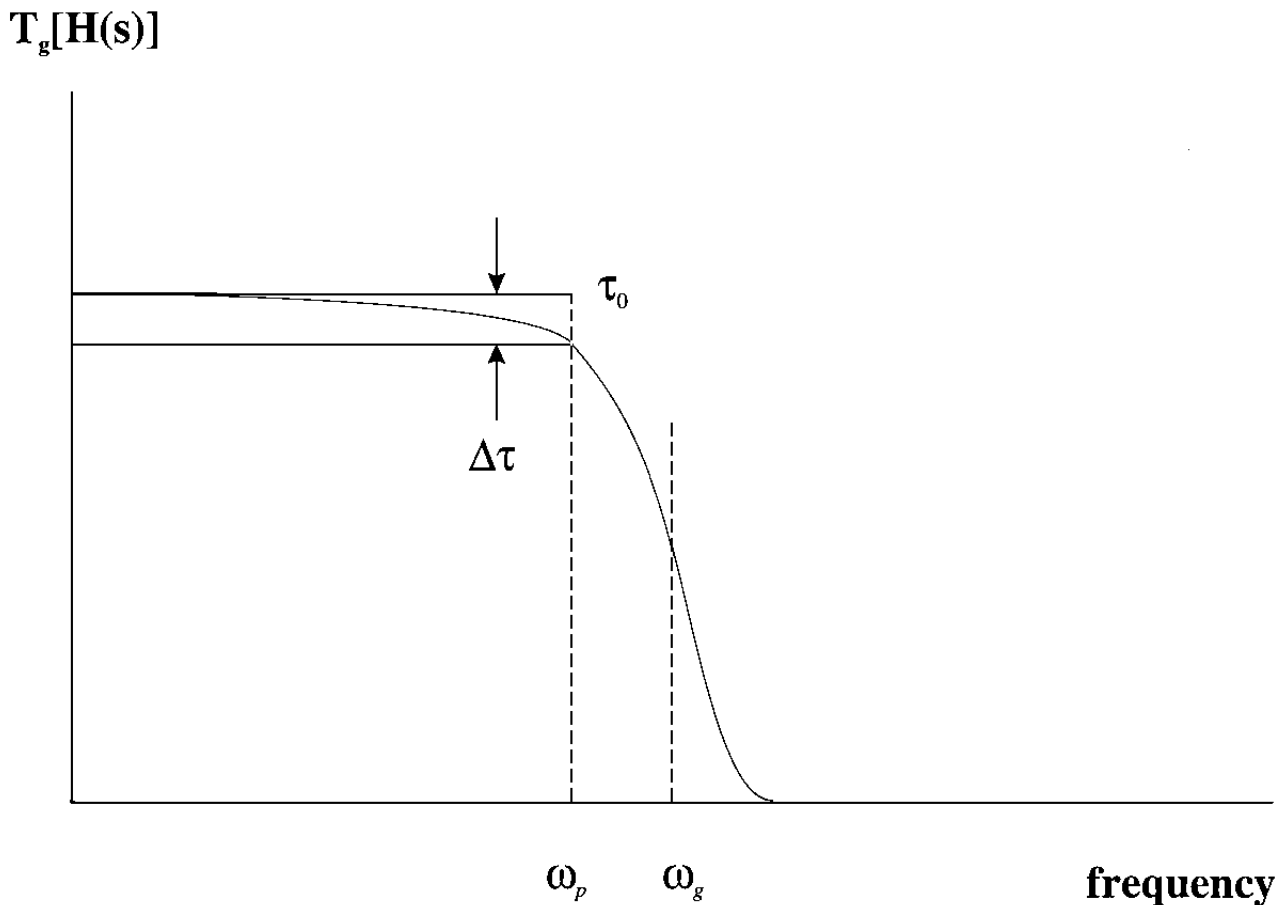
**Figure 111.1** Magnitude specifications for a low-pass filter.



The specifications of Fig. 111.1 are only half the story, however. In addition to the magnitude characteristics, filters have group delay characteristics related to the phase response of the filter. Figure 111.2 shows a typical group delay specification where (1)  $\tau_0$  is the nominal group delay, (2)  $\Delta\tau$  is the passband group delay tolerance, and (3)  $\omega_g$  is the frequency at which the group delay has decayed to 50% of  $\tau_0$ . Unfortunately, magnitude characteristics cannot be specified independent of group delay characteristics. Thus designers must make a choice between which set of specifications they wish to implement and must settle for whatever they get with the other specifications. Filter approximations are specific techniques for achieving one or the other set of

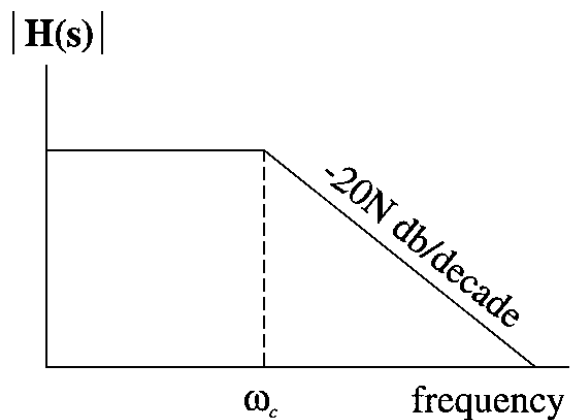
specifications. Each approximation offers a different compromise between meeting the magnitude specification and meeting the group delay specification.

**Figure 111.2** Delay specifications for a low-pass filter.

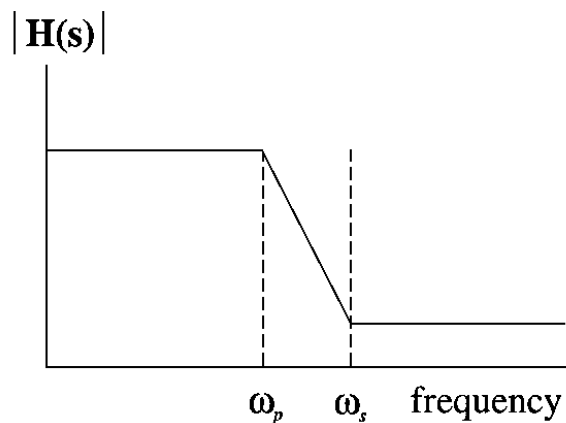


Filters can be classified into two broad classes: (1) all-pole filters and (2) filters with finite zeros. [Figure 111.3](#) shows the basic asymptotic response for all-pole filters in the left column and the basic asymptotic response for filters with finite zeros in the right column. In each column the response of the three basic filters, low-pass, high-pass, and band-pass, are shown. A fourth filter type, called a notch filter or band-elimination filter, is not shown because this type of filter exhibits an asymptotic response that is 0 dB at all frequencies, with a notch (area of high attenuation) around the corner frequency  $\omega_c$ . Most filter design procedures concentrate on designing low-pass filters and use standard transformations to convert these low-pass filters to high-pass, band-pass, or notch filters (a good explanation of how to do this can be found in chapter 6 of Sedra and Brackett [1978]).

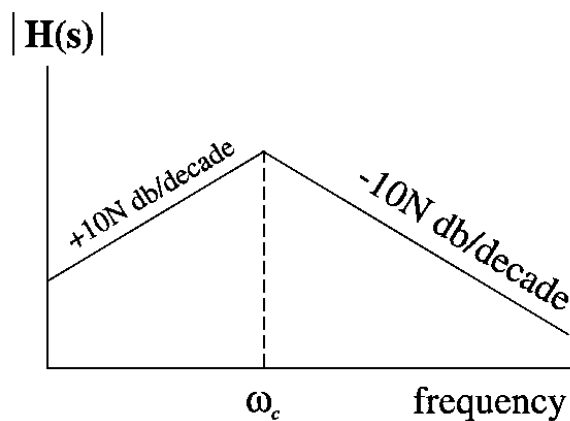
**Figure 111.3** Asymptotic response of filters.



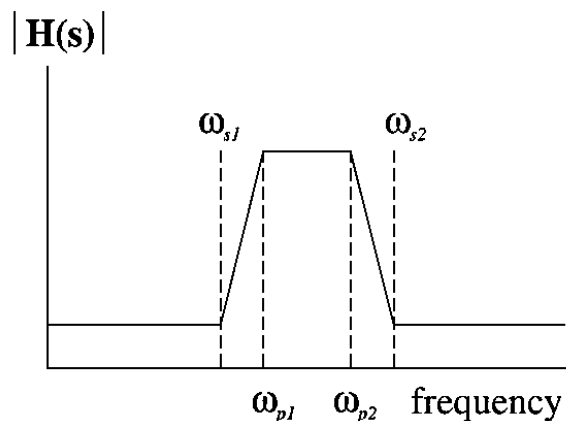
**Low-Pass All-Pole Filter**



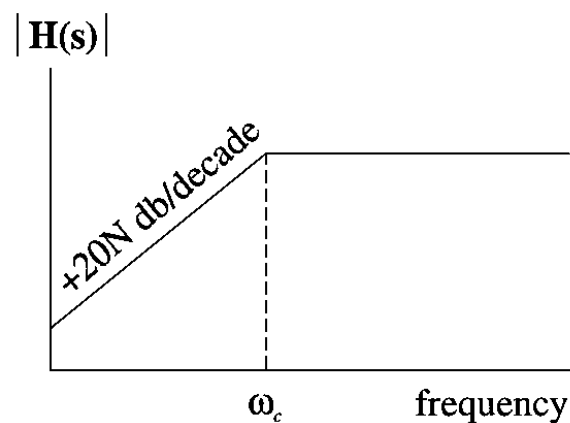
**Low-Pass Finite Zero Filter**



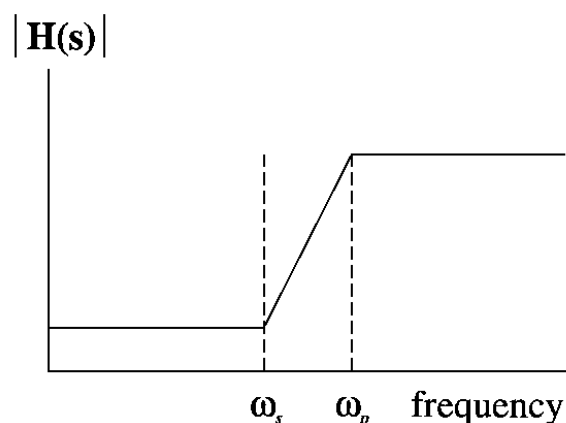
**Band-Pass All-Pole Filter**



**Band-Pass Finite Zero Filter**



**High-Pass All-Pole Filter**



**High-Pass Finite Zero Filter**

All passive RLC and active RC filters, regardless of the filter approximation or the method of implementation, have the same asymptotes for the same-order filter (*order* refers to the number of effective reactive elements—inductors and capacitors in a passive RLC filter, and capacitors in an active RC filter). The various filter approximations that we are about to introduce primarily affect how the filter performs in the **transition band** between frequencies  $\omega_p$  and  $\omega_s$ . In general, the wider the transition band, the better the group delay performance, but at the expense of significant deviation from the asymptotic response in the transition band. Note that the corner frequency  $\omega_c$  is an important parameter for all-pole filters because the asymptotic response is defined by this corner frequency along with the order of the filter  $N$  and the filter gain  $K$ .

## Bessel-Thompson Filters

The focus of the **Bessel-Thompson filters** is on meeting the group delay specification. The group delay is fixed at  $\tau_0$  for DC, and as many derivatives of the group delay function as possible are set to zero at DC. This yields what is called a *maximally flat* delay response. Mathematically it can be shown that this will obtain the closest possible match to the ideal "brick wall" delay response without any delay in excess of the prescribed group delay  $\tau_0$ . These filters were originally developed by Thompson and make use of Bessel functions—hence the name Bessel-Thompson filters. Bessel-Thompson filters exhibit the low-pass all-pole asymptotic response of [Fig. 111.3](#).

Although Bessel-Thompson filters offer the best possible group delay response, they do this at the expense of very poor magnitude response in the transition band. Furthermore, for a given order of filter, the transition band is much larger than for the other filter approximations. Other group delay approximations such as the equal-ripple group delay approximation provide only slightly better magnitude performance.

It is often very useful to calculate the required order of a Bessel-Thompson filter from the specifications for that filter. Since Bessel-Thompson filters are primarily dependent on the group delay specifications, the exact formula does not involve the standard magnitude specifications. However, the following approximate formula estimates the order of a Bessel-Thompson filter:

$$N \approx \frac{\alpha_s}{20 \log_{10}(\omega_s/\omega_c)} \quad (111.1)$$

with  $\alpha_s$  in dB and  $\omega_c$  and  $\omega_s$  in radians (or hertz).

## Butterworth Filters

**Butterworth filters** provide a maximally flat passband response for the magnitude of the transfer function at the expense of some peaking in the group delay response around the passband frequency  $\omega_p$ . In most cases the large improvement in magnitude response compared with Bessel-Thompson filters more than compensates for the peaking of the group delay response. Butterworth filters exhibit the low-pass all-pole asymptotic response of [Fig. 111.3](#) with exactly 3 dB of attenuation at the corner frequency  $\omega_c$ . This is less droop (i.e., attenuation at the corner frequency) than Bessel-Thompson filters, but more than the other filter approximations we will

discuss.

It is often very useful to calculate the required order of a Butterworth filter from the specifications. The following formula gives this relationship:

$$N \geq \frac{\log_{10}(K_\alpha)}{\log_{10}(K_\omega)} \quad (111.2)$$

where

$$K_\alpha = \sqrt{\frac{10^{\alpha_s/10} - 1}{10^{\alpha_p/10} - 1}} \quad (111.3)$$

$$K_\omega = \omega_s / \omega_p \quad (111.4)$$

with  $\alpha_p$  and  $\alpha_s$  in dB and  $\omega_p$  and  $\omega_s$  in radians (or hertz).

## Chebyshev Filters

**Chebyshev filters** provide an equal-ripple response in the passband (i.e., the magnitude of  $H(s)$  oscillates or ripples between zero and  $\alpha_p$  in the passband). Chebyshev filters are all-pole filters that can be shown to provide the narrowest transition band for a given filter order  $N$  of any of the all-pole filters. This narrow transition band, however, comes at the expense of significant and often unacceptable peaking and oscillation in the group delay response.

It is often very useful to calculate the required order of a Chebyshev filter from the specifications for that filter. The following formula gives this relationship:

$$N \geq \frac{\cosh^{-1}(K_\alpha)}{\cosh^{-1}(K_\omega)} \quad (111.5)$$

or

$$N \geq \frac{\log_{10} \left( K_\alpha + \sqrt{K_\alpha^2 - 1} \right)}{\log_{10} \left( K_\omega + \sqrt{K_\omega^2 - 1} \right)} \quad (111.6)$$

where  $K_\alpha$  and  $K_\omega$  are as defined in Eqs. (111.3) and (111.4), respectively.

## Inverse Chebyshev Filters

**Inverse Chebyshev filters** provide the same maximally flat magnitude response of Butterworth filters, including the slight, but usually acceptable, peaking in the group delay response, with a transition band exactly the same as for the Chebyshev filters. However, the inverse Chebyshev

filters exhibit the low-pass finite-zero asymptotic response of Fig. 111.3 with equal ripple in the stopband.

It is often very useful to calculate the required order of an inverse Chebyshev filter from the specifications. Since the inverse Chebyshev filter has the same order as the Chebyshev filter, Eqs. (111.5) or (111.6) can also be used to calculate the order of inverse Chebyshev filters.

## Elliptical-Cauer Filters

The **elliptical-Cauer filters** exhibit the low-pass finite-zero asymptotic response of Fig. 111.3 with equal ripple in both the passband (like Chebyshev filters) and the stopband (like inverse Chebyshev filters). While the group delay of these filters is poor, like that of the Chebyshev filters, elliptical-Cauer filters can be shown to have the narrowest transition band of any of the filter approximations.

It is often very useful to calculate the required order of an elliptical-Cauer filter. The following formula gives this relationship:

$$N \geq \frac{CEI(1/K_\omega)CEI\left(\sqrt{1-1/K_\alpha^2}\right)}{CEI(1/K_\alpha)CEI\left(\sqrt{1-1/K_\omega^2}\right)} \quad (111.7)$$

where  $CEI$  is the *complete elliptic integral* function.■

The  $CEI$  function is denoted by  $K$  or  $q$  in most textbooks, but we have chosen  $CEI$  so as not to be confused with other filter parameters.

Although the  $CEI$  function is not typically provided on calculators or in computer mathematics libraries, computer programs are readily available for calculating the  $CEI$  function.■

For example, MATLAB has the algorithm *ellipord* for calculating the order of both digital and analog elliptic filters. Also, Daniels [1974] contains a computer algorithm for calculating  $CEI$  (p. 79), and Lindquist [1977] contains an approximate formula for  $N$ .

$K_\alpha$  and  $K_\omega$  are as defined in Eqs. (111.3) and (111.4), respectively.

## 111.4 Filter Design

---

Figure 111.4 represents a biquad circuit that can be used for low-pass all-pole filters, and Fig. 111.5 shows a biquad that can be used for low-pass finite-zero filters. Both biquads exhibit optimum performance with regard to both passive and active sensitivities [Sedra and Brackett, 1978]. The element values for the circuit of Fig. 111.4 are given by

$$R = Q \quad C = 1/\omega_0 \quad (111.8)$$

where  $\omega_0$  and  $Q$  are the normalized **pole frequency** and **Q** of the second-order filter section. Practical element values are obtained by impedance scaling to practical impedance levels (i.e., multiply each resistor by the desired impedance level and divide each capacitor by the desired impedance level) and frequency scaling to the desired  $\omega_p$  (i.e., divide each capacitor by the desired

**Figure 111.5** Biquad circuit for finite-zero low-pass filters.

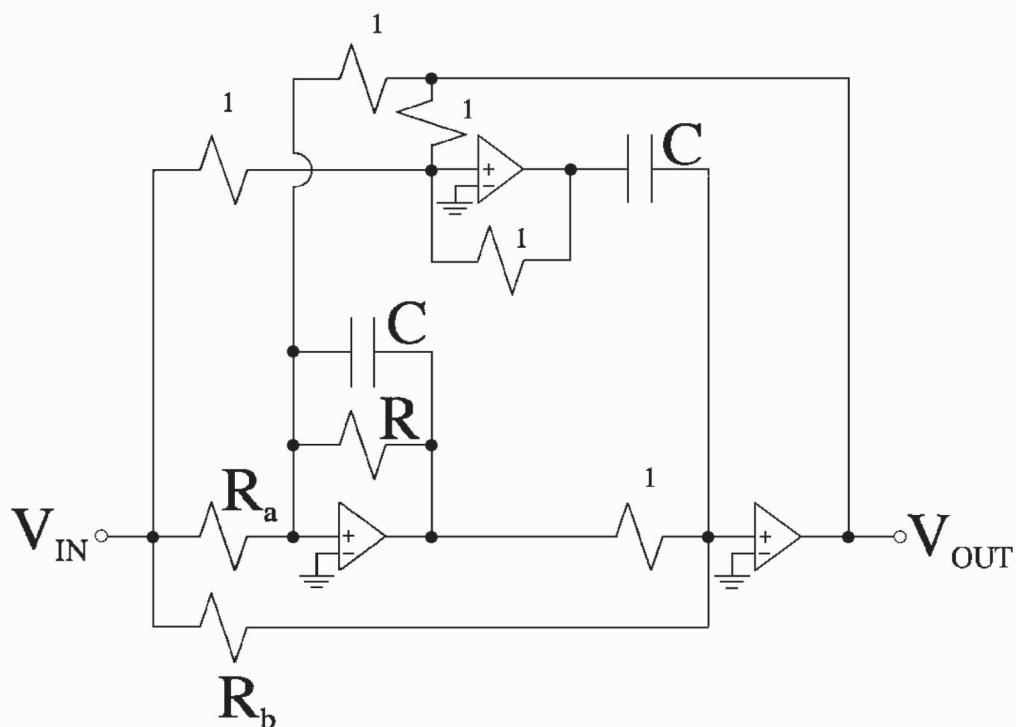


Figure 111.5 is used for the filters with finite zeros. The element values for the circuit of Fig. 111.5 are given by

$$\begin{aligned} R_a &= \frac{Q^2 \omega_n^2}{Q^2 \omega_n^2 + \omega_0^2} & R_b &= \frac{Q \omega_n^2}{\omega_0^2} & R_c &= \frac{\omega_n^2}{\omega_0^2} \\ R &= Q & C &= 1/\omega_0 \end{aligned} \quad (111.9)$$

where  $\omega_0$  and  $Q$  are the normalized resonant frequency and  $Q$  for the second-order filter section and  $\omega_n$  is the normalized frequency of the finite zero. As for the all-pole filter section, our tables have been normalized to  $\omega_p = 1$ . Practical resistor and capacitor values are obtained by impedance scaling to the desired impedance level and frequency scaling to the desired  $\omega_p$ .

In order to determine what the resonant frequencies,  $Q$ 's, and zero frequencies are for various types of filters, one may either use a computer program or a table. Many commercial computer programs are available for this purpose and should be used if a significant number of filters are going to be designed. For the casual filter designer, public domain software is available. The programs *filter* [Wilamowski, 1994] and *ladder* [Koller and Wilamowski, 1994] can be obtained by anonymous FTP to PLAINS.UWYO.EDU in directory "electrical." Copy files FILTERxx.ZIP and LADDERyy.ZIP (where xx and yy are the version numbers; currently xx = 35 and yy = 31) in *binary* mode, as these are in "zip" format. (Note: You must use an "unzip" program to expand them.) or the tables included with this chapter may be used. In what follows, we will make use of the tables to do a design example.

## Low-Pass All-Pole Filter Design Example

Table 111.1 applies to the circuit of Fig. 111.4 and is used for the design of the all-pole filters: Bessel-Thompson, Butterworth, and Chebyshev. Consider the design of an all-pole filter with the specifications

$$\alpha_p = 1 \text{ dB}, \quad \omega_p = 1 \text{ kHz}, \quad \alpha_s = 40 \text{ dB}, \quad \text{and} \quad \omega_s = 3 \text{ kHz}$$

Plugging these values into Eqs. (111.3) and (111.4), we obtain  $K_\alpha = 197$  and  $K_\omega = 3$ . Using Eqs. (111.2) and (111.5), we see that the order of filter necessary to meet these specifications is  $N_B = 5$  for a Butterworth filter and  $N_C = 4$  for a Chebyshev filter. Note that these specifications are too restrictive to be met by a Bessel-Thompson filter. (From Table 111.1 we see that  $\omega_c \approx 3$  for a sixth-order Bessel-Thompson filter and  $\omega_c > 3$  for higher-order Bessel-Thompson filters. From Eq. (111.1) for the order of a Bessel-Thompson filter, we see that  $\omega_s = 3 > \omega_c$  to obtain a practical value for  $N$ .)



**Table 111.1** Design Table for Low-Pass All-Pole Filters

N	$\alpha_p$	S	Bessel-Thompson				Butterworth				Chebyshev						
			R	C	$\alpha_c$	$\omega_c$	R	C	$\alpha_c$	$\omega_c$	R	C	$\alpha_c$	$\omega_c$			
2	0.10	1	0.5774	0.1511	-4.77	6.620	0.7071	0.3914	-3.00	2.555	0.7674	0.5505	-2.30	1.817			
	0.25	1		0.2367		4.225		0.4948		2.021	0.8093	0.6878	-1.84	1.454			
	0.50	1		0.3323		3.009		0.5910		1.692	0.8637	0.8121	-1.27	1.231			
	1.00	1		0.4619		2.165		0.7133		1.402	0.9565	0.9524	-0.39	1.050			
	2.00	1		0.6432		1.555		0.8745		1.143	1.1286	1.1023	+1.05	0.907			
	3.00	1		0.7842		1.275		1.0000		1.000	1.3047	1.1885	+2.31	0.841			
3	0.10	0 1	1.0000 0.6910	0.1459 0.1333	-6.25	7.295	1.0000 1.0000	0.5344 0.5344	-3.00	1.871 1.3409	1.0000 0.7693	1.0316 0.6963	-0.83 +2.81	1.178			
	0.25	0 1	1.0000 0.6910	0.2299 0.2098		4.634	1.0000 1.0000	0.6243 0.6243		1.602	1.0000 1.5080	1.3034 0.8643	-0.32 +0.31	1.016			
	0.50	0 1	1.0000 0.6910	0.3229 0.2950		3.293	1.0000 1.0000	0.7045 0.7045		1.419	1.0000 1.7062	1.5963 0.9356	-0.03 +0.35	0.902			
	1.00	0 1	1.0000 0.6910	0.4545 0.4125		2.352	1.0000 1.0000	0.7986 0.7986		1.252	1.0000 2.0177	2.0236 1.0029	-0.18 +0.07	0.788			
	2.00	0 1	1.0000 0.6910	0.6264 0.5723		1.696	1.0000 1.0000	0.9148 0.9148		1.093	1.0000 2.5516	2.7107 1.0623	-1.25 +0.06	0.690			
	3.00	0 1	1.0000 0.6910	0.7578 0.6906		1.409	1.0000 1.0000	1.0000 1.0000		1.000	1.0000 3.0677	3.3487 1.0916	-2.51 +0.06	0.632			
	4	0.10	1 2	0.5219 0.8055		0.1740 0.2625	-8.13	4.666		0.5412 1.3066	0.6250 0.6250	-3.00	1.600	0.6188 2.1829	1.2670 0.8671	+0.09	0.956
		0.25	1 2	0.5219 0.8055		0.2512 0.3790		3.232		0.5412 1.3066	0.7024 0.7024		1.424	0.6572 2.5361	1.4828 0.9277	+0.16	0.858
		0.50	1 2	0.5219 0.8055		0.3222 0.4861		2.520		0.5412 1.3066	0.7688 0.7688		1.301	0.7051 2.9406	1.6750 0.9697	+0.09	0.788
		1.00	1 2	0.5219 0.8055		0.4047 0.6106		2.006		0.5412 1.3066	0.8446 0.8446		1.184	0.7845 3.5590	1.8919 1.0068	+0.01	0.727
2.00		1 2	0.5219 0.8055	0.5025 0.7581	1.616	0.5412 1.3066		0.9352 0.9352	1.069	0.9294 4.5939	2.1244 1.0377		+0.04	0.675			
3.00		1 2	0.5219 0.8055	0.5706 0.8608	1.513	0.5412 1.3066		1.0000 1.0000	1.000	1.0765 5.5789	2.2589 1.0523		+0.19	0.649			
5		0.10	0 1 2	1.0000 0.5635 0.9165	0.1248 0.1205 0.1068	-8.80		8.682	1.0000 0.6180 1.6180	0.6866 0.6866 0.6866	-3.00		1.457	1.0000 0.9145 3.2820	1.8556 1.2540 0.9148	-0.09	0.841
		0.25	0 1 2	1.0000 0.5635 0.9165	0.1970 0.1902 0.1686			5.495	1.0000 0.6180 1.6180	0.7538 0.7538 0.7538			1.327	1.0000 1.0359 3.8757	2.2886 1.3654 0.9554	-0.22	0.765
		0.50	0 1 2	1.0000 0.5635 0.9165	0.2775 0.2678 0.2375			3.887	1.0000 0.6180 1.6180	0.8103 0.8103 0.8103			1.234	1.0000 1.1778 4.5450	2.7600 1.4483 0.9826	-0.24	0.716
		1.00	0 1 2	1.0000 0.5635 0.9165	0.3914 0.3778 0.3349			2.755	1.0000 0.6180 1.6180	0.8736 0.8736 0.8736			1.145	1.0000 1.3988 5.5564	3.4543 1.5262 1.0059	-0.25	0.662
	2.00	0 1 2	1.0000 0.5635 0.9165	0.5482 0.5292 0.4692	1.945		1.0000 0.6180 1.6180	0.9478 0.9478 0.9478	1.055	1.0000 1.7751 7.2323		4.5807 1.5949 1.0248	-0.07	0.611			
	3.00	0 1 2	1.0000 0.5635 0.9165	0.6645 0.6415 0.5687	1.614		1.0000 0.6180 1.6180	1.0000 1.0000 1.0000	1.000	1.0000 2.1375 8.8178		5.6329 1.6286 1.0336	-0.02	0.578			

N	$\alpha_p$	S	Bessel-Thompson				Butterworth				Chebyshev			
			R	C	$\alpha_c$	$\omega_c$	R	C	$\alpha_c$	$\omega_c$	R	C	$\alpha_c$	$\omega_c$
6	-0.10	1	0.5103	0.1159	-10.1	9.289	0.5176	0.7310	-3.00	1.368	0.5995	1.9486	+0.06	0.775
		2	0.6112	0.1101			0.7071	0.7310			1.3316	1.1983		
		3	1.0233	0.0976			1.9319	0.7310			4.6329	0.9410		
	-0.25	1	0.5103	0.1831		5.875	0.5176	0.7902	1.266	0.6370	2.2519	+0.25	0.713	
		2	0.6112	0.1738			0.7071	0.7902		1.5557	1.2597			
		3	1.0233	0.1542			1.9319	0.7902		5.5204	0.9698			
	-0.50	1	0.5103	0.2583		4.150	0.5176	0.8392	1.192	0.6836	2.5238	+0.47	0.679	
		2	0.6112	0.2453			0.7071	0.8392		1.8104	1.3019			
		3	1.0233	0.2175			1.9319	0.8392		6.5128	0.9887			
	-1.00	1	0.5103	0.3647		2.938	0.5176	0.8935	1.119	0.7609	2.8317	+0.75	0.645	
		2	0.6112	0.3463			0.7071	0.8935		2.1980	1.3390			
		3	1.0233	0.3071			1.9319	0.8935		8.0037	1.0047			
	-2.00	1	0.5103	0.5110		2.104	0.5176	0.9563	1.046	0.9016	3.1634	+0.99	0.614	
		2	0.6112	0.4852			0.7071	0.9563		2.8443	1.3698			
		3	1.0233	0.4303			1.9319	0.9563		10.4616	1.0175			
	-3.00	1	0.5103	0.6232		1.725	0.5176	1.0000	1.000	1.0443	3.3557	+1.05	0.600	
		2	0.6112	0.5917			0.7071	1.0000		3.4581	1.3843			
		3	1.0233	0.5247			1.9319	1.0000		12.7801	1.0234			
7	-0.10	0	1.0000	0.1100	-11.3	9.875	1.0000	0.7645	-3.00	1.308	1.0000	2.6541	-0.27	0.726
		1	0.5324	0.1079			0.5550	0.7645			0.8464	1.7402		
		2	0.6608	0.1017			0.8019	0.7645			1.8472	1.1522		
		3	1.1263	0.0904			2.2470	0.7645		6.2332	0.9568			
	-0.25	0	1.0000	0.1738		6.254	1.0000	0.8172	1.224	1.0000	3.2510	-0.19	0.679	
		1	0.5324	0.1705			0.5550	0.8172		0.9596	1.8802			
		2	0.6608	0.1606			0.8019	0.8172		2.1904	1.1902			
		3	1.1263	0.1428			2.2470	0.8172		7.4678	0.9782			
	-0.50	0	1.0000	0.2454		4.428	1.0000	0.8605	1.162	1.0000	3.9037	-0.48	0.646	
		1	0.5324	0.2409			0.5550	0.8605		1.0916	1.9847			
		2	0.6608	0.2268			0.8019	0.8605		2.5755	1.2155			
		3	1.1263	0.2017			2.2470	0.8605		8.8418	0.9920			
	-1.00	0	1.0000	0.3461		3.140	1.0000	0.9080	1.101	1.0000	4.8682	-0.49	0.612	
		1	0.5324	0.3397			0.5550	0.9080		1.2969	2.0831			
		2	0.6608	0.3199			0.8019	0.9080		3.1559	1.2371			
		3	1.1263	0.2845			2.2470	0.9080		10.8967	1.0037			
	-2.00	0	1.0000	0.4868		2.232	1.0000	0.9624	1.039	1.0000	6.4375	-1.76	0.578	
		1	0.5324	0.4778			0.5550	0.9624		1.6464	2.1699			
2		0.6608	0.4500	0.8019	0.9624		4.1151	1.2545						
	3	1.1263	0.4001		2.2470	0.9624		14.2801	1.0129					
-3.00	0	1.0000	0.5928	1.834	1.0000	1.0000	1.000	1.0000	7.9061	-2.34	0.557			
	1	0.5324	0.5817		0.5550	1.0000		1.9829	2.2127					
	2	0.6608	0.5479		0.8019	1.0000		5.0214	1.2626					
	3	1.1263	0.4872		2.2470	1.0000		17.4645	1.0172					

From Table 111.1 we obtain the following values for the Butterworth fifth-order and Chebyshev fourth-order filters that meet the specifications

*Butterworth fifth-order* ( $\alpha_p = 1.00$  dB):

Stage 0	$R_0 = 1.0000$	$C_0 = 0.8736$
Stage 1	$R_1 = 0.6180$	$C_1 = 0.8736$
Stage 2	$R_2 = 1.6180$	$C_2 = 0.8736$

*Chebyshev fourth-order* ( $\alpha_p = 1.00$  dB):

Stage 1	$R_1 = 0.7845$	$C_1 = 1.8919$
Stage 2	$R_2 = 3.5590$	$C_2 = 1.0068$

To obtain practical values, we must denormalize by impedance scaling to a practical impedance level (we have chosen 10 k $\Omega$ ) and frequency scaling to  $\omega_p = 1$  kHz. Thus we multiply all resistors by  $10^4$  and divide all capacitors by  $10^4$  to impedance-scale, and we then divide all capacitors by  $2\pi \times 10^3$  to frequency-scale. For active filter implementation, we recommend *metal film* resistors and *polystyrene* capacitors. *Mica* capacitors are slightly superior to polystyrene capacitors but have a more restricted range of values. *Polycarbonate* capacitors have a larger range of values but a poorer "retrace" property. *Mylar* capacitors provide large capacitance values in a small package, but at the expense of a much poorer temperature coefficient. The practical values for the resistors and the capacitors of our example after scaling are as follows:

*Butterworth fifth-order* ( $\alpha_p = 1.00$  dB):

Stage 0	$R_0 = 10.0$ k- $\Omega$	$C_0 = 13.9$ nF
Stage 1	$R_1 = 6.18$ k- $\Omega$	$C_1 = 13.9$ nF
Stage 2	$R_2 = 16.2$ k- $\Omega$	$C_2 = 13.9$ nF

*Chebyshev fourth-order* ( $\alpha_p = 1.00$  dB):

Stage 1	$R_1 = 7.85$ k- $\Omega$	$C_1 = 30.1$ nF
Stage 2	$R_2 = 35.6$ k- $\Omega$	$C_2 = 16.0$ nF

Note: All resistors labeled 1 - in Fig. 111.4 become 10 k $\Omega$ , and the two resistors labeled 2 - become 20 k $\Omega$ . These values are well suited for metal film resistors and polystyrene capacitors.

## Low-Pass Finite-Zero Filter Design Example

Table 111.2 applies to the circuit of Fig. 111.5 and is used for the design of the finite-zero filters: inverse Chebyshev and elliptical-Cauer. Consider the design of a finite-zero filter with the same specifications as in the first example:

$$\alpha_p = 1 \text{ dB}, \quad \omega_p = 1 \text{ kHz}, \quad \alpha_s = 40 \text{ dB}, \quad \text{and} \quad \omega_s = 3 \text{ kHz}$$

Plugging these values into Eqs. (111.3) and (111.4), we obtain  $K_\alpha = 197$  and  $K_\omega = 3$ . Using Eqs. (111.6) and (111.7), respectively, we see that the order of filter necessary to meet these specifications is  $N_{IC} = 4$  for an inverse Chebyshev filter and  $N_{EC} = 3$  for an elliptical-Cauer filter. From Table 111.2 we obtain the following values for the inverse Chebyshev fourth-order and elliptical-Cauer third-order filters that meet the specifications:

*Inverse Chebyshev fourth-order* ( $\alpha_s = 40 \text{ dB}$  and  $\alpha_p = 1.00 \text{ dB}$ ):

$$\text{Stage 1} \quad Q_1 = 0.5540 \quad \omega_1 = 1.3074 \quad \omega_{n1} = 2.5312$$

$$\text{Stage 2} \quad Q_2 = 1.4780 \quad \omega_2 = 1.1832 \quad \omega_{n2} = 6.1109$$

*Elliptical-Cauer third-order* ( $\alpha_s = 40 \text{ dB}$  and  $\alpha_p = 1.00 \text{ dB}$ ):

$$\text{Stage 0} \quad \omega_0 = 0.5237$$

$$\text{Stage 1} \quad Q_1 = 2.2060 \quad \omega_1 = 1.0027 \quad \omega_{n1} = 2.7584$$

**Table 111.2** Design Table for Low-Pass Finite-Zero Filters

Specifications		N	S	Inverse Chebyshev			Cauer (Elliptical)				
$\alpha_s$	$\alpha_p$			Q	$\omega_p$	$\omega_s$	Q	$\omega_p$	$\omega_s$	$\omega_s$	
40 dB	0.10 dB	2	1	0.7107	2.5616	25.6162	18.11336	0.7719	1.8218	18.1134	12.81790
		3	0		1.9437		5.51714		1.0100		3.51952
		3	1	1.0455	1.8591	6.3706		1.4269	1.2936	4.0429	
		4	1	0.5540	1.7282	3.3461	3.09139	0.6420	0.8643	2.0662	1.92976
		4	2	1.4780	1.5640	8.0782		2.6744	1.1362	4.7252	
		5	0		1.7409		2.22010		0.6706		1.41762
		5	1	0.6811	1.5870	2.3344		1.0852	0.8940	1.4691	
		5	2	2.0218	1.3997	3.7771		5.0104	1.0690	2.1727	
		6	1	0.5346	1.7437	1.8689	1.80515	0.6349	0.6933	1.2276	1.20454
		6	2	0.8653	1.4716	2.3529		1.9787	0.9366	1.5033	
		6	3	2.6828	1.2968	6.9748		9.3816	1.0360	3.6245	
		0.25 dB	2	1	0.7107	2.0293	20.2927	14.34906	0.8144	1.4557	14.3491
	3	0		1.6686		4.73634		0.8031		3.03512	
	3	1	1.0455	1.5960	5.4691		1.6166	1.1552	3.4800	3.50512	
	4	1	0.5540	1.5487	2.9986	2.77030	0.6847	0.7472	1.8693	1.75051	
	4	2	1.4750	1.4016	7.2392		3.1856	1.0700	4.2159		
	5	0		1.6087		2.04445		0.5554		1.33131	
	5	1	0.6811	1.4515	2.1497		1.2523	0.8351	1.3741		
	5	2	2.0218	1.2889	3.4782		6.2260	1.0344	2.0009		
	6	1	0.5346	1.6360	1.7534	1.69357	0.6783	0.6162	1.1779	1.15808	
	6	2	0.8653	1.3806	2.3951		2.4101	0.9079	1.4204		
	6	3	2.6828	1.2106	6.5436		12.1365	1.0174	1.3636		
	0.50 dB	2	1	0.7107	1.6949	16.9489	11.98464	0.8699	1.2335	11.9847	8.48923
	3	0		1.4839		4.21200		0.6991		2.71147	
	3	1	1.0455	1.4193	4.8636		1.8439	1.0706	3.1031		
	4	1	0.5540	1.4245	2.7579	2.54801	0.7366	0.6669	1.7568	1.62843	
	4	2	1.4780	1.2891	6.6583		3.7584	1.0303	3.9200		
	5	0		1.5115		1.82086		0.4700		1.27264	
	5	1	0.6811	1.3731	2.0197		1.4499	0.7995	1.3126		
	5	2	2.0218	1.2110	3.2680		7.6747	1.0144	1.8800		
	6	1	0.5346	1.5994	1.6713	1.61432	0.7318	0.5635	1.1443	1.12698	
	6	2	0.8653	1.3160	2.2830		2.9188	0.8925	1.3623		
	6	3	2.6828	1.1597	6.2374		15.4850	1.0071	3.1569		
	1.00 dB	2	1	0.7107	1.4054	14.0540	9.93763	0.9644	1.0526	9.9377	7.04485
	3	0		1.3143		3.73075		0.5237		2.41619	
	3	1	1.0455	1.2571	4.3079		2.2060	1.0027	2.7584		
	4	1	0.5540	1.3074	2.5312	2.33855	0.8255	0.6015	1.6096	1.51549	
	4	2	1.4780	1.1832	6.1109		4.7456	0.9993	3.5253		
	5	0		1.4186		1.80280		0.3654		1.21869	
	5	1	0.6811	1.2887	1.8956		1.7634	0.7727	1.2538		
	5	2	2.0218	1.1366	3.0671		10.0100	0.9994	1.7643		
	6	1	0.5346	1.4857	1.5923	1.53798	0.8200	0.5174	1.1138	1.09888	
	6	2	0.8653	1.2538	2.1751		3.7287	0.8831	1.3070		
	6	3	2.6828	1.1049	5.9425		21.0133	0.9997	2.9628		
	2.00 dB	2	1	0.7107	1.1478	11.4782	8.11633	1.1402	0.9107	8.1164	5.76106
	3	0		1.1540		3.27555		0.3953		2.13924	
	3	1	1.0455	1.1037	3.7823		2.8401	0.9513	2.4338		
	4	1	0.5540	1.1936	2.3109	2.13499	0.9861	0.5469	1.4902	1.40843	
	4	2	1.4780	1.0802	5.5790		6.4348	0.9771	3.1959		
	5	0		1.3270		1.68641		0.3008		1.16811	
	5	1	0.6811	1.2055	1.7732		2.3131	0.7567	1.1982		
	5	2	2.0218	1.0652	2.6691		14.2439	0.9895	1.6501		
	6	1	0.5346	1.4124	1.5137	1.46212	0.9809	0.4812	1.0856	1.07317	
	6	2	0.8653	1.1920	2.0678		5.1692	0.8808	1.2529		
	6	3	2.6828	1.0504	5.6493		31.3724	0.9952	2.7640		

Specifications		N	S	Inverse Chebyshev			Cauer (Elliptical)				
$\alpha_s$	$\alpha_p$			Q	$\omega_p$	$\omega_s$	Q	$\omega_p$	$\omega_s$	$\omega_s$	
60 dB	0.10 dB	2	1	0.7075	2.5599	80.9518	57.24160	0.7678	1.8206	57.2417	40.4791
		3	0		1.8865		11.80927		0.9780		7.45970
		3	1	1.0095	1.8687	13.8362		1.3588	1.2986	8.6040	
		4	1	0.5449	1.6396	5.8410	5.39637	0.6255	0.8121	3.5549	3.25974
		4	2	1.3577	1.5886	14.1015		2.3212	1.1480	8.4379	
		5	0		1.5664		3.40515		0.5883		2.04438
		5	1	0.6402	1.5100	3.5804		0.9746	0.8373	2.1363	
		5	2	1.7657	1.4333	5.7933		3.8628	1.0830	3.3302	
		6	1	0.5241	1.5299	2.6201	2.53074	0.6138	0.5928	1.5903	1.54869
		6	2	0.7705	1.4215	3.5791		1.5802	0.8868	2.0567	
		6	3	2.2442	1.3334	9.7783		6.3316	1.0491	5.3017	
		0.25 dB	2	1	0.7075	2.0271	64.1023	45.32720	0.8098	1.4541	45.3272
	3	0		1.6157		10.11407		0.7748		6.39522	
	3	1	1.0095	1.6804	11.6788		1.5305	1.1567	7.3732		
	4	1	0.5449	1.4623	5.2094	4.81286	0.6654	0.6971	3.1456	2.91918	
	4	2	1.3577	1.4168	12.5767		2.7212	1.0755	7.4297		
	5	0		1.4332		3.11565		0.4814		1.88699	
	5	1	0.6402	1.3816	3.2760		1.1118	0.7747	1.9694		
	5	2	1.7657	1.3114	5.3007		4.6482	1.0416	3.0475		
	6	1	0.5241	1.4250	2.4406	2.35734	0.6538	0.5203	1.4996	1.46216	
	6	2	0.7705	1.3241	3.3339		1.8792	0.8521	1.8929		
	6	3	2.2442	1.2420	9.1083		7.8068	1.0242	4.9008		
	0.50 dB	2	1	0.7075	1.6923	53.5142	37.84029	0.8643	1.2316	37.8403	26.7618
		3	0		1.4334		8.97286		0.6336		5.67935
		3	1	1.0095	1.4199	10.3610		1.7346	1.0693	6.5451	
		4	1	0.5449	1.3390	4.7699	4.40683	0.7147	0.6195	2.8889	2.68325
		4	2	1.3577	1.2973	11.5157		3.1795	1.0311	6.7941	
		5	0		1.3388		2.91045		0.4028		1.77664
		5	1	0.6402	1.2906	3.0603		1.2727	0.7354	1.8523	
		5	2	1.7657	1.2250	4.9516		5.5513	1.0166	2.8471	
6		1	0.5241	1.3499	2.3119	2.23304	0.7032	0.4704	1.4357	1.40138	
6		2	0.7705	1.2543	3.1581		2.2239	0.8317	1.8275		
6		3	2.2442	1.1785	8.6280		9.5116	1.0080	4.6250		
1.00 dB		2	1	0.7075	1.4022	44.3421	31.35464	0.9573	1.0503	31.3548	22.17880
		3	0		1.2656		7.92242		0.5004		5.02121
		3	1	1.0095	1.2536	9.1480		2.0563	0.9984	5.7834	
		4	1	0.5449	1.2221	4.3535	4.02211	0.7966	0.5513	2.6465	2.46079
		4	2	1.3577	1.1840	10.5103		3.8880	0.9954	6.1909	
		5	0		1.2479		2.71285		0.3255		1.67162
2.00 dB		3	0		1.2479		2.71285		0.3255		1.67162
	3	1	1.0095	1.0958	7.9962		2.6103	0.9436	5.0609		
	4	1	0.5449	1.1076	3.9459	3.64547	0.9461	0.4945	2.4102	2.24440	
	4	2	1.3577	1.0732	9.5261		5.0926	0.9846	0.9637		
	5	0		1.1157		2.51614		0.3985		1.58660	
	5	1	0.6402	1.1158	2.6457		1.9585	0.6807	1.6037		
	5	2	1.7657	1.0591	4.2808		9.3369	0.9823	2.4619		
	6	1	0.5241	1.2033	2.0609	1.99600	0.9334	0.3894	1.3150	1.28693	
	6	2	0.7705	1.1181	2.8152		3.6694	0.8081	1.6428		
	6	3	2.2442	1.0481	7.6911		16.3999	0.9846	3.9637		

With Table 111.2, an extra step is required in order to find the normalized element values for the circuit of Fig. 111.5. We must use Eq. (111.9) to find the element values for each stage from the values found in the table for  $Q$  ( $Q_1, Q_2, \dots$ ),  $\omega_0$  ( $\omega_1, \omega_2, \dots$ ), and  $\omega_n$  ( $\omega_{n1}, \omega_{n2}, \dots$ ). For odd-order filters, we must also add a first-order RC low-pass filter to the cascade with normalized values  $R = 1.0$  and  $C = \omega_0$ . To obtain practical values, we then denormalize by impedance scaling to a practical impedance level (we have chosen 10 k $\Omega$ ) and frequency scaling to  $\omega_p = 1$  kHz. Thus we multiply all resistors by  $10^4$  and divide all capacitors by  $10^4$  to impedance-scale, and we then divide all capacitors by  $2\pi \times 10^3$  to frequency-scale. The practical values for the resistors and the capacitors become:

*Inverse Chebyshev fourth-order* ( $\alpha_p = 1.00$  dB):

Stage 1	$R_a = 5.35 \text{ k}\Omega$	$R_b = 20.8 \text{ k}\Omega$	$R_c = 37.5 \text{ k}\Omega$
	$R = 5.54 \text{ k}\Omega$	$C = 12.2 \text{ nF}$	
Stage 2	$R_a = 9.83 \text{ k}\Omega$	$R_b = 394 \text{ k}\Omega$	$R_c = 267 \text{ k}\Omega$
	$R = 14.8 \text{ k}\Omega$	$C = 13.5 \text{ nF}$	

*Elliptical-Cauer third-order* ( $\alpha_p = 1.00$  dB)

Stage 0	$R_0 = 10.0 \text{ k}\Omega$	$C_0 = 30.4 \text{ nF}$	
Stage 1	$R_a = 9.74 \text{ k}\Omega$	$R_b = 167 \text{ k}\Omega$	$R_c = 75.7 \text{ k}\Omega$
	$R = 22.1 \text{ k}\Omega$	$C = 15.9 \text{ nF}$	

Note: All resistors labeled 1 - in Fig. 111.5 become 10 k $\Omega$ .

## Defining Terms

**Active RC filter:** A filter circuit that uses an active device (FET, transistor, op-amp, transconductance amp, etc.) with resistors and capacitors without the need for inductors. Active RC filters are particularly suited for integrated circuit applications.

**Bessel-Thompson filter:** A mathematical approximation to an ideal filter in which the group delay of the transfer function in the frequency domain is maximally flat. Bessel-Thompson filters are optimum in the sense that they offer the least droop without overshoot in the group delay.

**Biquad:** A filter circuit that implements a second-order filter function. High-order filters are often built from a cascade of second-order blocks. Each of these second-order blocks is referred to as a biquad because in the  $s$  domain these filters are characterized by a quadratic function of  $s$  in both the numerator and denominator.

**Butterworth filter:** A mathematical approximation to an ideal filter in which the magnitude of the transfer function in the frequency domain is maximally flat. Butterworth filters are optimum in the sense that they provide the least droop without overshoot in the magnitude response.

**Cascade:** A circuit topology in which a series of simple circuits (often biquads) are connected with the output of the first to the input of the second, the output of the second to the input of the third, and so forth. Using this configuration, a complex filter can be constructed from a "cascade" of simpler circuits.

**Chebyshev filter:** A mathematical approximation to an ideal filter in which the magnitude of the transfer function has a series of ripples in the passband that are of equal amplitude. Chebyshev filters are optimum in the sense that they offer the sharpest transition band for a given filter order for an all-pole filter.

**Elliptical-Cauer filter:** A mathematical approximation of an ideal filter in which the magnitude of the transfer function has a series of ripples in both the passband and stopband that are of equal amplitude. Elliptical-Cauer filters are optimum in the sense that they provide the sharpest possible transition band for a given filter order.

**Filter circuit:** An electronic circuit that passes some frequencies and rejects others so as to separate signals according to their frequency content. Common filter types are low-pass, band-pass, and high-pass, which refer to whether they pass low, middle, or high frequencies, respectively.

**Inverse Chebyshev filters:** A mathematical approximation to an ideal filter exhibiting a Butterworth magnitude response in the passband and an elliptical-Cauer magnitude response in the stopband.

**LC ladder filter circuits:** A planar circuit topology resembling a ladder in which the rungs of the ladder (referred to as the *parallel circuit elements*) consist of inductors or capacitors, one rail of the ladder is the ground plane, and the other rail (referred to as the *series circuit elements*) consists of a series connection of inductors and capacitors.

**Passband:** The frequencies for which a filter passes the signal from the input to the output without significant attenuation.

**Passive RLC filter:** A filter circuit that uses only passive devices, resistors, capacitors, inductors, and mutual inductors.

**Pole frequency ( $\omega_0$ ):** A parameter of a biquad filter used along with  $Q$  to define the location of the complex second-order pole pair in the  $s$  domain. The pole frequency  $\omega_0$  is the distance of the second-order pole pair from the origin in the  $s$  plane. This frequency is approximately equal to the cutoff frequency for low-pass and high-pass filters and to the resonant frequency for band-pass filters (see  $Q$ ).

**$Q$ :** A parameter of a biquad circuit used along with  $\omega_0$  to define the location of the complex second-order pole pair in the  $s$  plane. The real part of the pole pair is located in the  $s$  plane at  $\omega_0/2Q$ . The two poles are at a distance  $\omega_0$  from the origin in the  $s$  plane.

**Stopband:** The frequencies for which a filter provides significant attenuation for signals between the input and the output.

**Transition band:** The frequencies between a passband and stopband for which there are no filter specifications that must be met. Transition bands are necessary to allow for practical filter circuits, but it is desirable to keep transition bands as narrow as possible (i.e., a sharp

transition band).

## References

- Antoniou, A. 1967. Gyrator using operational amplifiers. *Electronic Letters*, vol. 3, pp. 350–352.
- Bowron, P. and Stephenson, F. W. 1979. *Active Filters for Communications and Instrumentation*. McGraw-Hill Book Co., London.
- Bruton, L. T., Pederson, R. T., and Treleaven, D. H. 1972. Low-frequency compensation of FDNR low-pass filters. *Proc. IEEE*. 60:444–445.
- Daniels, R. W. 1974. *Approximation Methods for Electronic Filter Design*. McGraw-Hill Book Co., New York.
- Ghausi, M. S. and Laker, K. R. 1981. *Modern Filter Design*. Prentice-Hall, Inc., Englewood Cliffs, NJ.
- Hamilton, T. A. and Sedra, A. S. 1971. A novel application of a gyrator-type circuit. *Fifth IEEE Asilomar Conference on Circuits and Systems*, Pacific Grove, CA, pp. 343–348.
- Hansell, G. E. 1969. *Filter Design and Evaluation*. Van Nostrand Reinhold Co., New York.
- Haritantis, I., Constantinides, A. G., and Deliyannis, T. 1976. Wave active filters. *Proc. IEE (England)*. 123(7):676–682.
- Koller, R. D. and Wilamowski, B. M. 1994. "Ladder" Filter Design Program. Electrical Engineering Dept., University of Wyoming, Laramie, WY 82071 (email: koller@uwyo.edu).
- Lindquist, C. S. 1977. *Active Network Design with Signal Filtering Applications*. Steward & Sons, Long Beach, CA.
- Mitra, S. K. 1971. *Active Inductorless Filters*. IEEE Press, New York.
- Sallen, R. P. and Key, E. L. 1955. A practical method of designing RC active filters. *IRE Trans. Circuits Syst.*, CT-2:74–85.
- Schaumann, R., Ghausi, M. S., and Laker, K. R. 1990. *Design of Analog Filters: Passive, Active-RC and Switched Capacitors*. Prentice-Hall, Inc., Englewood Cliffs, NJ.
- Schaumann, R., Soderstrand, M. A., and Laker, K. 1976. *Modern Active Filter Design*. IEEE Press, New York.
- Sedra, A. S. and Brackett, P. O. 1978. *Filter Theory and Design: Active and Passive*. Matrix Publishers, Inc., Champaign, IL.
- Tsividis, Y. P. and Voorman, J. O. 1993. *Integrated Continuous-Time Filters: Principles, Design, and Applications*. IEEE Press, New York.
- Wilamowski, B. M. 1994. "Filter" Filter Design Program. Electrical Engineering Dept., University of Wyoming, Laramie, WY 82071 (email: wilam@uwyo.edu).
- Williams, A. B. 1975. *Active Filter Design*. Artech House, Inc., Dedham, MA.
- Zverev, A. I. 1967. *Handbook of Filter Synthesis*. John Wiley & Sons, New York.

## Further Information

The following are recommended for further information on active filter design:

- Schaumann, R., Ghausi, M. S., and Laker, K. R. 1990. *Design of Analog Filters: Passive, Active-RC and Switched Capacitors*. Prentice-Hall, Inc., Englewood Cliffs, NJ.
- Sedra, A. S. and Brackett, P. O. 1978. *Filter Theory and Design: Active and Passive*. Matrix

Publishers, Inc., Champaign, IL.

Tsividis, Y. P. and Voorman, J. O. 1993. *Integrated Continuous-Time Filters: Principles, Design, and Applications*. IEEE Press, New York.



Soclof, S. "Diodes and Transistors"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## 112.1 Semiconductors

### 112.2 Bipolar Junction Transistors

### 112.3 Junction Field-Effect Transistors

JFET as an Amplifier—Small-Signal AC Voltage Gain

### 112.4 Metal-Oxide Silicon Field-Effect Transistors

MOSFET as an Amplifier—Small-Signal AC Voltage Gain • MOSFETs for Digital Circuits

## Sidney Soclof

*California State University*

Transistors form the basis of all modern electronic devices and systems, including the integrated circuits used in systems ranging from radio and TVs to computers. Transistors are solid-state electron devices made out of a category of materials called *semiconductors*. The most widely used semiconductor for transistors, by far, is silicon, although gallium arsenide, which is a compound semiconductor, is used for some very-high-speed applications. We will start off with a very brief discussion of semiconductors. Then there will be a short discussion of PN junctions and diodes, followed by a section on the three major types of transistors, the bipolar junction transistor (BJT), the junction field-effect transistor (JFET), and the metal-oxide silicon field-effect transistor (MOSFET).

## 112.1 Semiconductors

---

Semiconductors are a category of materials with an electrical conductivity that is intermediate between that of the good conductors and the insulators. The good conductors, which are all metals, have electrical resistivities down in the range of  $10^{-6}$ – $\cdot\text{cm}$ . The insulators have electrical resistivities that are up in the range of  $10^6$  to as much as about  $10^{12}$ – $\cdot\text{cm}$ . Semiconductors have resistivities that are generally in the range of  $10^{-4}$  to  $10^4$ – $\cdot\text{cm}$ . The resistivity of a semiconductor is strongly influenced by impurities, called **dopants**, that are purposely added to the material to change the electronic characteristics.

We will first consider the case of the pure, or **intrinsic**, semiconductor. As a result of the thermal energy present in the material, electrons can break loose from covalent bonds and become free electrons able to move through the solid and contribute to the electrical conductivity. The covalent bonds left behind have an electron vacancy called a **hole**. Electrons from neighboring covalent bonds can easily move into an adjacent bond with an electron vacancy, or hole, and thus the hole can move from one covalent bond to an adjacent bond. As this process continues, we can say that

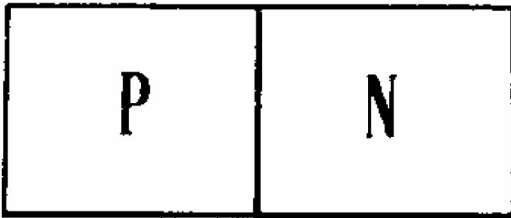
the hole is moving through the material. These holes act as if they have a positive charge equal in magnitude to the electron charge, and they can also contribute to the electrical conductivity. Thus, in a semiconductor there are two types of mobile electrical charge carriers that can contribute to the electrical conductivity: the free electrons and the holes. Since the electrons and holes are generated in equal numbers and recombine in equal numbers, the free electron and hole populations are equal.

In the **extrinsic** or **doped semiconductor**, impurities are purposely added to modify the electronic characteristics. In the case of silicon, every silicon atom shares its four valence electrons with each of its four nearest neighbors in covalent bonds. If an impurity or "dopant" atom with a valency of five, such as phosphorus, is substituted for silicon, four of the five valence electrons of the dopant atom will be held in covalent bonds. The extra, or fifth, electron will not be in a covalent bond and is loosely held. At room temperature, almost all of these extra electrons will have broken loose from their parent atoms and become free electrons. These pentavalent dopants thus donate free electrons to the semiconductor and are called **donors**. These donated electrons upset the balance between the electron and hole populations, so there are now more electrons than holes. This is now called an **N-type semiconductor**, in which the electrons are the **majority carriers** and holes are the **minority carriers**. In an N-type semiconductor the free electron concentration is generally many orders of magnitude larger than the hole concentration.

If an impurity or dopant atom with a valency of three, such as boron, is substituted for silicon, three of the four valence electrons of the dopant atom will be held in covalent bonds. One of the covalent bonds will be missing an electron. An electron from a neighboring silicon-to-silicon covalent bond, however, can easily jump into this electron vacancy, thereby creating a vacancy, or hole, in the silicon-to-silicon covalent bond. These trivalent dopants thus accept free electrons, thereby generating holes, and are called **acceptors**. These additional holes upset the balance between the electron and hole populations, so there are now more holes than electrons. This is now called a **P-type semiconductor**, in which the holes are the majority carriers and the electrons are the minority carriers. In a P-type semiconductor the hole concentration is generally many orders of magnitude larger than the electron concentration.

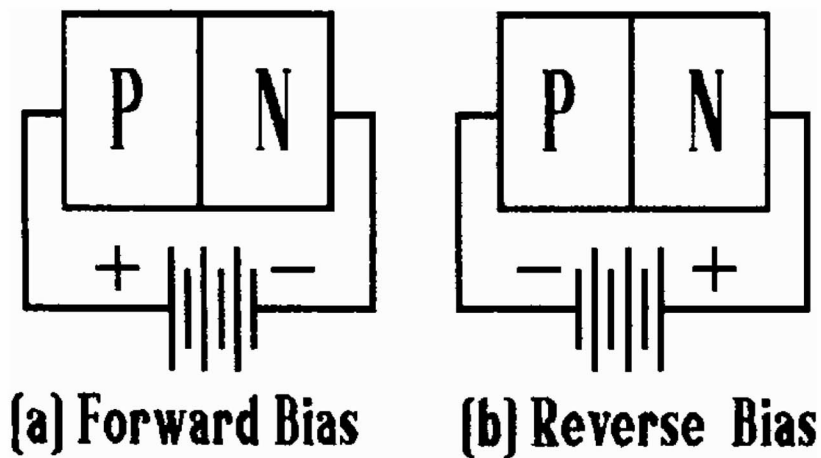
Figure 112.1 shows a single crystal chip of silicon that is doped with acceptors to make it a P-type on one side and doped with donors to make it N-type on the other side. The transition between the two sides is called the *PN junction*. As a result of the concentration difference of the free electrons and holes there will be an initial flow of these charge carriers across the junction, which will result in the N-type side attaining a net positive charge with respect to the P-type side. This results in the formation of an electric potential "hill" or barrier at the junction. Under equilibrium conditions the height of this potential hill, called the **contact potential**, is such that the flow of the majority carrier holes from the P-type side up the hill to the N-type side is reduced to the extent that it becomes equal to the flow of the minority carrier holes from the N-type side down the hill to the P-type side. Similarly, the flow of the majority carrier free electrons from the N-type side is reduced to the extent that it becomes equal to the flow of the minority carrier electrons from the P-type side. Thus the net current flow across the junction under equilibrium conditions is zero.

**Figure 112.1** PN junction.



In Fig. 112.2 the silicon chip is connected as a *diode*, or two-terminal electrode device. The situation in which a bias voltage is applied is shown. In Fig. 112.2(a) the bias voltage is a **forward bias**, which produces a reduction in the height of the potential hill at the junction. This allows for a large increase in the flow of electrons and holes across the junction. As the forward bias voltage increases, the **forward current** will increase at approximately an exponential rate and can become very large.

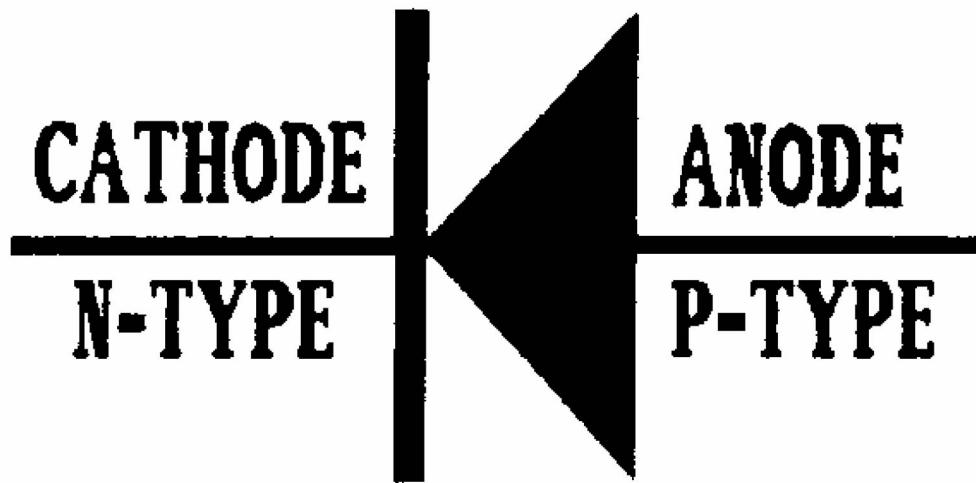
**Figure 112.2** Diode.



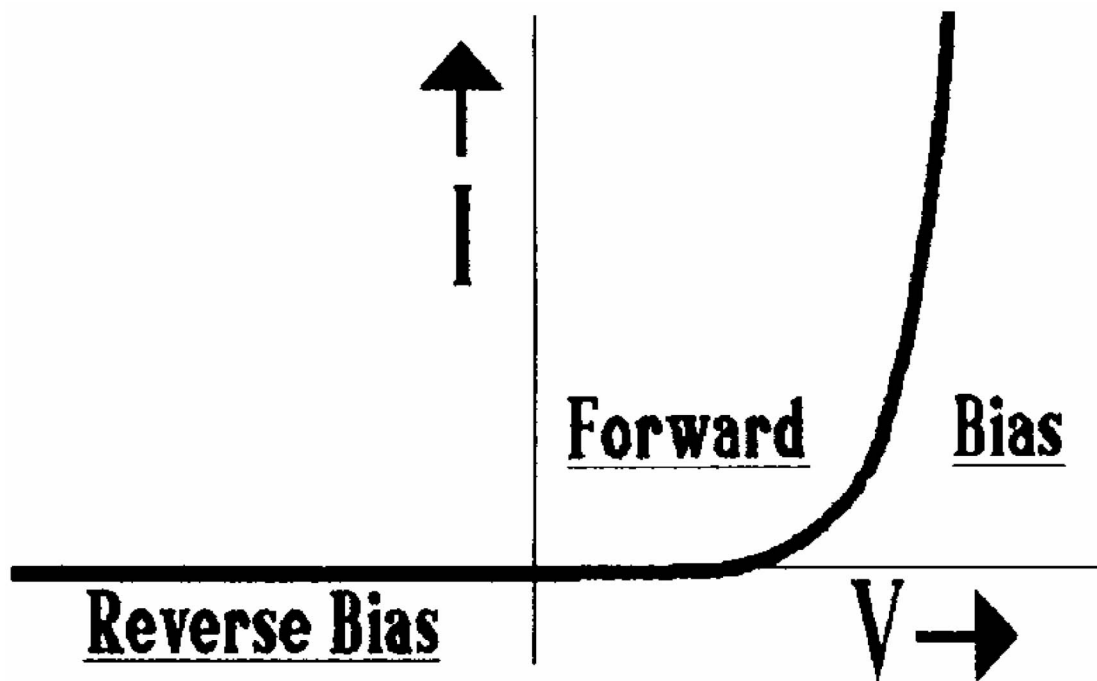
In Fig. 112.2(b) the bias voltage is a **reverse bias**, which produces an increase in the height of the potential hill at the junction. This essentially chokes off the flow of (1) electrons from the N-type side to the P-type side, and (2) holes from the P-type side to the N-type side. The only thing left is the very small trickle of electrons from the P-type side and holes from the N-type side. Thus the **reverse current** of the diode will be very small.

In Fig. 112.3 the circuit schematic symbol for the diode is shown, and in Fig. 112.4 a graph of the current versus voltage curve for the diode is presented. The P-type side of the diode is called the **anode**, and the N-type side is the **cathode** of the diode. The forward current of diodes can be very large—in the case of large power diodes, up into the range of 10 to 100 A. The reverse current is generally very small, often down in the low nanoampere or even picoampere range.

**Figure 112.3** Diode symbol.



**Figure 112.4** Current versus voltage curve.



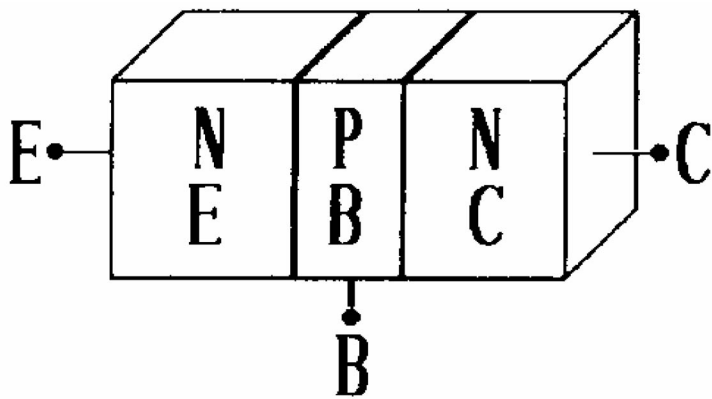
The diode is basically a one-way voltage-controlled current switch. It allows current to flow in the forward direction when a forward bias voltage is applied, but when a reverse bias is applied the current flow becomes extremely small. Diodes are used extensively in electronic circuits. Applications include rectifiers to convert AC to DC, waveshaping circuits, peak detectors, DC level shifting circuits, and signal transmission gates. Diodes are also used for the demodulation of amplitude-modulated (AM) signals.

## 112.2 Bipolar Junction Transistors

---

A basic diagram of the bipolar junction transistor, or BJT, is shown in [Fig. 112.5](#). Whereas the diode has one PN junction, the BJT has two PN junctions. The three regions of the BJT are the emitter, base, and collector. The middle, or base, region is very thin, generally less than 1 micrometer wide. This middle electrode, or base, can be considered as the control electrode that controls the current flow through the device between emitter and collector. A small voltage applied to the base (i.e., between base and emitter) can produce a large change in the current flow through the BJT.

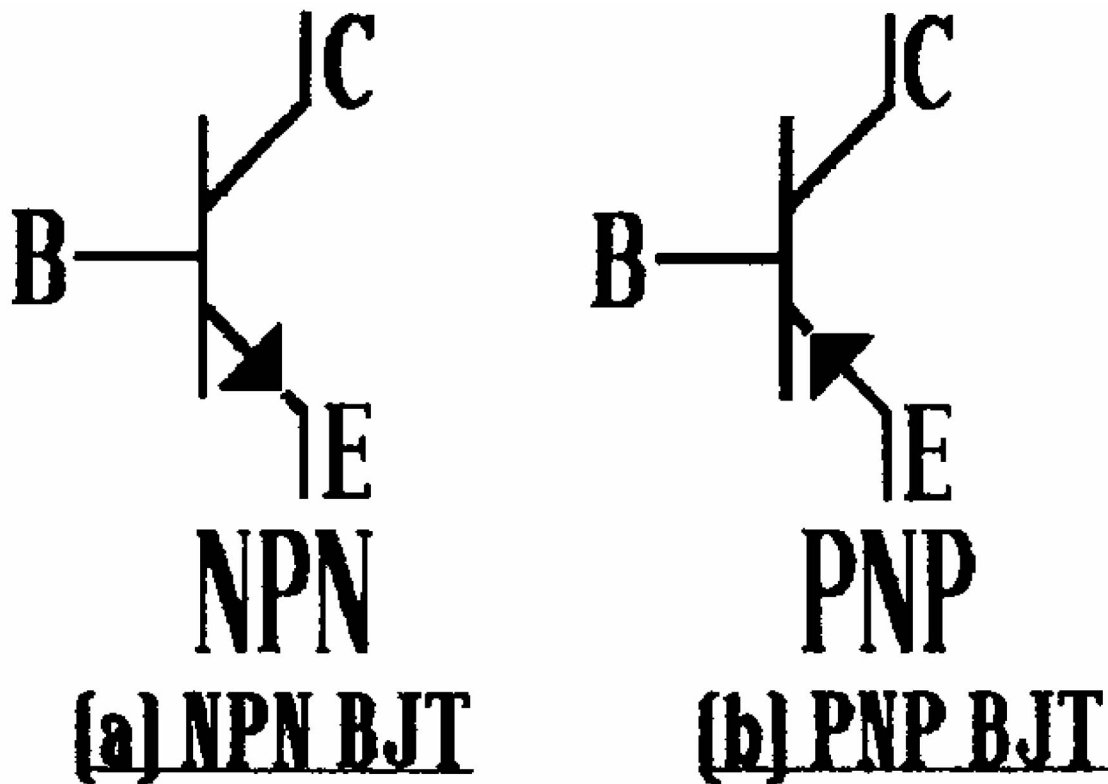
**Figure 112.5** Bipolar junction transistor.



BJTs are often used for the amplification of electrical signals. In these applications the emitter-base PN junction is turned "on" (forward biased) and the collector-base PN junction is "off" (reverse biased). For the NPN BJT shown in Fig. 112.5, the emitter will emit electrons into the base region. Since the P-type base region is so very thin, most of these electrons will survive the trip across the base and reach the collector-base junction. When the electrons reach the collector-base junction they will "roll downhill" into the collector and thus be collected by the collector to become the collector current,  $I_C$ . The emitter and collector currents will be approximately equal, so  $I_C \cong I_E$ . There will be a small base current,  $I_B$ , resulting from the emission of holes from the base across the emitter-base junction into the emitter. There will also be a small component of the base current due to the recombination of electrons and holes in the base. The ratio of collector current to base current, given by the parameter  $\beta$  or  $h_{FE}$ , is  $\beta = I_C/I_B$ , which will be very large, generally up in the range of 50 to 300 for most BJTs.

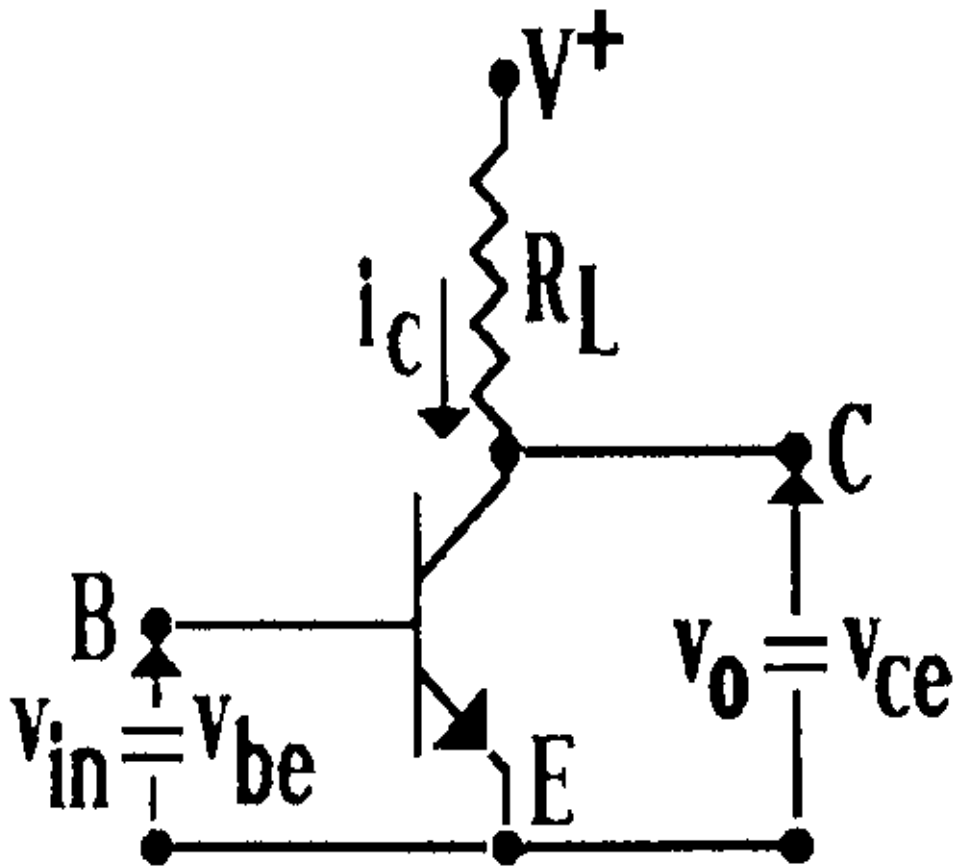
In Fig. 112.6(a) the circuit schematic symbol for the NPN transistor is shown, and in Fig. 112.6(b) the corresponding symbol for the PNP transistor is given. The basic operation of the PNP transistor is similar to that of the NPN, except for a reversal of the polarity of the algebraic signs of all DC currents and voltages.

**Figure 112.6** BJT schematic symbols.



In Fig. 112.7 the operation of a BJT as an amplifier is shown. When the BJT is operated as an amplifier the emitter-base PN junction is turned "on" (forward biased) and the collector-base PN junction is "off" (reverse biased). An AC input voltage applied between base and emitter,  $v_{in} = v_{be}$ , can produce an AC component,  $i_c$ , of the collector current. Since  $i_c$  flows through a load resistor,  $R_L$ , an AC voltage,  $v_o = v_{ce} = -i_c \cdot R_L$ , will be produced at the collector. The AC small-signal voltage gain is  $A_V = v_o/v_{in} = v_{ce}/v_{be}$ .

**Figure 112.7** BJT amplifier.

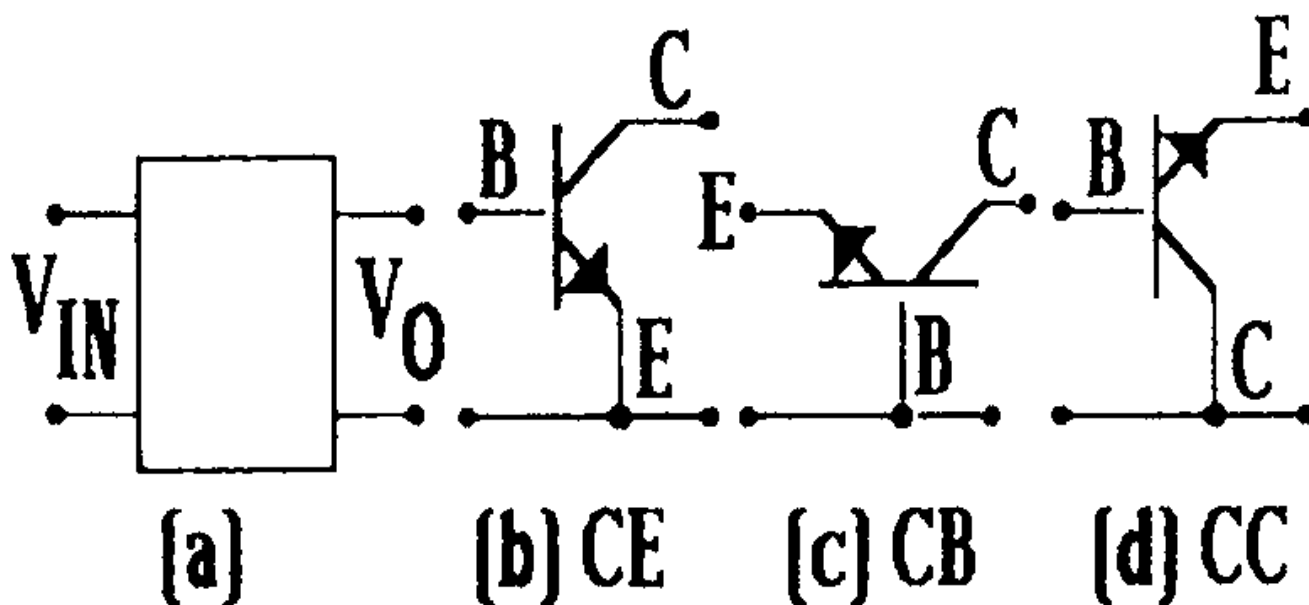




The collector current,  $I_C$ , of a BJT when operated as an amplifier is related to the base-to-emitter voltage,  $V_{BE}$ , by the exponential relationship  $I_C = I_{CO} \cdot \exp(V_{BE}/V_T)$ , where  $I_{CO}$  is a constant and  $V_T = \text{thermal voltage} = 25 \text{ mV}$ . The rate of change of  $I_C$  with respect to  $V_{BE}$  is given by the **transfer conductance**,  $g_m \geq dI_C/dV_{BE} = I_C/V_T$ . If the net load driven by the collector of the transistor is  $R_L$ , the AC small-signal voltage gain is  $A_V = v_{ce}/v_{be} = -g_m \cdot R_L$ . The negative sign indicates that the output voltage will be an amplified but inverted replica of the input signal. If, for example, the transistor is biased at a DC collector current level of  $I_C = 1 \text{ mA}$  and drives a net load of  $R_L = 10 \text{ k}\Omega$ , then  $g_m = I_C/V_T = 1 \text{ mA}/25 \text{ mV} = 40 \text{ mS}$ , and  $A_V = v_{ce}/v_{be} = -g_m \cdot R_L = -40 \text{ mS} \cdot 10 \text{ k}\Omega = -400$ . Thus we see that the voltage gain of a single BJT amplifier stage can be very large, often up in the range of 100 or more.

The BJT is a three-electrode or *triode* electron device. When connected in a circuit it is usually operated as a two-port, or two-terminal, pair device, as shown in Fig. 112.8. Therefore, one of the three electrodes of the BJT must be common to both the input and output ports. Thus there are three basic BJT configurations: common-emitter (CE), common-base (CB), and common-collector (CC). The most often used configuration, especially for amplifiers, is the common-emitter (CE), although the other two configurations are used in some applications.

**Figure 112.8** The BJT as a two-port device.

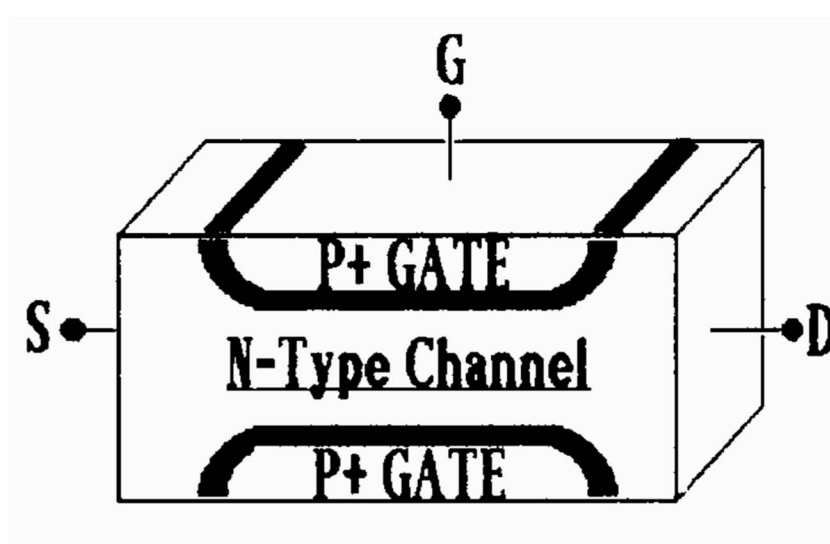


The BJT is often used as a switching device, especially in digital circuits, and in high-power applications. When used as a switching device, the transistor is switched between the *cutoff region*, in which both junctions are off, and the *saturation region*, in which both junctions are on. In the cutoff region the collector current is reduced to a very small value, down in the low nanoampere range, so the transistor looks essentially like an open circuit. In the saturation region the voltage drop between collector and emitter becomes very small, usually less than 0.1 volts, and the transistor looks like a very small resistance.

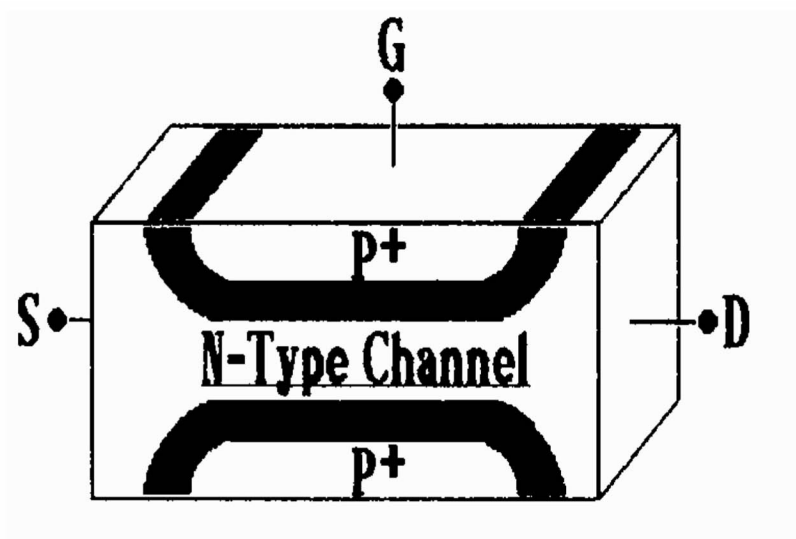
## 112.3 Junction Field-Effect Transistors

A *junction field-effect transistor* (JFET) is a type of transistor in which the current flow through the device between the drain and source electrodes is controlled by the voltage applied to the gate electrode. A simple physical model of the JFET is shown in Fig. 112.9. In this JFET an N-type conducting channel exists between drain and source. The gate is a heavily doped P-type region (designated as P<sup>+</sup> that surrounds the N-type channel. The gate-to-channel PN junction is normally kept reverse biased. As the reverse bias voltage between gate and channel increases, the depletion region width increases, as shown in Fig. 112.10. The depletion region extends mostly into the N-type channel because of the heavy doping on the P<sup>+</sup> side. The depletion region is depleted of mobile charge carriers and thus cannot contribute to the conduction of current between drain and source. Thus, as the gate voltage increases, the cross-sectional area of the N-type channel available for current flow decreases. This reduces the current flow between drain and source. As the gate voltage increases, the channel gets further constricted and the current flow gets smaller. Finally when the depletion regions meet in the middle of the channel, as shown in Fig. 112.11, the channel is pinched off in its entirety, all of the way between the source and the drain. At this point the current flow between drain and source is reduced to essentially zero. This voltage is called the **pinch-off voltage,  $V_p$** . The pinch-off voltage is also represented as  $V_{GS}(\text{off})$ , as being the gate-to-source voltage that turns the drain-to-source current,  $I_{DS}$ , off.

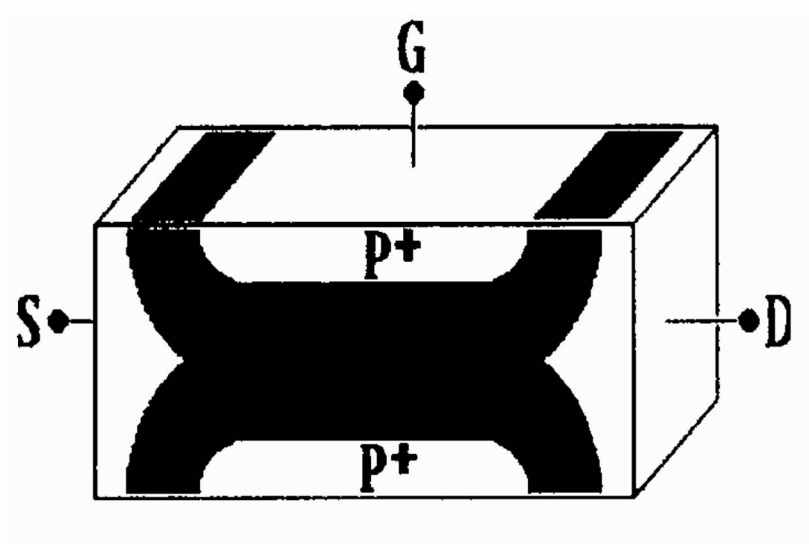
**Figure 112.9** JFET model.



**Figure 112.10** JFET with increased gate voltage.



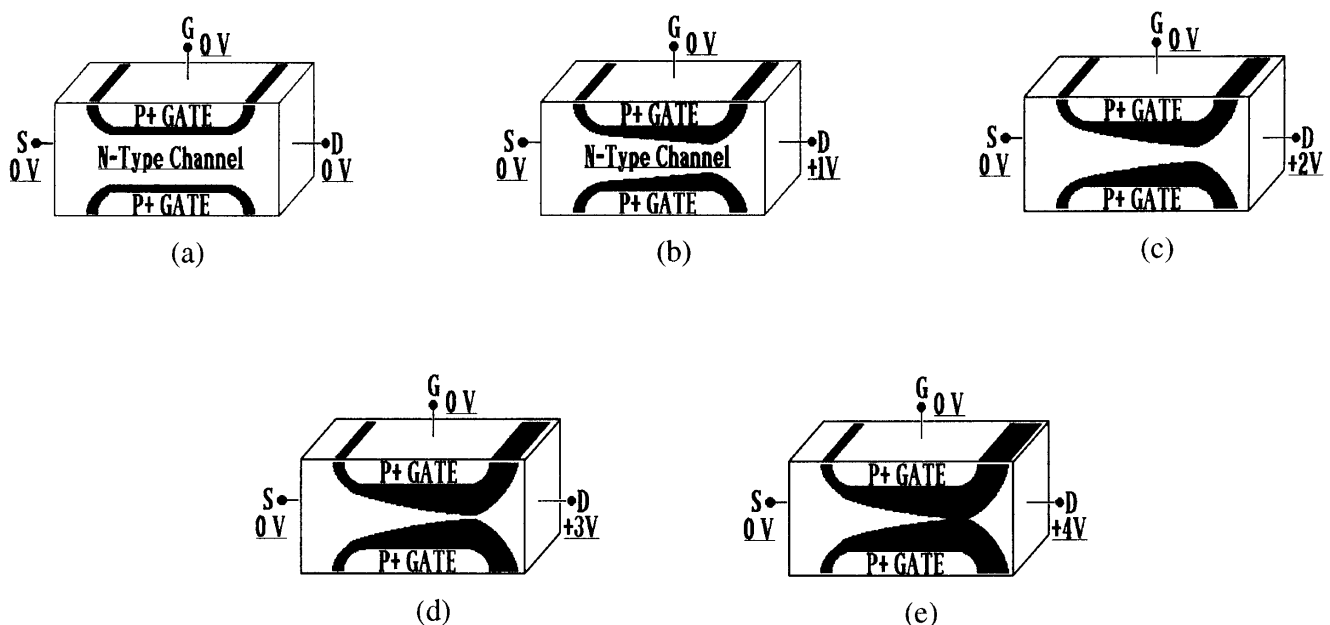
**Figure 112.11** JFET with pinched-off channel.



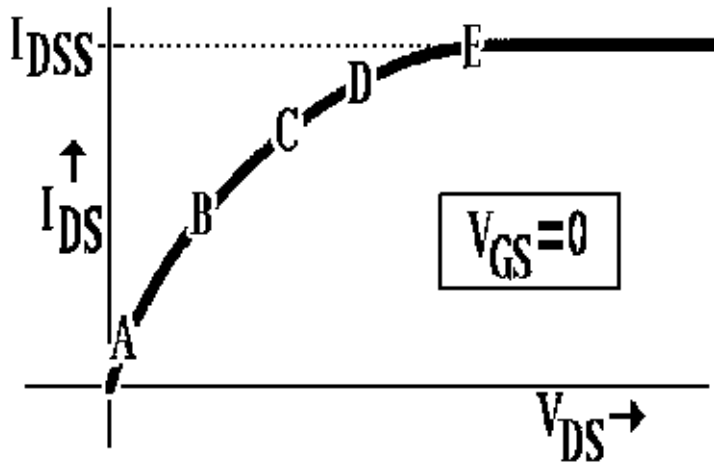
We have been considering here an N-channel JFET. The complementary device is the P-channel JFET, which has a heavily doped N-type P<sup>+</sup> gate region surrounding a P-type channel. The operation of a P-channel JFET is the same as for an N-channel device, except the algebraic signs of all DC voltages and currents are reversed.

We have been considering the case for  $V_{DS}$  small compared to the pinch-off voltage, such that the channel is essentially uniform from drain to source, as shown in Fig. 112.12(a). Now let's see what happens as  $V_{DS}$  increases. As an example, let's assume an N-channel JFET with a pinch-off voltage of  $V_P = -4$  V. We will see what happens for the case of  $V_{GS} = 0$  as  $V_{DS}$  increases. In Fig. 112.12(a) the situation is shown for the case of  $V_{DS} = 0$  in which the JFET is fully on and there is a uniform channel from source to drain. This is at point A on the  $I_{DS}$  versus  $V_{DS}$  curve of Fig. 112.13. The drain-to-source conductance is at its maximum value of  $g_{ds}(on)$ , and the drain-to-source resistance is correspondingly at its minimum value of  $r_{ds}(on)$ . Now let's consider the case of  $V_{DS} = +1$  V, as shown in Fig. 112.12(b). The gate-to-channel bias voltage at the source end is still  $V_{GS} = 0$ . The gate-to-channel bias voltage at the drain end is  $V_{GD} = V_{GS} - V_{DS} = -1$  V, so the depletion region will be wider at the drain end of the channel than at the source end. The channel will thus be narrower at the drain end than at the source end and this will result in a decrease in the channel conductance,  $g_{ds}$ , and correspondingly, an increase in the channel resistance,  $r_{ds}$ . So the slope of  $I_{DS}$  versus  $V_{DS}$  curve that corresponds to the channel conductance will be smaller at  $V_{DS} = 1$  V than it was at  $V_{DS} = 0$ , as shown at point B on the  $I_{DS}$  versus  $V_{DS}$  curve of Fig. 112.13.

**Figure 112.12** N-type channel.

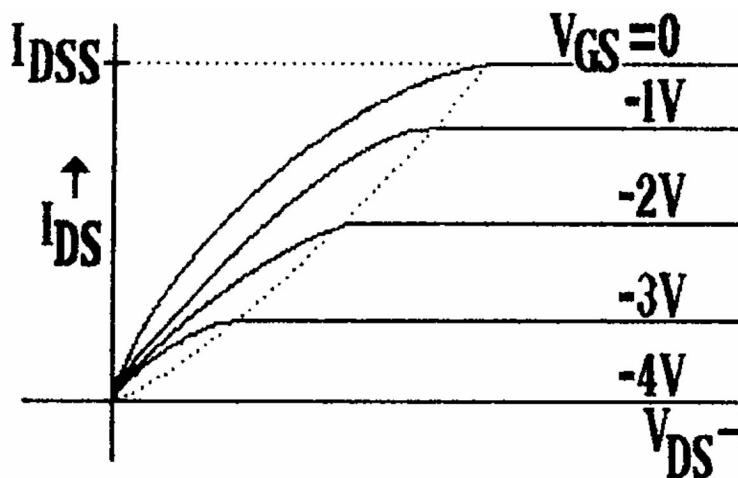


**Figure 112.13**  $I_{DS}$  versus  $V_{DS}$  curve.



In Fig. 112.12(c) the situation for  $V_{DS} = +2$  V is shown. The gate-to-channel bias voltage at the source end is still  $V_{GS} = 0$ , but the gate-to-channel bias voltage at the drain end is now  $V_{GD} = V_{GS} - V_{DS} = -2$  V, so the depletion region will now be substantially wider at the drain end of the channel than at the source end. This leads to a further constriction of the channel at the drain end and this will again result in a decrease in the channel conductance,  $g_{ds}$ , and correspondingly, as increase in the channel resistance,  $r_{ds}$ . So, the slope of the  $I_{DS}$  versus  $V_{DS}$  curve will be smaller at  $V_{DS} = 2$  V than it was at  $V_{DS} = 1$  V, as shown at point C on the  $I_{DS}$  versus  $V_{DS}$  curve of Fig. 112.13.

**Figure 112.14** JFET drain characteristics.



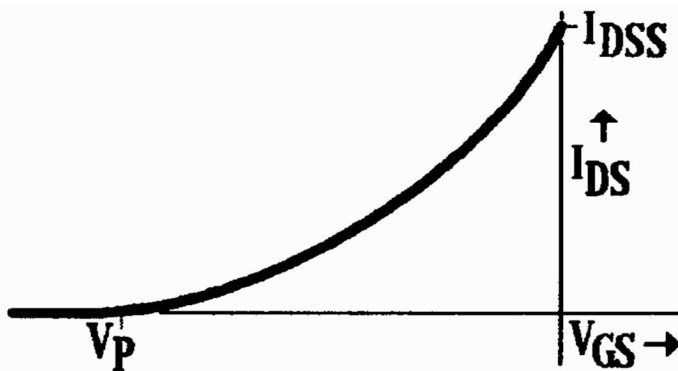
In Fig. 112.12(d) the situation for  $V_{DS} = +3 \text{ V}$  is shown, and this corresponds to point D on the  $I_{DS}$  versus  $V_{DS}$  curve of Fig. 112.13.

When  $V_{DS} = +4 \text{ V}$  the gate-to-channel bias voltage will be  $V_{GD} = V_{GS} - V_{DS} = 0 - 4 \text{ V} = -4 \text{ V} = V_P$ . As a result the channel is now pinched off at the drain end, but it is still wide open at the source end since  $V_{GS} = 0$ , as shown in Fig. 112.12(e). It is very important to note that the channel is pinched off just for a very short distance at the drain end, so that the drain-to-source current,  $I_{DS}$ , can still continue to flow. This is not at all the same situation as for the case of  $V_{GS} = V_P$ , wherein the channel is pinched off in its entirety, all of the way from source to drain. When this happens, it is like having a big block of insulator the entire distance between source and drain, and  $I_{DS}$  is reduced to essentially zero. The situation for  $V_{DS} = +4 \text{ V} = -V_P$  is shown at point E on the  $I_{DS}$  versus  $V_{DS}$  curve of Fig. 112.13.

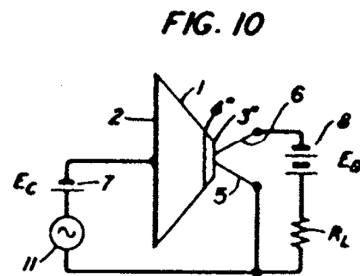
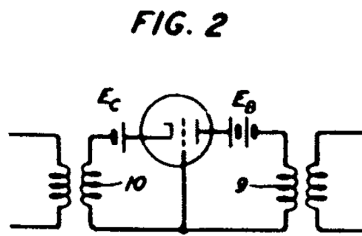
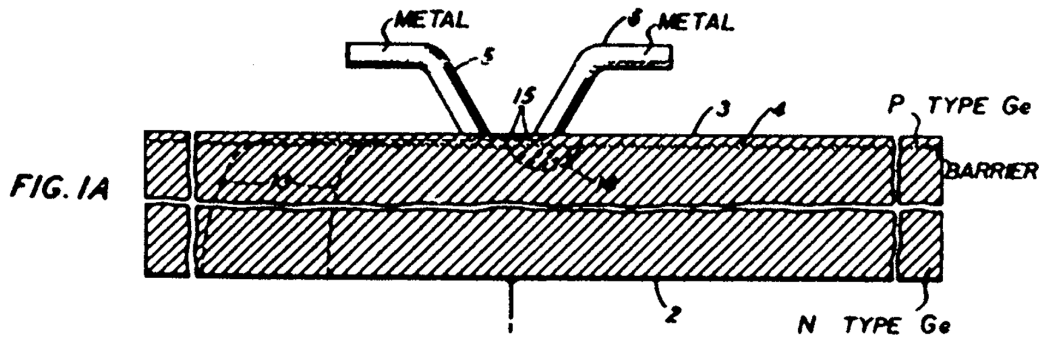
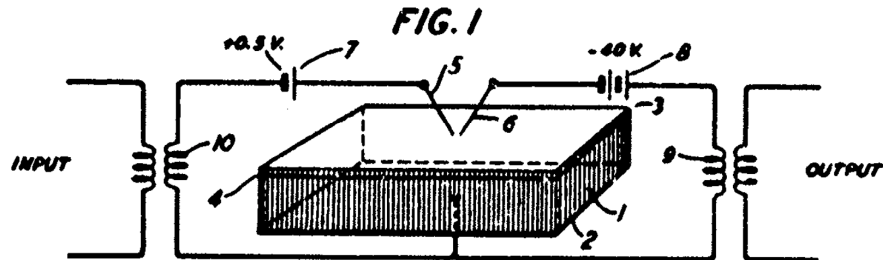
The region below the active region where  $V_{DS} < +4 \text{ V} = -V_P$  has several names. It is called the **nonsaturated region**, the **triode region**, and the **ohmic region**. The term *triode region* apparently originates from the similarity of the shape of the curves to that of the vacuum tube triode. The term *ohmic region* is due to the variation of  $I_{DS}$  with  $V_{DS}$  as in Ohm's law, although this variation is nonlinear except for the region where  $V_{DS}$  is small compared to the pinch-off voltage, where  $I_{DS}$  will have an approximately linear variation with  $V_{DS}$ .

So far we have looked at the  $I_{DS}$  versus  $V_{DS}$  curve only for the case of  $V_{GS} = 0$ . In Fig. 112.14 a family of curves  $I_{DS}$  versus  $V_{DS}$  for various constant values of  $V_{GS}$  is presented. These are called the *drain characteristics*, also known as the **output characteristics**, since the output side of the JFET is usually the drain side. In the active region where  $I_{DS}$  is relatively independent of  $V_{DS}$ , there is a simple approximate equation relating  $I_{DS}$  to  $V_{GS}$ . This is the "square law" **transfer equation** as given by  $I_{DS} = I_{DSS} [1 - (V_{GS}/V_P)]^2$ . In Fig. 12.15 a graph of the  $I_{DS}$  versus  $V_{GS}$  *transfer characteristics* for the JFET is presented. When  $V_{GS} = 0$ ,  $I_{DS} = I_{DSS}$  as expected, and as  $V_{GS} \rightarrow V_P$ ,  $I_{DS} \rightarrow 0$ . The lower boundary of the active region is controlled by the condition that the channel be pinched off at the drain end. To meet this condition the basic requirement is that the gate-to-channel bias voltage at the drain end of the channel,  $V_{GD}$ , be greater than the pinch-off voltage  $V_P$ . For the example under consideration with  $V_P = -4 \text{ V}$ , this means that  $V_{GD} = V_{GS} - V_{DS}$  be more negative than  $-4 \text{ V}$ . Therefore,  $V_{DS} - V_{GS} \geq +4 \text{ V}$ . Thus, for  $V_{GS} = 0$ , the active region will begin at  $V_{DS} = +4 \text{ V}$ . When  $V_{GS} = -1 \text{ V}$ , the active region will begin at  $V_{DS} = +3 \text{ V}$ , for now  $V_{GD} = -4 \text{ V}$ . When  $V_{GS} = -2 \text{ V}$ , the active region begins at  $V_{DS} = +2 \text{ V}$ , and when  $V_{GS} = -3 \text{ V}$ , the active region begins at  $V_{DS} = +1 \text{ V}$ . The dotted line in Fig. 112.14 marks the boundary between the nonsaturated and active regions.

**Figure 112.15** JFET transfer characteristics.



In the nonsaturated region  $I_{DS}$  is a function of both  $V_{GS}$  and  $I_{DS}$ , and in the lower portion of the nonsaturated region where  $V_{DS}$  is small compared to  $V_P$ ,  $I_{DS}$  becomes an approximately linear function of  $V_{DS}$ . This linear portion of the nonsaturated is called the *voltage-variable resistance* (VVR) region, for in this region the JFET acts like a linear resistance element between source and drain. The resistance is variable in that it is controlled by the gate voltage.



### THREE-ELECTRODE CIRCUIT ELEMENT UTILIZING SEMICONDUCTIVE MATERIALS

John Bardeen and

Walter H. Brattain

Patented October 3, 1950

#2,524,035

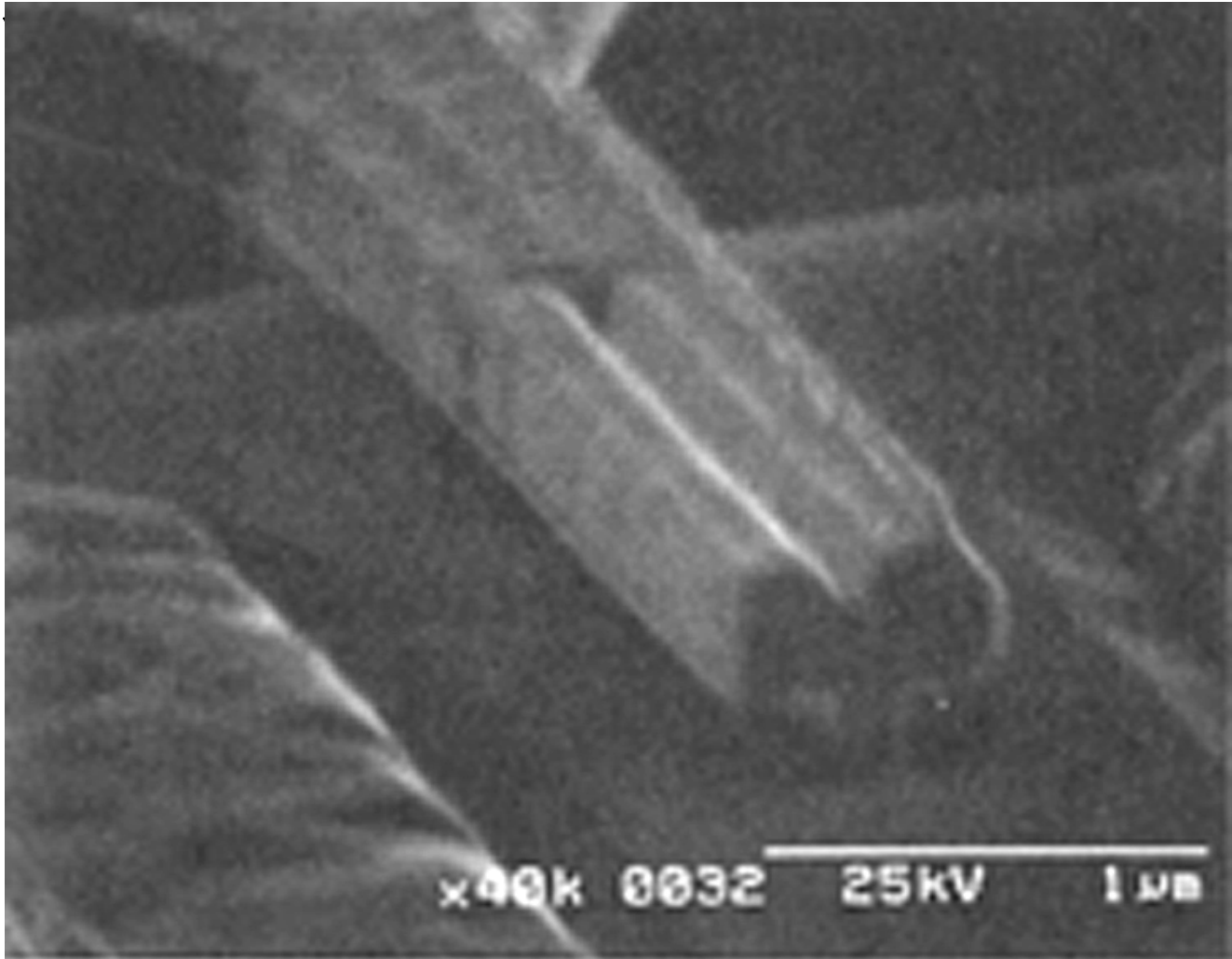
An excerpt:

The principal object of the invention is to amplify or otherwise translate electric signals or variations by use of compact, simple, and rugged apparatus of novel type.

Another object is to provide a circuit element for use as an amplifier or the like which does not require a heated thermionic cathode for its operation, and which therefore is immediately operative when turned on. A related object is to provide such a circuit element which requires no evacuated

or gas-filled envelope.

What Bardeen, Brattain, and Shockley invented at Bell Laboratories was the Transistor; the fundamental element in our modern electronic world. It soon replaced vacuum tubes in applications calling for amplifiers, oscillators and even relays in digital switching. (©1992, DewRay Products, Inc. Used with permission.)



This Schottky-collector Resonant Tunneling Diode (RTD) is believed to be the fastest active semiconductor-based device. Microwave and DC measurements indicate that it has a maximum frequency of oscillation ( $f_{\max}$ ) of 2.2 THz ( $2.2 \times 10^{12}$  Hz). Devices capable of such fast oscillation are potentially useful for extremely compact, high-frequency radiometers, radar systems, or space-based communications systems.

The semiconductor material consists of heavily doped n-type InGaAs beneath an AlAs/InGaAs quantum well structure. The T-shaped, air-bridged anode is in contact with a very thin top layer of fully depleted InGaAs. The novel top contact, together with its very small size of approximately



0.1 micron, results in much lower resistance than other, similar devices. This low resistance is responsible for the extremely high  $f_{\max}$  of the RTD.

The semiconductor layer structure was grown on an InP substrate by MBE (Molecular Beam Epitaxy). The T-anode was fabricated using an electron-beam lithography system and a multi-layer resist.

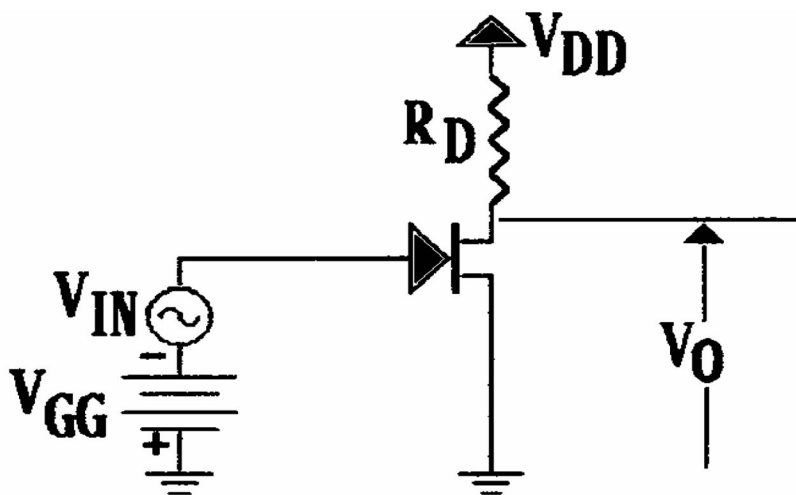
This device was developed by M. Reddy, M. J. Mondry, and M. J. W. Rodwell, U.C.S.B.; S. C. Martin, R. E. Muller, and R. P. Smith, CSMT, JPL/Caltech; and D. H. Chow and J. N. Schulman, Hughes Research Laboratories. (Photo courtesy of Madhukar Reddy.)

The work at U.C.S.B. was supported by ONR under contract #N00014-93-1-0378, NSF/PYI, and a JPL President's Fund. The work at JPL was performed by the Center for Space Microelectronics Technology, Jet Propulsion Laboratory, California Institute of Technology, and was sponsored by the National Aeronautics and Space Administration, Office of Advanced Concepts and Technology, and by the Innovative Science and Technology Office of BMDO through an agreement with NASA.

## JFET as an Amplifier - Small-Signal AC Voltage Gain

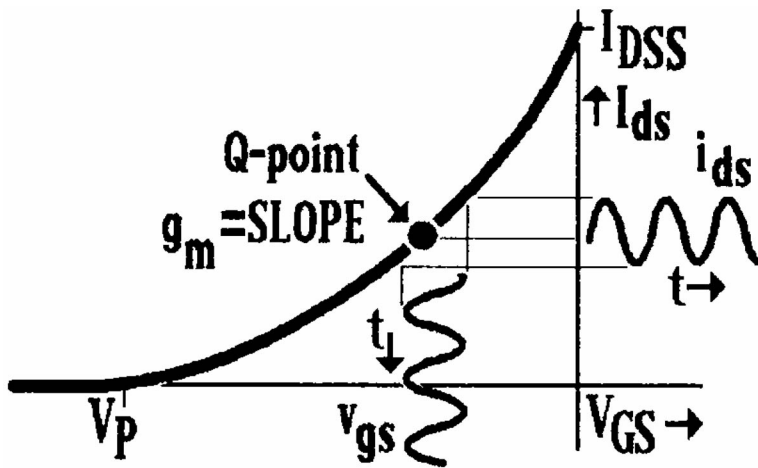
Let's consider the common-source amplifier of Fig. 112.16. The input AC signal is applied between gate and source, and the output AC voltage between is taken between drain and source. Thus the source electrode of this triode device is common to input and output, hence the designation of this JFET configuration as a common-source (CS) amplifier.

**Figure 112.16** Common source amplifier.



A good choice of the DC operating point or quiescent point ("Q-point") for an amplifier is in the middle of the active region at  $I_{DS} = I_{DSS} / 2$ . This allows for the maximum symmetrical drain current swing, from the quiescent level of  $I_{DSQ} = I_{DSS} / 2$ , down to a minimum of  $I_{DS} \cong 0$ , and up to a maximum of  $I_{DS} = I_{DSS}$ . This choice for the Q-point is also a good one from the standpoint of allowing for an adequate safety margin for the location of the actual Q-point due to the inevitable variations in device and component characteristics and values. This safety margin should keep the Q-point well away from the extreme limits of the active region, and thus ensure operation of the JFET in the active region under most conditions. If  $I_{DSS} = +10$  mA, then a good choice for the Q-point would thus be around +5 mA. The AC component of the drain current,  $i_{ds}$ , is related to the AC component of the gate voltage,  $v_{gs}$  by  $i_{ds} = g_m \cdot v_{gs}$  where  $g_m$  is the **dynamic transfer conductance** and is given by  $g_m = 2\sqrt{I_{DS} \cdot I_{DSS}} / (-V_P)$ . If  $V_P = -4$  V, then  $g_m = \sqrt{5 \text{ mA} \cdot 10 \text{ mA}} / 4 \text{ V} = 3.54 \text{ mA/V} = 3.54 \text{ mS}$ . If a small AC signal voltage,  $v_{gs}$ , is superimposed on the quiescent DS gate bias voltage  $V_{GSQ} = V_{GG}$  only a small segment of the transfer characteristic adjacent to the Q-point will be traversed, as shown in Fig. 112.17. This small segment will be close to a straight line, and as a result the AC drain current,  $i_{ds}$ , will have a waveform close to that of the AC voltage applied to the gate. The ratio of  $i_{ds}$  to  $v_{gs}$  will be the slope of the transfer curve as given by  $i_{ds}/v_{gs} \cong dI_{DS}/dV_{GS} = g_m$ . Thus  $i_{ds} \cong g_m \cdot v_{gs}$ . If the net load driven by the drain of the JFET is the drain load resistor,  $R_D$ , as shown in Fig. 112.16, then the AC drain current  $i_{ds}$  will produce an AC drain voltage of  $v_{ds} = -i_{ds} \cdot R_D$ . Since  $i_{ds} = g_m \cdot v_{gs}$ , this becomes  $v_{ds} = -g_m v_{GS} \cdot R_D$ . The AC small-signal voltage gain from gate to drain thus becomes  $A_V = v_O/v_{IN} = v_{ds}/v_{gs} = -g_m \cdot R_D$ . The negative sign indicates signal inversion as is the case for a common-source amplifier.

**Figure 112.17** JFET transfer characteristic.



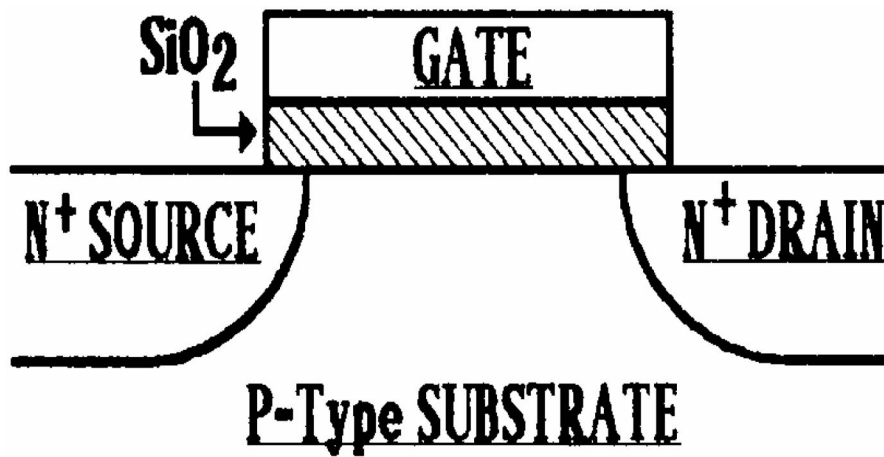
If the DC drain supply voltage is  $V_{DD} = +20$  V, a quiescent drain-to-source voltage of  $V_{DSQ} = V_{DD} / 2 = 10$  V will result in the JFET being biased in the middle of the active region. Since  $I_{DSQ} = 5$  mA in the example under consideration, the voltage drop across the drain load resistor,  $R_D$ , is 10 V. Thus  $R_D = 10 \text{ V} / 5 \text{ mA} = 2 \text{ k}\Omega$ . The AC small-signal voltage gain,  $A_V$ , thus becomes  $A_V = -g_m \cdot R_D = -3.54 \text{ mS} \cdot 2 \text{ k}\Omega = -7.07$ . Note that the voltage gain is relatively modest compared to the much larger voltage gains that can be obtained in a

bipolar-junction transistor (BJT) common-emitter amplifier. This is due to the lower transfer conductance of both JFETs and MOSFETs compared to BJTs. For a BJT the transfer conductance is given by  $g_m = I_C/V_T$ , where  $I_C$  is the quiescent collector current and  $V_T = kT/q \cong 25$  mV is the "thermal voltage." At  $I_C = 5$  mA,  $g_m = 5$  mA / 25 mV = 200 mS for the BJT, as compared to only 3.5 mS for the JFET in this example. With a net load of 2 k $\Omega$ , the BJT voltage gain will be  $-400$  as compared to the JFET voltage gain of only 7.1. Thus FETs have the disadvantage of a much lower transfer conductance and therefore lower voltage gain than BJTs operating under similar quiescent current levels, but they do have the major advantage of a much higher input impedance and a much lower input current. In the case of a JFET the input signal is applied to the reverse-biased gate-to-channel PN junction and thus sees a very high impedance. In the case of a common-emitter BJT amplifier the input signal is applied to the forward-biased base-emitter junction and the input impedance is given approximately by  $r_{IN} = r_{BE} \cong 1.5 \cdot \beta \cdot V_T/I_C$ . If  $I_C = 5$  mA and  $\beta = 200$ , for example, then  $r_{IN} \cong 1500\Omega$ . This moderate input resistance value of 1.5 k $\Omega$  is certainly no problem if the signal source resistance is less than around 100 $\Omega$ . However, if the source resistance is above 1 k $\Omega$ , then there will be a substantial signal loss in the coupling of the signal from the signal source to the base of the transistor. If the source resistance is in the range of above 100 k $\Omega$ , and certainly if it is above 1 M $\Omega$ , there will be severe signal attenuation due to the BJT input impedance, and a FET amplifier will probably offer a greater overall voltage gain. Indeed, when high impedance signal sources are encountered, a multistage amplifier with a FET input stage, followed by cascaded BJT stages, is often used.

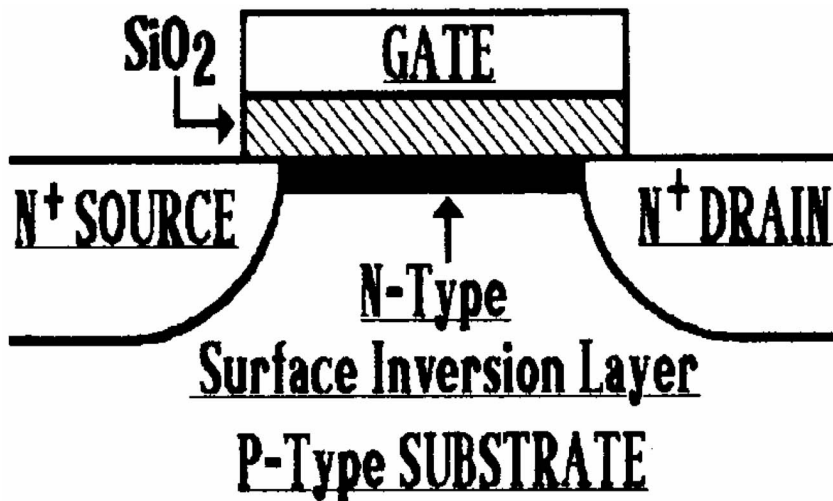
## 112.4 Metal-Oxide Silicon Field-Effect Transistors

A **metal-oxide silicon field-effect transistor** (MOSFET) is similar to a JFET, in that it is a type of transistor in which the current flow through the device between the drain and source electrodes is controlled by the voltage applied to the gate electrode. A simple physical model of the MOSFET is shown in Fig. 112.18. The gate electrode is electrically insulated from the rest of the device by a thin layer of silicon dioxide (SiO<sub>2</sub>). In the absence of any gate voltage there is no conducting channel between source and drain, so the device is off and  $I_{DS} \cong 0$ . If now a positive voltage is applied to the gate, electrons will be drawn into the silicon surface region immediately underneath the gate oxide. If the gate voltage is above the **threshold voltage**,  $V_T$ , there will be enough electrons drawn into this silicon surface region to make the electron population greater than the hole population. This surface region under the oxide will thus become N-type; this region will be called an *N-type surface inversion layer*. This N-type surface inversion layer will now constitute an N-type conducting channel between source and drain, as shown in Fig. 112.19, so that now current can flow and  $I_{DS} > 0$ . Further increases in the gate voltage,  $V_{GS}$ , above the threshold voltage,  $V_T$ , will cause more electrons to be drawn into the channel. The increase in the electron population in the channel will result in an increase in the conductance of the channel and thus an increase in  $I_{DS}$ .

**Figure 112.18** MOSFET physical model.

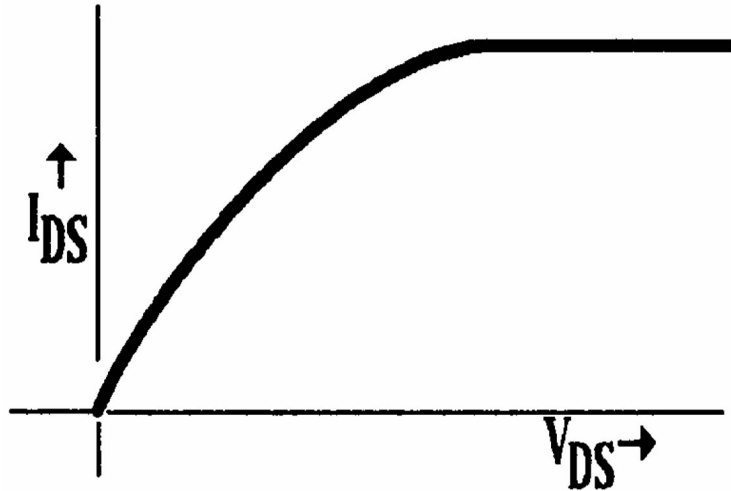


**Figure 112.19**  $V_{GS} > V_T$ .



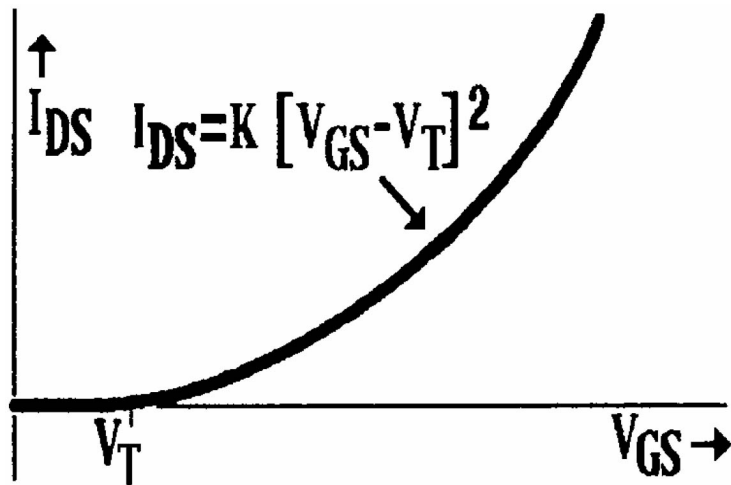
At small values of the drain-to-source voltage,  $V_{DS}$ , the drain current,  $I_{DS}$ , will increase linearly with  $V_{DS}$ , but as  $V_{DS}$  increases the channel will become constricted at the drain end and the curve of  $I_{DS}$  versus  $V_{DS}$  will start to bend over. Finally, if  $V_{DS}$  is large enough such that  $V_{GS} - V_{DS} < V_T$ , the channel becomes pinched off at the drain end and the curve of  $I_{DS}$  versus  $V_{DS}$  will become almost horizontal, as shown in Fig. 112.20. This region where  $I_{DS}$  becomes relatively independent of  $V_{DS}$  is called the *active region*. When the MOSFET is used as an amplifier it should be operated in the active region.

**Figure 112.20**  $I_{DS}$  versus  $V_{DS}$  curve.



In the active region  $I_{DS}$  is relatively independent of  $V_{DS}$  but is a strong function  $V_{GS}$ . The transfer relationship between  $I_{DS}$  and  $V_{GS}$  in the active region is given approximately by the square law equation  $I_{DS} = K(V_{GS} - V_T)^2$ , where  $K$  is a constant. In Fig. 112.21, a graph of the  $I_{DS}$  versus  $V_{GS}$  transfer characteristics of an N-channel MOSFET is shown.

**Figure 112.21** Transfer characteristic.

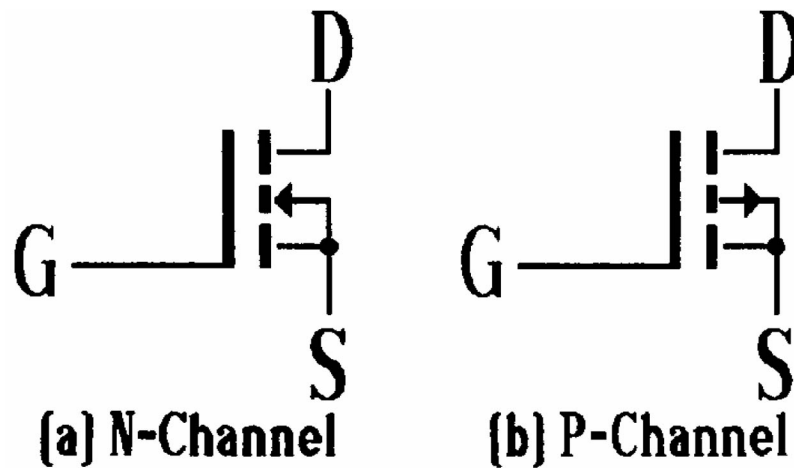


In the active region, under small-signal conditions, the AC component of the drain current is related to the AC component of the gate voltage by  $i_{ds} = g_m \cdot v_{gs}$  where  $g_m$  is the *dynamic transfer conductance*. We have that  $g_m = dI_{DS}/dV_{GS} = 2K(V_{GS} - V_T)$  Since  $K = I_{DS}/(V_{GS} - V_T)^2$  this can be expressed as  $g_m = 2I_{DS}/(V_{GS} - V_T)$  Since  $(V_{GS} - V_T) = \sqrt{I_{DS}/K}$  this can also be rewritten as  $g_m = 2\sqrt{K \cdot I_{DS}}$

## MOSFET as an Amplifier - Small-Signal AC Voltage Gain

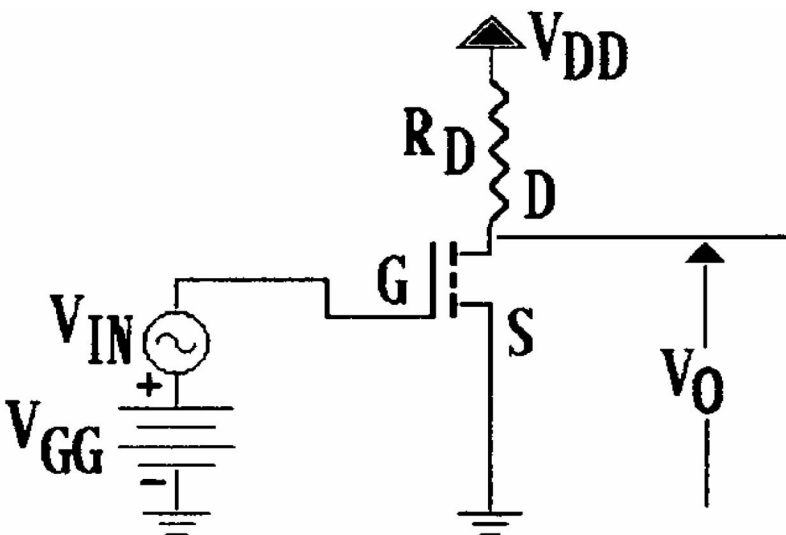
In Fig. 112.22 some MOSFET symbols are shown for both N-channel and P-channel devices.

**Figure 112.22** MOSFET symbols.



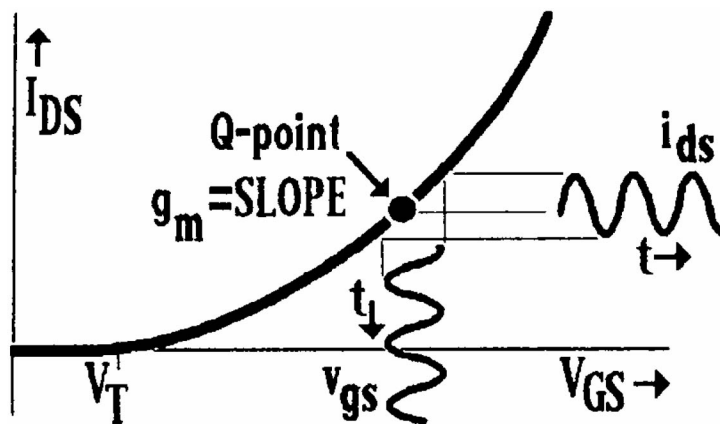
Let's consider the MOSFET common-source amplifier circuit of Fig. 112.23. The input AC signal is applied between the gate and the source as  $v_{gs}$ , and the output AC voltage is taken between the drain and the source as  $v_{ds}$ . Thus the source electrode of this triode device is common to input and output—hence the designation of this MOSFET configuration as common-source (CS) amplifier.

**Figure 112.23** Common source amplifier.



Let us assume a quiescent current level of  $I_{DSQ} = 10 \text{ mA}$  and  $V_{GS} - V_T = 1 \text{ V}$ . The AC component of the drain current,  $i_{ds}$ , is related to the AC component of the gate voltage,  $v_{gs}$ , by  $i_{ds} = g_m \cdot v_{gs} \Omega$ , where  $g_m$  is the *dynamic transfer conductance*, and is given by  $g_m = 2K(V_{GS} - V_T) = 2I_{DS}/(V_{GS} - V_T) = 2 \cdot 10 \text{ mA} / 1 \text{ V} = 20 \text{ mS}$ . If a small AC signal voltage,  $v_{gs}$ , is superimposed on the DC gate bias voltage  $V_{GS}$ , only a small segment of the transfer characteristic adjacent to the Q-point will be traversed, as shown in Fig. 112.24. This small segment will be close to a straight line, and as a result the AC drain current,  $i_{ds} \Omega$ , will have a waveform close to that of the AC voltage applied to the gate. The ratio of  $i_{ds}$  to  $v_{gs}$  will be the slope of the transfer curve, as given by  $i_{ds}/v_{gs} \cong dI_{DS}/dV_{GS} = g_m \Omega$ . Thus,  $i_{ds} \cong g_m \cdot v_{gs}$ . If the net load driven by the drain of the MOSFET is the drain load resistor,  $R_D$ , as shown in Fig. 112.23, then the AC drain current  $i_{ds}$  will produce an AC drain voltage of  $v_{ds} = -i_{ds} \cdot R_D \Omega$ . Since  $i_{ds} = g_m \cdot v_{gs}$ , this becomes  $v_{ds} = -g_m v_{gs} \cdot R_D \Omega$ . The AC small-signal voltage gain from gate to drain thus becomes  $A_V = v_o/v_{in} = v_{ds}/v_{gs} = -g_m \cdot R_D \Omega$ . The negative sign indicates signal inversion as is the case for a common-source amplifier.

**Figure 112.24** Transfer characteristic.



If the DC drain supply voltage is  $V_{DD} = +20 \text{ V}$ , a quiescent drain-to-source voltage of  $V_{DSQ} = V_{DD}/2 = +10 \text{ V}$  will result in the MOSFET being biased in the middle of the active region. Since  $I_{DSQ} = 10 \text{ mA}$  in the example under consideration and the voltage drop across the drain load resistor ( $R_D$ ) is  $10 \text{ V}$ , we obtain  $R_D = 10 \text{ V} / 10 \text{ mA} = 1 \text{ k}\Omega$ . The AC small-signal voltage gain,  $A_V$ , thus becomes  $A_V = -g_m \cdot R_D = -20 \text{ mS} \cdot 1 \text{ k}\Omega = -20$ . Note that the voltage gain is relatively modest, as compared to the much larger voltage gains that can be obtained in a bipolar-junction transistor (BJT) common-emitter amplifier. This is due to the lower transfer conductance of both JFETs and MOSFETs compared to BJTs. For a BJT the transfer conductance is given by  $g_m = I_C/V_T \Omega$ , where  $I_C$  is the quiescent collector current and  $V_T = kT/q \cong 25 \text{ mV}$  is the thermal voltage. At  $I_C = 10 \text{ mA}$ ,  $g_m = 10 \text{ mA} / 25 \text{ mV} = 400 \text{ mS}$ , compared to only  $20 \text{ mS}$  for the MOSFET in this example. With a net load of  $1 \text{ k}\Omega$  the BJT voltage gain will be  $-400 \Omega$  compared to the MOSFET voltage gain of only  $-20 \Omega$ . Thus FETs do have the disadvantage of a much lower transfer conductance and therefore lower voltage gain than BJTs operating under similar quiescent current levels, but they do have the major advantage of a much higher input impedance and a much lower input current. In the case of a MOSFET the input signal is applied to



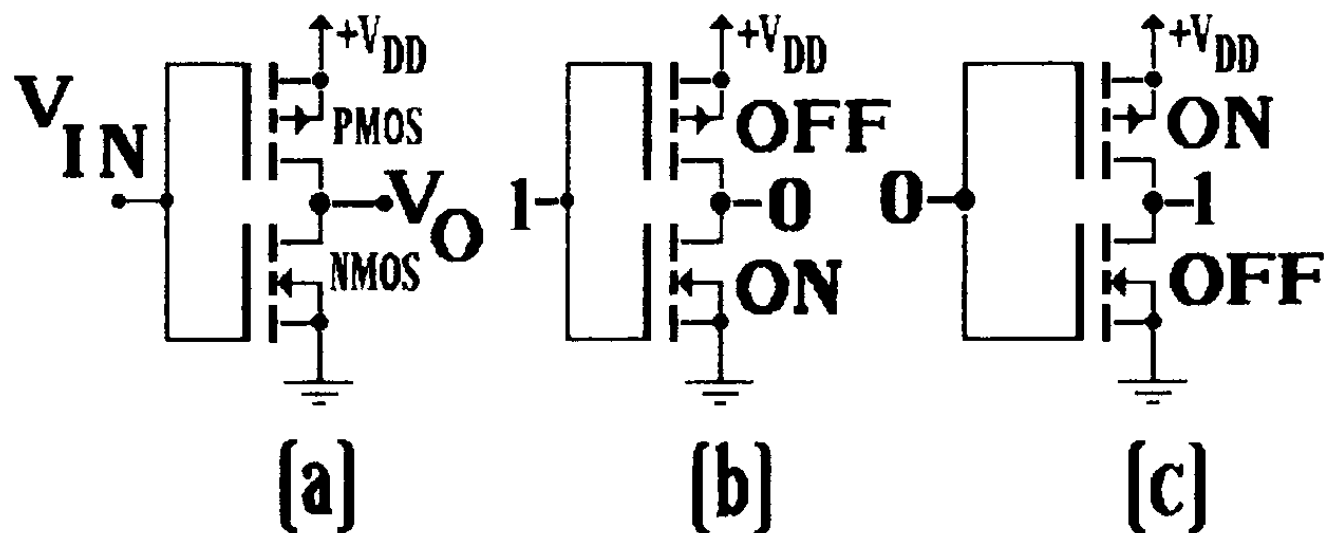
the gate, which is insulated from the rest of the device by the thin  $\text{SiO}_2$  layer, and thus sees a very high impedance. In the case of a common-emitter BJT amplifier, the input signal is applied to the forward-biased base-emitter junction and the input impedance is given approximately by  $r_{\text{IN}} = r_{\text{BE}} \cong 1.5 \cdot \beta \cdot V_T / I_C$ , where  $V_T$  is the thermal voltage of about 25 mV. If  $I_C = 5$  mA and  $\beta = 200$ , for example, then  $r_{\text{IN}} \cong 1500 \Omega$ . This moderate input resistance value of 1.5 k $\Omega$  is certainly no problem if the signal source resistance is less than around 100  $\Omega$ . However, if the source resistance is above 1 k $\Omega$ , there will be a substantial signal loss in the coupling of the signal from the signal source to the base of the transistor. If the source resistance is in the range of above 100 k $\Omega$ , and certainly if it is above 1 M $\Omega$ , there will be severe signal attenuation due to the BJT input impedance, and an FET amplifier will probably offer a greater overall voltage gain. Indeed, when high-impedance signal sources are encountered, a multistage amplifier with an FET input stage followed by cascaded BJT stages is often used.

## MOSFETs for Digital Circuits

MOSFETs are used very extensively for digital applications. For high-density integrated circuits MOSFETs offer great advantages over BJTs and JFETs from the standpoint of a much smaller size and lower power consumption.

A very important MOSFET configuration is the *complementary symmetry MOSFET*, or **CMOS** circuit. In Fig. 112.25, a CMOS inverter circuit is shown, comprising an N-channel MOSFET (NMOS) and a P-channel MOSFET (PMOS). In Fig. 112.25(b), the situation is shown with the input signal in the high or "1" state. The NMOS is now on and exhibits a moderately low resistance, typically on the order of 100  $\Omega$ . The PMOS is off and acts as a very high resistance. This situation results in the output voltage  $V_O$ , going low, close to ground potential (0 V).

**Figure 112.25** CMOS circuit.





In Fig. 112.25(c), the situation is shown with the input signal in the low, or "0" state. The NMOS is off and acts as a very high resistance, and the PMOS is on and exhibits a moderately low resistance, typically on the order of  $100\ \Omega$ . This situation results in the output voltage,  $V_O$ , being pulled up high (the "1" state), close to the  $V_{DD}$  supply.

We note that under both input conditions one of the transistors will be off, and as a result the current flow through the two transistors will be extremely small. Thus the power dissipation will also be very small, and, indeed, the only significant amount of power dissipation in the CMOS pair occurs during the short switching interval when both transistors are simultaneously on.

## Defining Terms

**Active region:** The region of transistor operation in which the output current is relatively independent of the output voltage. For the BJT this corresponds to the condition that the emitter-base junction is on, and the collector-base junction is off. For the FETs this corresponds to the condition that the channel is on, or open, at the source end, and pinched off at the drain end.

**Acceptors:** Impurity atoms that, when added to a semiconductor, contribute holes. In the case of silicon, acceptors are atoms from the third column of the periodic table, such as boron.

**Anode:** The P-type side of a diode.

**Cathode:** The N-type side of a diode.

**CMOS:** The complementary-symmetry MOSFET configuration composed of an N-channel MOSFET and a P-channel MOSFET, operated such that when one transistor is on, the other is off.

**Contact potential:** The internal voltage that exists across a PN junction under thermal equilibrium conditions, when no external bias voltage is applied.

**Donors:** Impurity atoms that, when added to a semiconductor, contribute free electrons. In the case of silicon, donors are atoms from the fifth column of the periodic table, such as phosphorus, arsenic, and antimony.

**Dopants:** Impurity atoms that are added to a semiconductor to modify the electrical conduction characteristics.

**Doped semiconductor:** A semiconductor that has had impurity atoms added to modify the electrical conduction characteristics.

**Extrinsic semiconductor:** A semiconductor that has been doped with impurities to modify the electrical conduction characteristics.

**Forward bias:** A bias voltage applied to the PN junction of a diode or transistor that makes the P-type side positive with respect to the N-type side.

**Forward current:** The large current flow in a diode that results from the application of a forward bias voltage.

**Hole:** An electron vacancy in a covalent bond between two atoms in a semiconductor. Holes are mobile charge carriers with an effective charge that is opposite to the charge on an electron.

**Intrinsic semiconductor:** A semiconductor with a degree of purity such that the electrical characteristics are not significantly affected.

**Majority carriers:** In a semiconductor, the type of charge carrier with the larger population. For example, in an N-type semiconductor, electrons are the majority carriers.

**Minority carriers:** In a semiconductor, the type of charge carrier with the smaller population. For example, in an N-type semiconductor, holes are the minority carriers.

**N-type semiconductor:** A semiconductor that has been doped with donor impurities to produce the condition that the population of free electrons is greater than the population of holes.

**Ohmic, nonsaturated, or triode region:** These three terms all refer to the region of FET operation in which a conducting channel exists all of the way between source and drain. In this region the drain current varies with both the gate voltage and the drain voltage.

**Output characteristics:** The family of curves of output current versus output voltage. For the BJT output characteristics are curves of collector current versus collector voltage for various constant values of base current or voltage and are also called the *collector characteristics*. For FETs these will be curves of drain current versus drain voltage for various constant values of gate voltage and are also called the *drain characteristics*.

**P-type semiconductor:** A semiconductor that has been doped with acceptor impurities to produce the condition that the population of holes is greater than the population of free electrons.

**Pinch-off voltage,  $V_P$  :** The voltage that, when applied across the gate-to-channel PN junction, will cause the conducting channel between drain and source to become pinched off. This is also represented as  $V_{GS}$  (off).

**Reverse bias:** A bias voltage applied to the PN junction of a diode or transistor that makes the P-type side negative with respect to the N-type side.

**Reverse current:** The small current flow in a diode that results from the application of a reverse bias voltage.

**Thermal voltage:** The quantity  $kT/q$  where  $k$  is Boltzmann's constant,  $T$  is absolute temperature, and  $q$  is electron charge. The thermal voltage has units of volts and is a function only of temperature, being approximately 25 mV at room temperature.

**Threshold voltage:** The voltage required to produce a conducting channel between source and drain in a MOSFET.

**Transfer conductance:** The AC or dynamic parameter of a device that is the ratio of the AC output current to the AC input voltage. The transfer conductance is also called the *mutual transconductance* and is usually designated by the symbol  $g_m$ .

**Transfer equation:** The equation that relates the output current (collector or drain current) to the input voltage (base-to-emitter or gate-to-source voltage).

**Triode:** The three-terminal electron device, such as a bipolar junction transistor or a field-effect transistor.

## References

Mauro, R. 1989. *Engineering Electronics*. Prentice Hall, Englewood Cliffs, NJ.

- Millman, J. and Grabel, A. 1987. *Microelectronics*, 2nd ed. McGraw-Hill, New York.
- Mitchell, F. H., Jr. and Mitchell, F. H., Sr. 1992. *Introduction to Electronics Design*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.
- Savant, C. J., Roden, M. S., and Carpenter, G. L. 1991. *Electronic Design*, 2nd ed. Benjamin-Cummings, Menlo Park, CA.
- Sedra, A. S. and Smith, K. C. 1991. *Microelectronics Circuits*, 3rd ed. Saunders, Philadelphia, PA.

Soclof, S. "Analog Integrated Circuits"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Analog Integrated Circuits

---

- 113.1 Operational Amplifiers
- 113.2 Voltage Comparators
- 113.3 Voltage Regulators
- 113.4 Power Amplifiers
- 113.5 Wide-Bandwidth (Video) Amplifiers
- 113.6 Modulators, Demodulators, and Phase Detectors
- 113.7 Voltage-Controlled Oscillators
- 113.8 Waveform Generators
- 113.9 Phase-Locked Loops
- 113.10 Digital-to-Analog and Analog-to-Digital Converters
- 113.11 Radio-Frequency Amplifiers
- 113.12 Integrated Circuit Transducers

Optoelectronic Devices

### Sidney Soclof

*California State University*

An **integrated circuit** is an electronic device in which there is more than one circuit component in the same package. Most integrated circuits contain many transistors, together with diodes, resistors, and capacitors. Integrated circuits may contain tens, hundreds, or even many thousands of transistors. Indeed, some integrated circuits for computer and image-sensing applications may have millions of transistors or diodes on a single silicon chip.

A **monolithic integrated circuit** is one in which all of the components are contained on a single-crystal chip of silicon. This silicon chip typically measures from  $1 \times 1 \times 0.25$  mm thick for the smallest integrated circuits to  $10 \times 10 \times 0.5$  mm for the larger integrated circuits.

A **hybrid integrated circuit** has more than one chip in the package. The chips may be monolithic integrated circuits and separate or "discrete" devices such as transistor or diode chips. There can also be discrete passive components such as capacitor and resistor chips. The chips are usually mounted on an insulating ceramic substrate, usually alumina ( $\text{Al}_2\text{O}_3$ ), and are interconnected by a thin-film or thick-film conductor pattern that has been deposited on the ceramic substrate. Thin-film patterns are deposited by vacuum evaporation techniques and are usually about 1 micrometer in thickness, whereas thick-film patterns are pastes that are printed on the substrate through a screen and usually range in thickness from 10 to 30 micrometers.

Integrated circuits can be classified according to function, the two principal categories being *digital* and *analog* (also called *linear*) integrated circuits. A digital integrated circuit is one in

which all of the transistors operate in the switching mode, being either off (in the cutoff mode of operation) or on (in the saturation mode) to represent the high and low (1 and 0) digital logic levels. The transistors during the switching transient pass very rapidly through the active region. Virtually all digital integrated circuits are of the monolithic type and are composed almost entirely of transistors, usually of the MOSFET type. Some digital integrated circuits contain more than one million transistors on a single silicon chip.

Analog or linear integrated circuits operate on signal voltages and currents that are in analog or continuous form. The transistors operate mostly in the active (or linear) mode of operation. There are many different types of analog integrated circuits, such as operational amplifiers, voltage comparators, audio power amplifiers, voltage regulators, voltage references, video (wide-bandwidth) amplifiers, radio-frequency amplifiers, modulators and demodulators for AM and FM, logarithmic converters, function generators, voltage-controlled oscillators, phase-locked loops, digital-to-analog and analog-to-digital converters, and other devices. The majority of analog integrated circuits are of the monolithic type, although there are many hybrid integrated circuits of importance.

In addition to the two basic functional categories of analog and digital integrated circuits, there are many integrated circuits that have both analog and digital circuitry in the same integrated circuit package, or even on the same chip. Some of the analog integrated circuits mentioned previously, such as the digital-to-analog and analog-to-digital converters, contain both types of circuitry.

Almost all integrated circuits are fabricated from silicon. The principal exceptions are some very high-speed digital integrated circuits that use gallium arsenide (GaAs) to take advantage of the very high electron mobility in that material. Integrated circuits are fabricated using the same basic processes as for other semiconductor devices. In integrated circuits the vast majority of devices are transistors and diodes, with relatively few passive components such as resistors and capacitors. In many cases no capacitors at all are used, and in some cases there are no resistors either. The active devices contained in an integrated circuit are transistors, including bipolar junction transistors (BJTs), junction field-effect transistors (JFETs), and metal-oxide silicon field-effect transistors (MOSFETs). Digital integrated circuits are made up predominantly of MOSFETs, with generally very few other types of components. Some analog integrated circuits use mostly, or even exclusively, BJTs, although many integrated circuits use a mixture of BJTs and field-effect transistors (either JFETs or MOSFETs), and some are even exclusively FETs, with no BJTs at all.

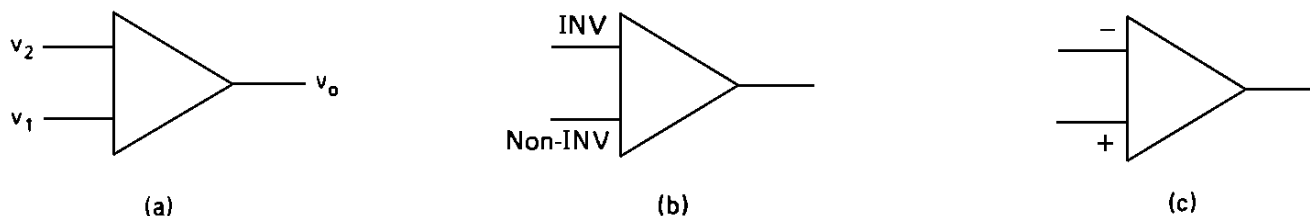
The sections that follow give very brief descriptions of some of the analog integrated circuits.

## 113.1 Operational Amplifiers

---

An operational amplifier is an integrated circuit that produces an output voltage,  $V_O$ , that is an amplified replica of the difference between two input voltages, as given by the equation  $V_O = A_{OL}(V_1 - V_2)$ , where  $A_{OL}$  is called the open-loop gain. The basic symbol for the operational amplifier is shown in Fig. 113.1. Most operational amplifiers are of the monolithic type, and there are hundreds of different types of operational amplifiers available from dozens of different manufacturers.

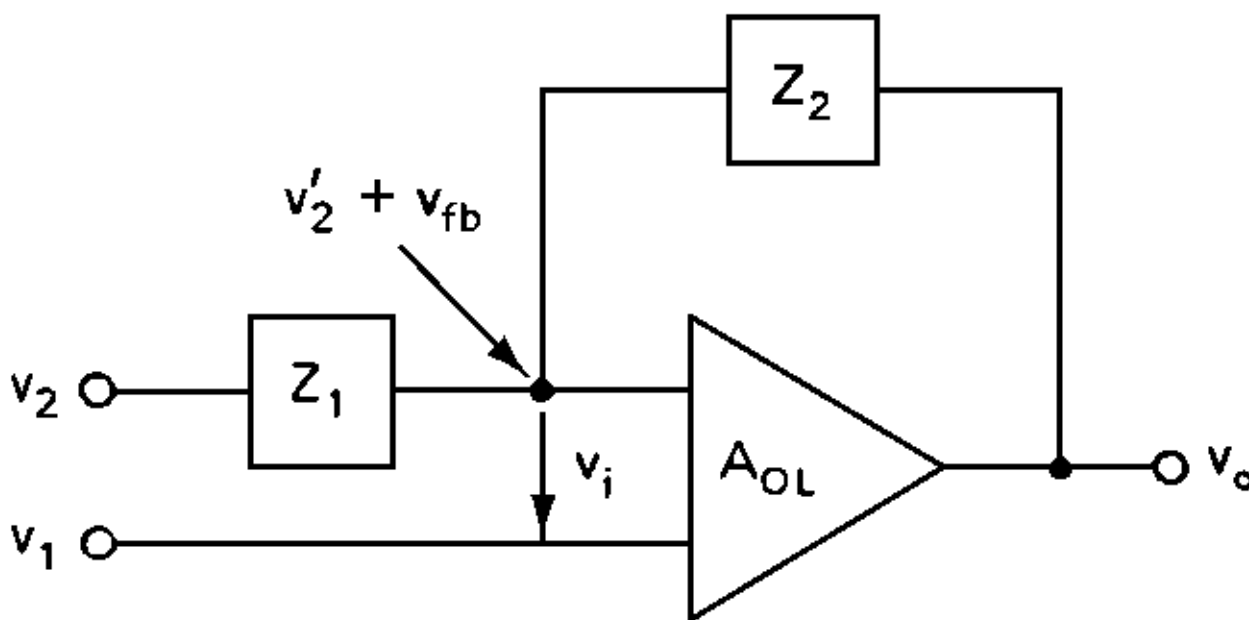
**Figure 113.1** Operational amplifier symbols: (a) basic operational amplifier symbol; (b) symbol with input polarities indicated explicitly; (c) symbol with input polarities indicated explicitly. (Source: Soclof, S. 1991. *Design and Applications of Analog Integrated Circuits*. Prentice Hall, Englewood Cliffs, NJ. With permission.)



The operational amplifier was one of the first types of analog integrated circuits developed; the term *operational amplifier* comes from one of the earliest uses of this type of circuit—in analog computers dating back to the early and middle 1960s. Operational amplifiers were used in conjunction with other circuit components, principally resistors and capacitors, to perform various mathematical operations, such as addition, subtraction, multiplication, integration, and differentiation—hence the name "operational amplifier." The range of applications of operational amplifiers has vastly expanded since these early beginnings; operational amplifiers are now used to perform a multitude of tasks through the entire field of electronics.

Operational amplifiers are usually used in a feedback, or closed-loop, configuration, as shown in Fig. 113.2. Under the assumption of a large open-loop gain,  $A_{OL}$ , the output voltage is given by  $V_o = V_1[1 + (Z_2/Z_1)] - V_2(Z_2/Z_1)$ .

**Figure 113.2** Closed-loop (negative feedback) operational-amplifier system. (Source: Soclof, S. 1991. *Design and Applications of Analog Integrated Circuits*. Prentice Hall, Englewood Cliffs, NJ. With permission.)



The following is a list of some important applications of operational amplifiers:

*Difference amplifier.* This produces an output voltage proportional to the difference of two input voltages.

*Summing amplifier.* This produces an output voltage that is a weighted summation of a number of input voltages.

*Current-to-voltage converter.* This produces an output voltage that is proportional to an input current.

*Voltage-to-current converter.* This produces an output current that is proportional to an input voltage but is independent of the load being driven.

*Active filters.* This is a very broad category of operational amplifier circuit that can be configured as low-pass, high-pass, band-pass, or band-stop filters.

*Precision rectifiers and clipping circuits.* This is a broad category of wave-shaping circuits that can be used to clip off or remove various portions of a waveform.

*Peak detectors.* This produces an output voltage proportional to the positive or negative peak value of an input voltage.

*Logarithmic converters.* This produces an output voltage proportional to the logarithm of an input voltage.

*Exponential or antilogarithmic converters.* This produces an output voltage that is an exponential function of an input voltage.

*Current integrator or charge amplifier.* This produces an output voltage proportional to net flow of charge in a circuit.

*Voltage regulators.* This produces an output voltage that is regulated to remain relatively constant with respect to changes in the input or supply voltage and with respect to changes in the output or load current.

*Constant current sources.* This produces an output current that is regulated to remain relatively constant with respect to changes in the input or supply voltage and with respect to changes in the output or load impedance or voltage.

*Amplifiers with electronic gain control.* These are amplifiers in which the gain can be controlled over a wide range by the application of an external voltage.

*Function generators.* These are circuits that can be used to generate various types of waveforms, including square waves and triangular waves.

*Clamping circuits.* These circuits produce an output voltage that has the same AC waveform as the input signal, but the DC level is shifted by an amount controlled by a fixed reference voltage.

*Analog signal multiplexer.* This circuit combines several input signals for transmission over a single communications link by means of *time-domain multiplexing*.

*Sample-and-hold circuit.* The input signal is sampled over a short period of time, and the sampled value is then held at that value until the next sample is taken.

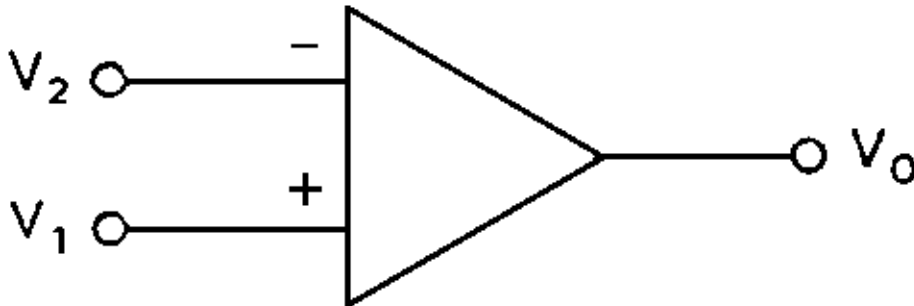
*Analog multiplier.* This produces an output voltage proportional to the product of two input voltages.



## 113.2 Voltage Comparators

A voltage comparator is an integrated circuit as shown in Fig. 113.3 that is used to compare two input voltages and produce an output voltage that is in the high (or "1") stage if  $V_1 > V_2$  and in the low (or "0") stage if  $V_1 < V_2$ . It is essentially a one-bit analog-to-digital converter.

**Figure 113.3** Voltage comparator. (Source: Soclof, S. 1991. *Design and Applications of Analog Integrated Circuits*. Prentice Hall, Englewood Cliffs, NJ. With permission.)



In many respects, voltage comparators are similar to operational amplifiers, and, indeed, operational amplifiers can be used as voltage comparators. A voltage comparator is, however, designed specifically to be operated under open-loop conditions, basically as a switching device. An operational amplifier, on the other hand, is almost always used in a closed-loop configuration and is usually operated as a linear amplifier.

Being designed to be used in a closed-loop configuration, the frequency response characteristics of an operational amplifier are generally designed to ensure an adequate measure of stability against an oscillatory type of response. This results in a sacrifice being made in the bandwidth, rise time, and slewing rate of the device. In contrast, since a voltage comparator operates as an open-loop device, no sacrifices have to be made in the frequency response characteristics, so a very fast response time can be obtained.

An operational amplifier is designed to produce a zero output voltage when the difference between the two input signals is zero. A voltage comparator, in contrast, operates between two fixed output voltage levels, so the output voltage is either in the high or low states.

The output voltage of an operational amplifier will saturate at levels that are generally about 1 or 2 V away from the positive and negative power supply voltage levels. The voltage comparator output is often designed to provide some degree of flexibility in fixing the high- and low-state output voltage levels and for ease in interfacing with digital logic circuits.

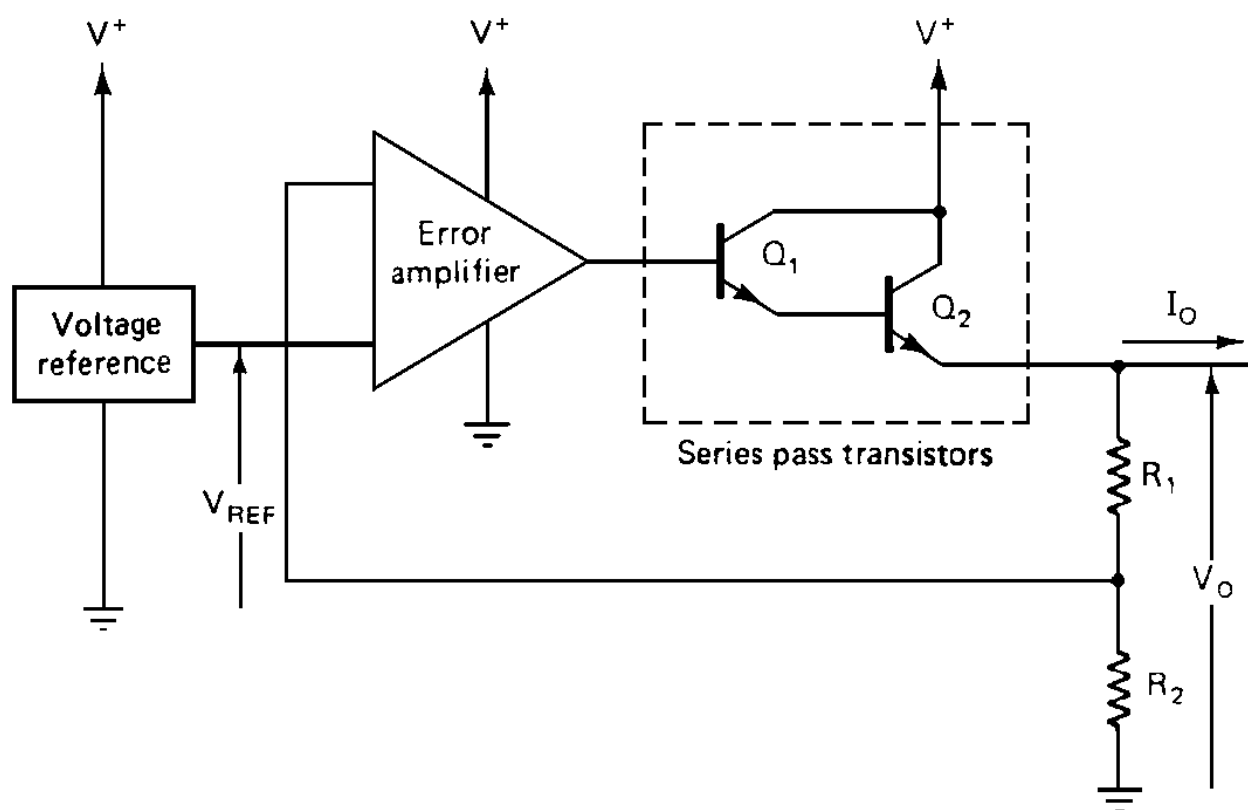
There are many applications of voltage comparators. These include pulse generators, square-wave and triangular-wave generators, pulse-width modulators, level and zero-crossing detectors, pulse regenerators, line receivers, limit comparators, voltage-controlled oscillators, analog-to-digital converters, and time-delay generators.

## 113.3 Voltage Regulators

A voltage regulator is an electronic device that supplies a constant voltage to a circuit or load. The output voltage of the voltage regulator is regulated by the internal circuitry of the device to be relatively independent of the current drawn by the load, the supply or line voltage, and the ambient temperature. A voltage regulator may be part of some larger electronic circuit but is often a separate unit or module, usually in the form of an integrated circuit. A voltage regulator, as shown in Fig. 113.4, is composed of three basic parts:

1. A voltage reference circuit that produces a reference voltage that is independent of the temperature and supply voltage
2. An amplifier to compare the reference voltage with the fraction of the output that is fed back from the voltage regulator output to the inverting input terminal of the amplifier
3. A series-pass transistor or combination of transistors to provide an adequate level of output current to the load being driven

**Figure 113.4** Voltage regulator: basic block diagram. (Source: Soclof, S. 1991. *Design and Applications of Analog Integrated Circuits*. Prentice Hall, Englewood Cliffs, NJ. With permission.)



Voltage regulators usually include protection circuitry such as current limiting and thermal limiting to protect the integrated circuit against overheating and possible damage.

An important type of voltage regulator is the *switching-mode regulator*, in which the series-pass transistors are not on continuously but, rather, are rapidly switched from being completely on to completely off. The output voltage level is controlled by the fraction of time that the series-pass transistors are on (i.e., the duty cycle). Switching-mode regulators are characterized by having very high efficiencies, often above 90%.

## 113.4 Power Amplifiers

---

Although most integrated circuit amplifiers can deliver only small amounts of power to a load, generally well under 1 watt, there are integrated circuits that are capable of supplying much larger amounts of power, up in the range of several watts, or even several tens of watts. There are a variety of integrated circuit audio power amplifiers available that are used in the range of frequencies up to about 10 or 20 kHz for amplification of audio signal for delivery to loudspeakers. These integrated circuit audio power amplifiers also are used for other applications, such as relay drivers and motor controllers.

There are also available a variety of power operational amplifiers. The maximum current available from most operational amplifiers is generally in the range of about 20 to 25 mA. The maximum supply voltage rating is usually around 36 V with a single supply or +18 V and −18 V when a split supply is used. For an operational amplifier with a 36 V total supply voltage, the maximum peak-to-peak output voltage swing available will be around 30 V. With a maximum output current rating of 20 mA, the maximum AC power that can be delivered to a load is  $P_L = 30 \text{ V} \times 20 \text{ mA} / 4 = 150 \text{ mW}$ . Although this level of output power is satisfactory for many applications, a considerably larger power output is required for some applications.

Larger AC power outputs from operational amplifiers can be obtained by adding external *current boost* power transistors that are driven by the operational amplifier to the circuit. There are also available *power operational amplifiers* that are capable of operation with supply voltages as high as 200 V, and with peak output current swings as large as 20 A. With the proper heat sinking for efficient transfer of heat from the integrated circuit to the ambient, the power operational amplifiers can deliver output power up in the range of tens of watts to a load.

### HOW A CHIP IS MADE

Semiconductor chips are made from silicon, an element found in ordinary beach sand. Silicon's conductive properties allow for the creation of on/off switches that equal a one or zero in a computer's binary language.

Through hundreds of complex processing steps, the switches are built and connected into circuits, millions of which can be placed on a single chip.

The exact nature and number of steps needed to manufacture a semiconductor chip varies with its design and complexity, but the basic process remains the same.

Ultra-pure silicon is processed into cylinders that are sliced into thin, 5- to 8-inch diameter wafers on which hundreds of individual computer chips can be made. The wafers are cleaned, inspected, and placed in high temperature furnaces where they are coated with a non-conducting oxide film.

A thin layer of light-sensitive plastic, called photoresist, is applied over the oxide. A glass "mask," containing the chip's circuit pattern, is placed over the wafer and precisely aligned. In a process called photolithography, which is similar to developing a photograph from a negative, light or an X-ray beam is projected through the mask to print each chip's circuit pattern on the wafer surface.

The unexposed photoresist is washed away in solvent baths, etching the protective oxide layer with the shape of the circuit pattern. Holes are also etched in the protective oxide layer.

The wafer is then bombarded with ions, or charged particles, that penetrate the holes etched in the oxide surface. The depth and concentration of these materials determine the specific electrical characteristics of the chip. The process of oxidation, photolithography, etching, and ion implanting are repeated to build transistors and other electronic circuitry that make up each chip.

Once the electronic components have been implanted in the silicon, interconnecting wiring is added to the chip by placing the wafer in a vacuum chamber and coating it with copper mixed with aluminum or other metals. The aluminum is etched away, leaving the desired wiring.

A thin layer of material is added to protect the wafer. The wafers are then cut into individual chips by diamond-bladed saws and mounted in metal or plastic packages, called modules. These modules are tested and plugged into printed circuit boards, which are eventually built into finished computers.

(Courtesy of IBM Microelectronics Division.)

## 113.5 Wide-Bandwidth (Video) Amplifiers

Video, or wide-bandwidth, amplifiers are designed to give a relatively flat gain versus frequency response characteristic over the frequency range that is generally required to transmit video information. This frequency range is from low frequencies, generally around 30 Hz, up to several megahertz. For standard television reception, the bandwidth required is around 4 MHz, but for other video display applications the bandwidth requirement may be as high as 20 MHz, and for some applications up in the range of 50 MHz.

In contrast, the bandwidths required for audio applications extend only over the frequency range corresponding to the range of the human ear—around 50 Hz to 15 kHz.

The principal technique that is used to obtain the large bandwidths that are required for video amplifiers is the trading off of reduced gain in each amplifier stage for increased bandwidth. This trade-off is accomplished by the use of reduced load resistances for the various gain stages of the amplifier and by the use of negative feedback. In many video amplifiers both techniques are employed. The reduction in the gain of the individual stages can be compensated for by adding additional gain stages.

Included in the category of video amplifiers are the very wide-bandwidth operational amplifiers. Most operational amplifiers are limited to a bandwidth of around 1 to 10 MHz, but there are wide-bandwidth operational amplifiers available that can be used up in the range of 100 to 200 MHz.

## 113.6 Modulators, Demodulators, and Phase Detectors

---

This is a category of integrated circuit that can be used to produce amplitude-modulated (AM) and frequency-modulated (FM) signals. These same integrated circuits can also be used for the demodulation, or detection, of AM and FM signals. In the AM case these integrated circuits can be used to generate and to demodulate *double-sideband/suppressed carrier* (DSB/SC) and *single-sideband/suppressed carrier* (SSB/SC) signals. Another application of this type of integrated circuit is as a *phase detector*, in which an output voltage proportional to the phase difference of two input signals is produced.

## 113.7 Voltage-Controlled Oscillators

---

A voltage-controlled oscillator (VCO) is an oscillator circuit in which the frequency of oscillation can be controlled by an externally applied voltage. VCOs are generally designed to operate over a wide frequency range, often with a frequency ratio of 100:1. One important feature that is often required for VCOs is a linear relationship between the oscillation frequency and the control voltage. Many VCOs have a maximum frequency of operation of around 1 MHz, but there are some emitter-coupled VCOs that can operate up to 50 MHz.

## 113.8 Waveform Generators

---

A waveform generator is an integrated circuit that generates the following three types of voltage waveforms: square waves, triangular waves, and sinusoidal waves. The square and triangular waves can be generated by the same type of circuits as used for VCOs.

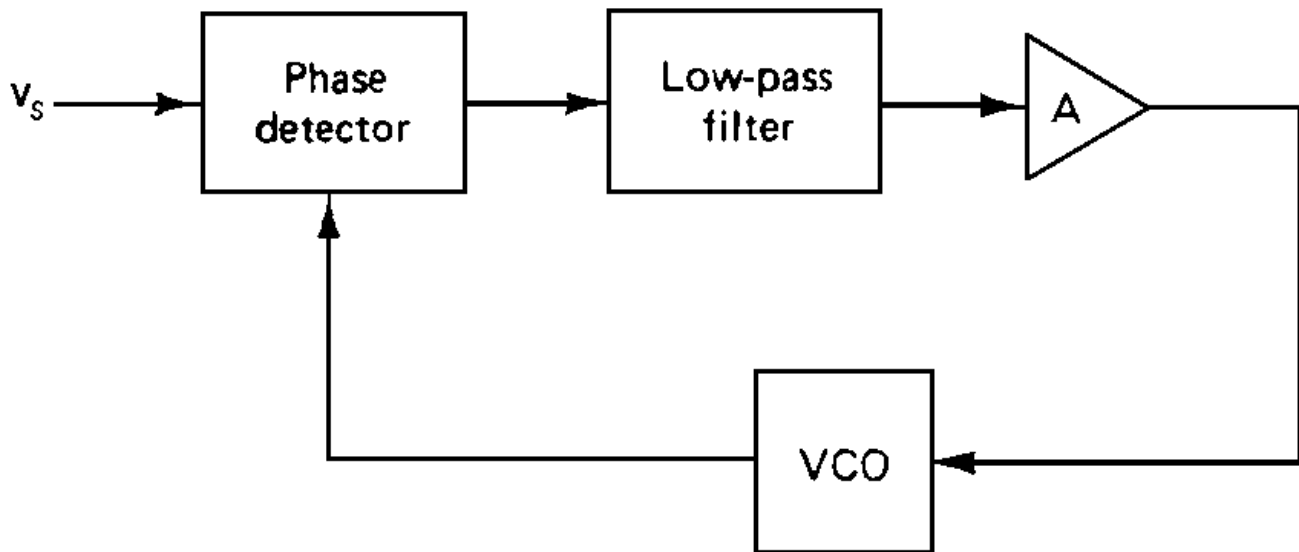
For the generation of a sinusoidal waveform a feedback amplifier using an LC tuned circuit or an RC phase shift network in the feedback loop can be used. These feedback oscillators can produce a very low-distortion sine wave, but it is difficult to modulate the oscillation frequency over a very wide range by means of a control voltage. The VCO, on the other hand, is capable of a frequency sweep ratio as large as 100:1, with very good linearity between the frequency and the control voltage. A sinusoidal waveform can be obtained from a VCO by using a waveshaping network to convert the triangular wave output to a sine wave.

## 113.9 Phase-Locked Loops

---

A phase-locked loop (PLL) is a feedback loop comprising a phase detector, low-pass filter, and a voltage-controlled oscillator (VCO) as shown in [Fig. 113.5](#). When the PLL has locked in on a signal, the frequency of the VCO will exactly follow the signal frequency.

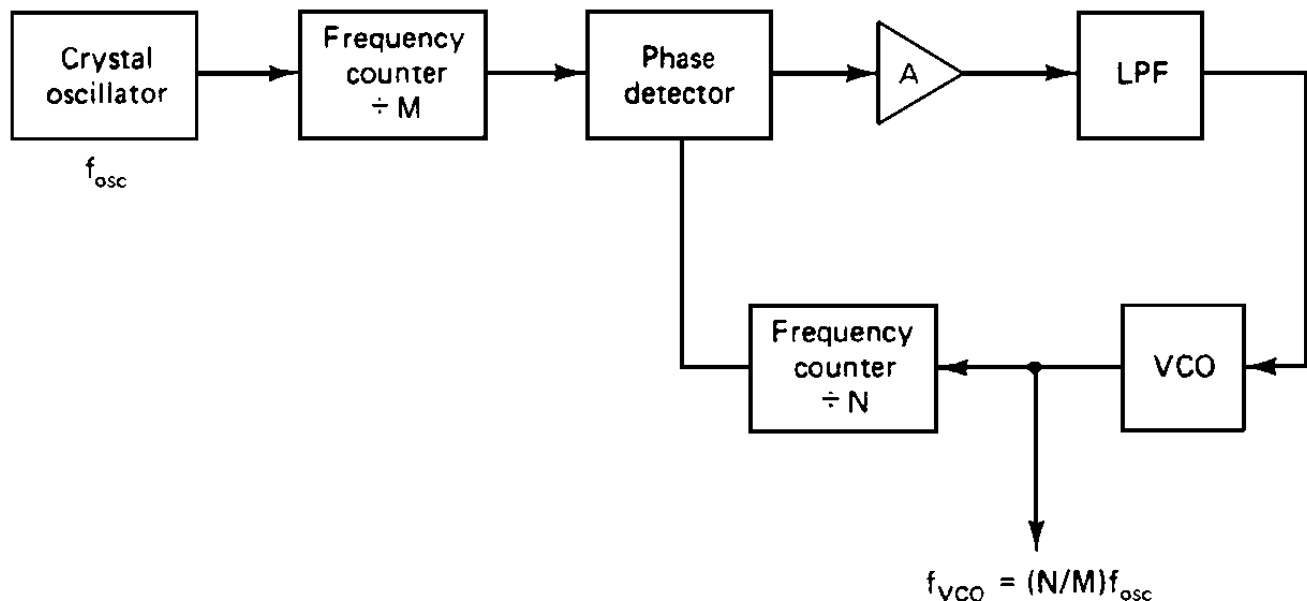
**Figure 113.5** Phase-locked loop. (Source: Soclof, S. 1991. *Design and Applications of Analog Integrated Circuits*. Prentice Hall, Englewood Cliffs, NJ. With permission.)



The PLL can be used as an FM demodulator or detector. In this case the VCO control voltage is proportional to the frequency deviation of the FM signal and represents the demodulated output voltage. Another closely related application is the demodulation of frequency-shift keying (FSK) signals, a process that is similar to FM except that the signal frequency is shifted between just two values.

One important application of PLLs is in frequency synthesis, in which a precise series of frequencies is produced, all derived from a stable crystal-controlled oscillator. In Fig. 113.6, a PLL frequency synthesizer circuit is shown.

**Figure 113.6** PLL frequency synthesizer. (Source: Soclof, S. 1991. *Design and Applications of Analog Integrated Circuits*. Prentice Hall, Englewood Cliffs, NJ. With permission.)



## **113.10 Digital-to-Analog and Analog-to-Digital Converters**

---

A digital-to-analog converter (D/A or DAC) is an integrated circuit that converts a digital input signal to an analog output voltage (or current) that is proportional to the digital signal. DACs vary in resolution from 4 to 16 bits.

An analog-to-digital converter (A/D or ADC) is an integrated circuit that converts an analog input signal to a digital output. ADCs vary in resolution from a simple voltage comparator used as a 1-bit ADC to 16-bit ADCs.

For some applications, such as storing entire frames of video information in one-thirtieth of a second, very high conversion rates are required. For these applications parallel comparator (or "flash") ADCs are used with conversion rates as high as 500 MHz for an 8-bit ADC.

## **113.11 Radio-Frequency Amplifiers**

---

There are radio-frequency (R-F) integrated circuits that use tuned circuits and operate as band-pass amplifiers. These are used in communications circuits, such as AM and FM radio, and in television for signal amplification and mixing.

## **113.12 Integrated Circuit Transducers**

---

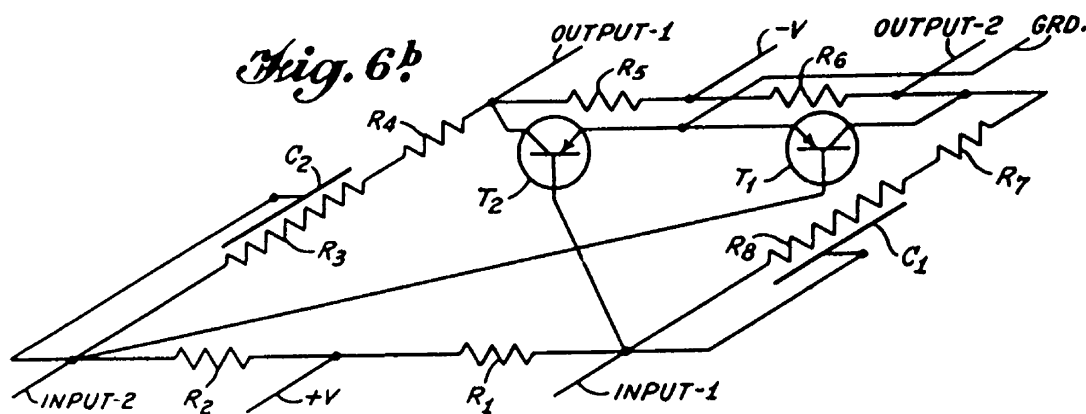
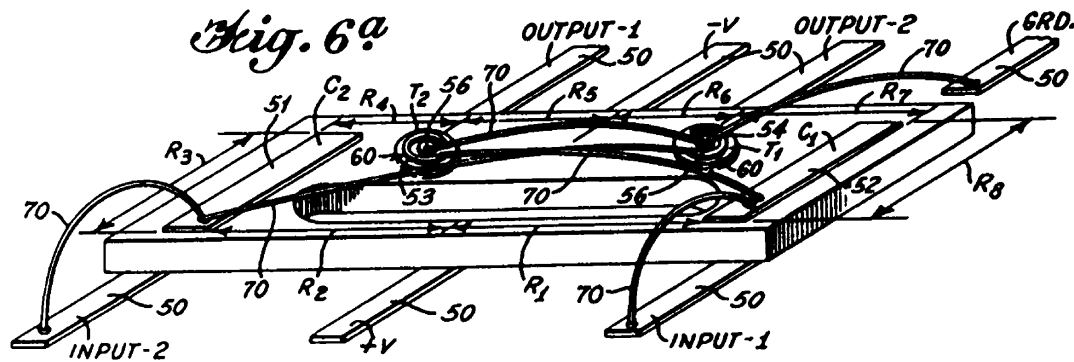
This is a broad category of integrated circuits that are used for the conversion of various physical inputs to an electrical signal. These circuits include the magnetic field sensor, based on the Hall effect, which produces an output voltage proportional to the magnetic field strength. There are also temperature sensor integrated circuits that can produce a voltage or current output that is proportional to the temperature. There are electromechanical integrated circuit transducers such as pressure sensors that produce an output voltage proportional to pressure. Miniature solid-state accelerometers based on integrated circuit sensors are also available.

## **Optoelectronic Devices**

A very important category of integrated circuit transducers is that of the optoelectronic devices. These devices range from photodiode or phototransistor-amplifier modules to image sensors containing in the range of one million individual photodiode image-sensing elements. These image-sensing integrated circuit chips also include additional circuitry to properly transfer out (line by line, in a serial output) the information from the two-dimensional array of the image sensor.

There are also integrated circuits used with light emitting diodes (LEDs) and laser diodes for the generation of light pulses for optoelectronic communications systems, including fiber optic systems.

In the case of a fiber optic communications system, a small diameter glass fiber is used as a conduit or waveguide to guide a beam of light from transmitter to receiver. Optoelectronic integrated circuits are used at the transmitting end with LEDs or laser diodes for the generation and emission of the optical signal. Optoelectronic integrated circuits are used at the receiving end with photodiodes or phototransistors for the detection, amplification, and processing of the received signal.



## MINIATURIZED ELECTRONIC CIRCUITS

Jack S. Kilby

Patented June 23, 1964

#3,138,743

An excerpt:

In contrast to the approaches to miniaturization that have been made in the past, the present invention has resulted from a new and totally different concept for miniaturization. Radically departing from the teachings of the art, it is proposed by the invention that miniaturization can best be attained by use of as few materials and operations as possible. In accordance with the principles of the invention, the ultimate in circuit miniaturization is attained by using only one material for all circuit elements and a limited number of compatible process steps for the production thereof.

Kilby patented what we now call the Integrated Circuit. He made whole circuits (with transistors, resistors, and capacitors) using a single semiconductor substrate and diffusing opposing materials onto it to form the circuit elements. This fundamental process has advanced to the point that now millions of transistors are put on a single "chip." (© 1992, DewRay Products, Inc. Used with permission.)



## Defining Terms

**Hybrid integrated circuit:** An electronic circuit package that contains more than one chip. These chips can be a mixture of monolithic ICs, diodes, transistors, capacitors, and resistors.

**Integrated circuit:** An electronic circuit package that contains more than one circuit element.

**Monolithic integrated circuit:** A single-crystal chip of a semiconductor, generally silicon, that contains a complete electronic circuit.

## References

Franco, S. 1988. *Design with Operational Amplifiers and Analog Integrated Circuits*. McGraw-Hill, New York.

Gray, P. R. and Meyer, R. G. 1992. *Analysis and Design of Analog Integrated Circuits*. John Wiley & Sons, New York.

Irvine, R. G. 1987. *Operational Amplifiers<sup>3/4</sup> Characteristics and Applications*. Prentice Hall, Englewood Cliffs, NJ.

Kennedy, E. J. 1988. *Operational Amplifier Circuits*. Holt, Rinehart and Winston, New York.

McMenamin, J. M. 1985. *Linear Integrated Circuits, Operation and Applications*. Prentice Hall, Englewood Cliffs, NJ.

Sedra, A. S. and Smith, K. C. 1982. *Microelectronic Circuits*. Holt, Rinehart and Winston, New York.

Seippel, R. G. 1983. *Operational Amplifiers*. Prentice Hall, Englewood Cliffs, NJ.

Soclof, S. 1991. *Design and Applications of Analog Integrated Circuits*. Prentice Hall, Englewood Cliffs, NJ.

Bhattacharya, P. "Optoelectronic Devices"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## 114.1 Light-Emitting Diodes

Device Structure and Efficiency • Heterojunction and Edge-Emitting LEDs

## 114.2 Lasers

Lasing Condition and Gain • Threshold Condition • Operating Principles • Power Output

## 114.3 Photodetectors

Quantum Efficiency and Responsivity • Photoconductor • PIN Photodiode • Heterojunction and Avalanche Photodiodes

## 114.4 Conclusion

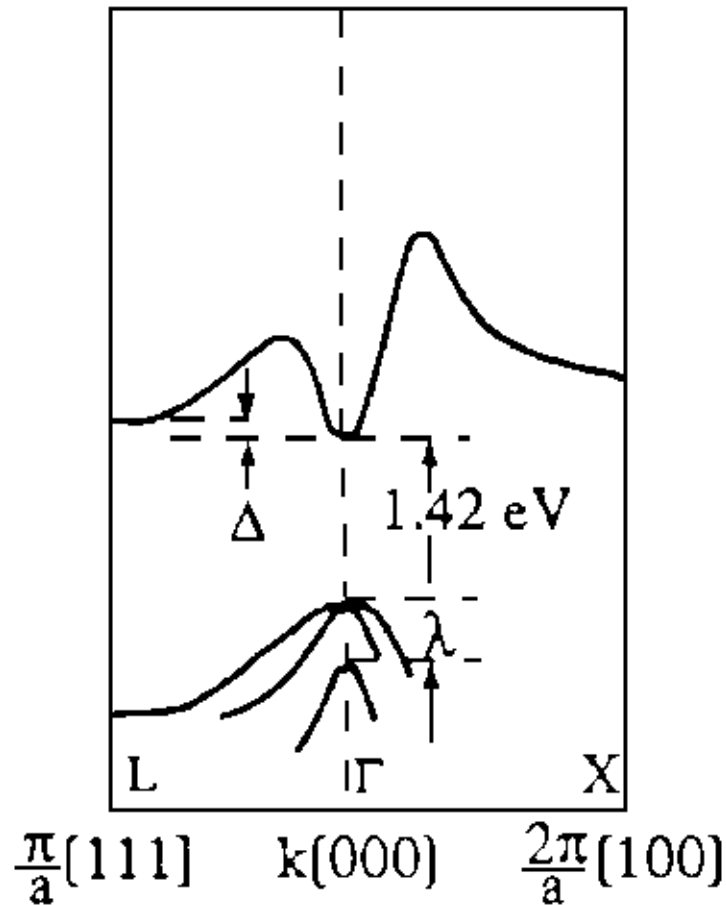
### Pallab Bhattacharya

*University of Michigan*

The phenomenal development of optoelectronics has been spurred by the needs of light-wave communication systems, optoelectronic computing, instrumentation, and alternate energy systems. The subject of optoelectronics deals with the interaction of light and optical processes with electronic processes. Devices in which such interactions take place—usually accompanied by an energy conversion process—are called *optoelectronic devices*. Examples are light-emitting diodes, lasers, and photodiodes.

The development of optoelectronic devices has gone hand in hand with the research and development of compound semiconductors. In compound semiconductors such as GaAs, InP, ZnS, and CdTe the **band gap** is *direct* and hence the quantum efficiency is very high. Therefore, these binary compounds and their ternary and quaternary derivatives are extensively used for the design and fabrication of high-performance optoelectronic devices. The detailed band structure of GaAs is shown in [Fig. 114.1](#).

**Figure 114.1** The band structure of gallium arsenide.



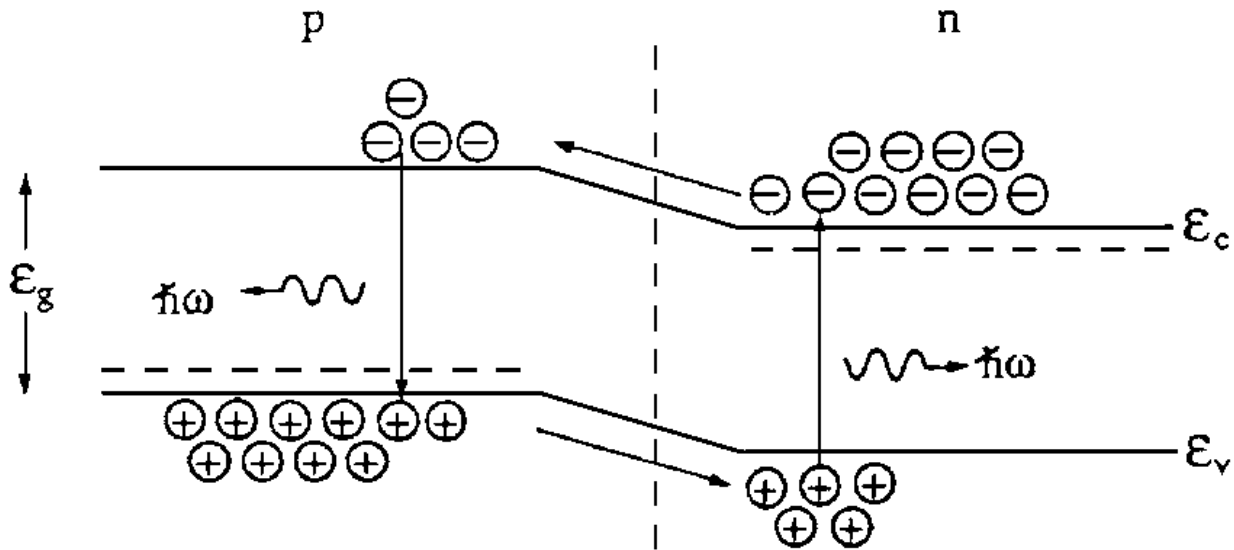
## 114.1 Light-Emitting Diodes

The light-emitting diode (LED) is a device that is used extensively for display and sensing applications and can emit light in the visible and infrared regions of the spectrum. A semiconductor LED is usually a forward-biased p-n junction. An LED is characterized by simpler design and fabrication procedures, lower cost, and simpler drive circuitry than a laser.

In a junction LED, photons of near-band-gap energy are generated by the process of *electroluminescence*, in which a large density of electrons injected into a normally empty conduction band by forward bias recombine with holes in the valence band to produce photons of approximately band-gap energy. The photons have random phases and therefore the LED is an

incoherent light source. The spectral line width of this emission is typically 20–50 nm, depending on the materials used. The process of injection luminescence is illustrated in Fig. 114.2.

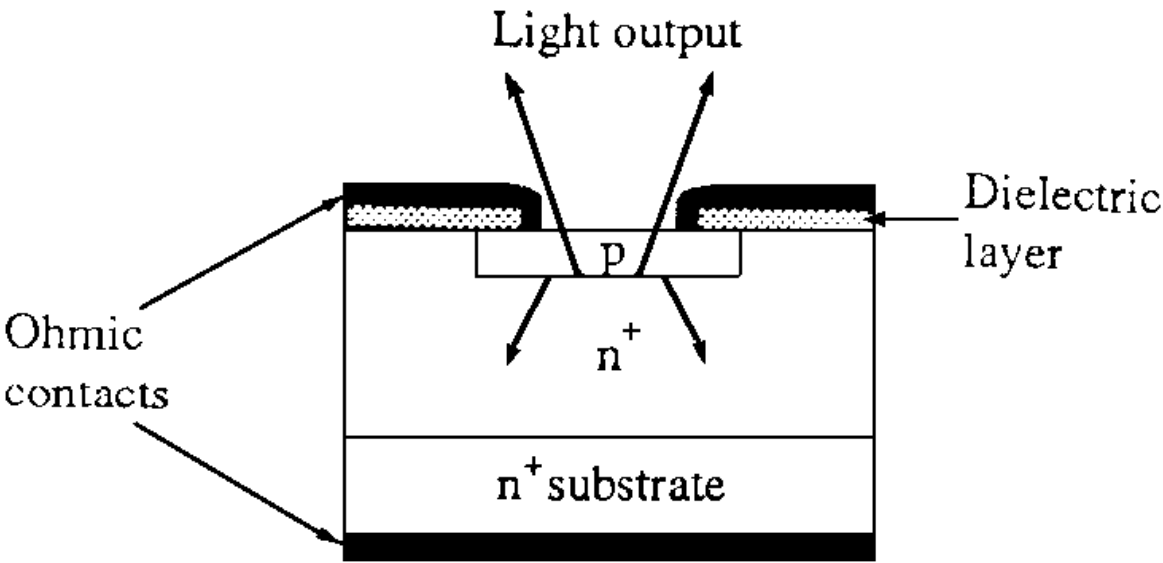
**Figure 114.2** Injection of minority carriers in a forward-biased p-n junction leading to spontaneous emission of photons.



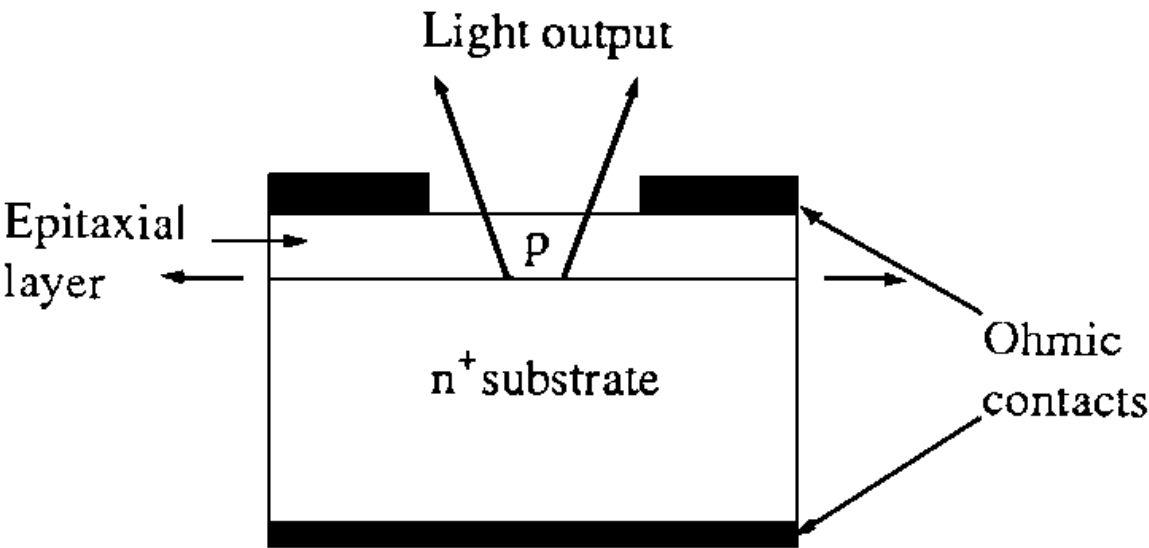
## Device Structure and Efficiency

The schematic of a typical surface-emitting LED is shown in Fig. 114.3. The p-n junction is formed by diffusion or epitaxy. The junction is placed close to the surface to reduce reabsorption of the emitted light. Also, the junction is configured such that most of the recombination takes place in the top layer. This condition is ensured by having the majority of the current flowing across the diode be accounted for by the motion of those carriers that are injected into the top layer from the bottom layer.

**Figure 114.3** Planar surface-emitting LED structure made by (a) diffusion and (b) epitaxial growth.



(a)



(b)

The overall efficiency,  $\eta_o$ , of an LED is given by

$$\eta_o = \eta_{in} \eta_r \eta_e \quad (114.1)$$

where  $\eta_{in}$ ,  $\eta_r$ , and  $\eta_e$  are, respectively, the injection, radiative recombination, and extraction efficiencies.

For electron injection in an  $n^+$ -p diode,

$$\eta_{in} = \left( 1 + \frac{\mu_h N_A L_e}{\mu_e N_D L_h} \right)^{-1} \quad (114.2)$$

where  $N_A$  and  $N_D$  are the acceptor and donor doping levels,  $\mu_e$  and  $\mu_h$  are the electron and hole mobilities, and  $L_e$  and  $L_h$  are the respective diffusion lengths.

The recombination efficiency  $\eta_r$  is the same as the internal quantum efficiency or radiative efficiency, and

$$\eta_r = \frac{R_r}{R_r + R_{nr}} \quad (114.3)$$

where  $R_r$  and  $R_{nr}$  are the radiative and nonradiative recombination rates, respectively. The extraction efficiency relates to the amount of light that can come out of the device and is characterized by a transmission factor,  $F_T$ , given by

$$F_T = \frac{1}{4} \left( \frac{n_{r2}}{n_{r1}} \right)^2 \left[ 1 - \left( \frac{n_{r1} - n_{r2}}{n_{r1} + n_{r2}} \right)^2 \right] \quad (114.4)$$

where  $n_{r2}$  is the refractive index of the semiconductor, and  $n_{r1} = 1$  (air). In standard practice  $F_T$  is increased by covering the active device with a material (usually plastic) with  $n_r > 1$  and making this encasing in the shape of a dome for maximum extraction of light.

The *responsivity* of an LED is defined as the ratio of the emitted optical power  $P_o$  to the injection current. In other words

$$R = \frac{P_o}{I} = \eta_o \frac{h\nu}{q} = \frac{1.24\eta_o}{\lambda(\mu m)} \text{ (W/A)} \quad (114.5)$$

where  $h\nu$  is the photon energy and the emitted optical power is given by

$$P_o = \frac{h\nu V_\ell}{\sqrt{2} \hbar^3 \tau_r} \left( \frac{m_r^*}{\pi} \right)^{3/2} (kT)^{3/2} \exp \left( \frac{\Delta E_f - E_g}{kT} \right) \text{ (W)} \quad (114.6)$$

where  $V_\ell$  is the volume of the active region,  $\hbar (= h/2\pi)$  is the reduced Planck constant,  $k$  is the

Boltzmann constant,  $\tau_r$  is the recombination lifetime, and  $m_r^*$  is the reduced effective mass, given by

$$m_r^* = \frac{m_e^+ m_h^*}{m_e^* + m_h^*} \quad (114.7)$$

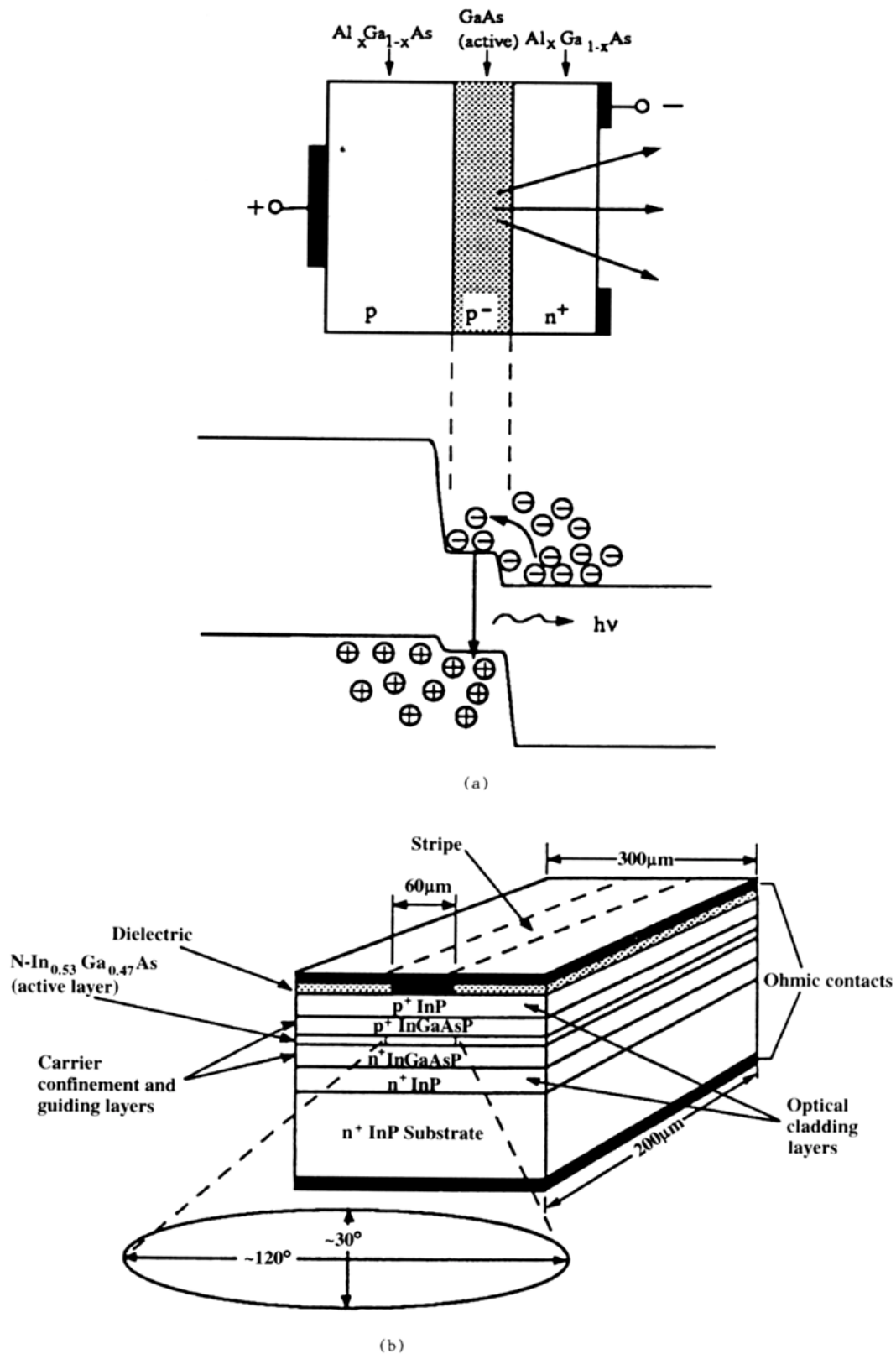
$\Delta E_f$  is the energy difference of the electron and hole quasi-Fermi levels and  $E_g$  is the band gap. The terms  $m_e^*$  and  $m_h^*$  represent the **effective mass** of electrons and holes, respectively.

## Heterojunction and Edge-Emitting LEDs

There are two problems with the simple LED structure described in the previous section. First, the recombination process can be dominated by surface states if the junction is too close to the surface. Second, if the junction is placed far from the surface, not only does the reabsorption of the emitted light increase, but the recombination volume is ill defined by the diffusion length of the injected carriers. Use of a **heterojunction** alleviates both these problems. This device is illustrated in [Fig. 114.4\(a\)](#). First, the density of defect states at a **lattice-matched** heterojunction can be far less than the surface state density, and, second, the heterojunction defines the active volume by confining the injected carriers, and therefore the recombination, in the low-band-gap active layer.



**Figure 114.4** (a) Device structure and band diagram of a forward-biased double heterostructure (DH) LED. (b) Schematic illustration of the structure of a stripe geometry DH InGaAs/InGaAs/InP edge-emitting LED.



A typical edge-emitting LED is illustrated in Fig. 114.4(b). Here the heterojunction serves the dual purpose of confining carriers and confining the optical mode, since the refractive index is inversely proportional to the band gap of a semiconductor. This is a guided wave device, very similar to a laser. However, the end facets are not made, and hence the LED does not have a resonant cavity and feedback. Edge-emitting LEDs are very useful for coupling into optical fibers.

## 114.2 Lasers

---

*Laser* is an acronym for light amplification by stimulated emission of radiation. The first semiconductor laser was invented and demonstrated in 1962. The semiconductor laser is essentially a dielectric optical waveguide terminated by facets, or mirrors, to form a resonant cavity. As opposed to an LED, whose output results from spontaneous emission, the output of a laser results from **stimulated emission** and is coherent.

### Lasing Condition and Gain

Figure 114.5(a) shows the process of creating a nonequilibrium population of electrons and holes by photon absorption. The energy of the absorbed photon is  $E_2 - E_1 = h\nu$ . The quasi-Fermi levels are denoted by  $E_{fn}$  and  $E_{fp}$ , respectively. For stimulated emission to occur from this nonequilibrium population distribution, the condition is

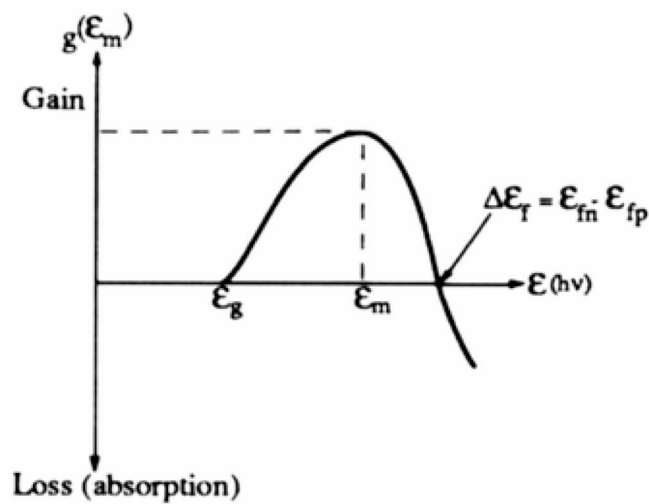
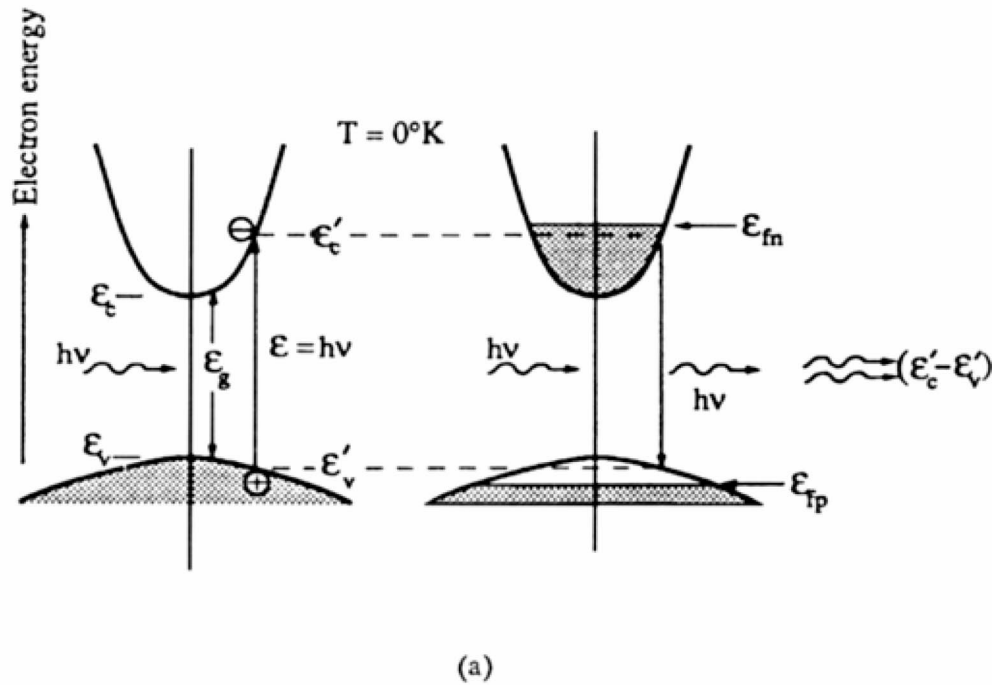
$$E_{fn} - E_{fp} > E \quad (114.8)$$

where  $E = h\nu$  is the energy of the emitted photon. This equation represents the condition for optical gain in a semiconductor. The spectral gain of a semiconductor is given by

$$g(E) = \frac{\sqrt{2}(m_r^*)^{3/2} q^2 p_{cv}^2}{3\pi n_r E_o m_o^2 \hbar^2 c E} (E - E_g)^{1/2} [f_n(E_2) - f_p(E_1)] \quad (\text{cm}^{-1}) \quad (114.9)$$

where  $f_n$  and  $f_p$  are the quasi-Fermi functions defined by the quasi-Fermi levels  $E_{fn}$  and  $E_{fp}$ , respectively, and  $p_{cv}$  is the momentum matrix element. The variation of gain with photon energy for different injection levels is schematically shown in Fig. 114.5(b).

**Figure 114.5** Illustration of absorption and stimulated emission processes in a direct band-gap semiconductor and the corresponding dependence of gain (loss) with photon energy.



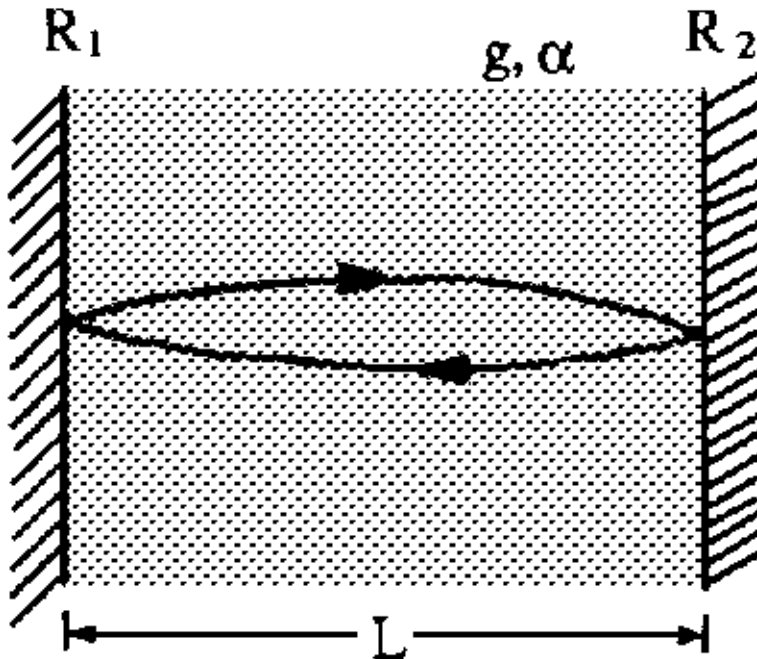
The first requirement for lasing is that the gain must be at least equal to the losses. The second requirement is that the radiation must be coherent. For this, the lasing medium must be placed in a resonant cavity in which *selective amplification* of one frequency can take place. In a semiconductor laser a Fabry-Perot cavity is obtained by cleaving the two ends of a waveguide.

## Threshold Condition

The threshold condition is derived in the context of a semiconductor laser in which a Fabry-Perot cavity is made by cleaving both sides of the waveguide so that the length of the cavity is given by  $l = m\lambda/2$ . Here,  $\lambda$  is a wavelength near the peak of the spontaneous emission spectrum and  $m$  is an integer. The cleaved edges—being optically flat to near perfection—serve as the mirrors of the cavity because of the large refractive index difference between the semiconductor material and air. The reflectivity of these cleaved facets is  $\sim 0.32$ , and therefore a good portion of the optical signal is also transmitted.

Consider the schematic in Fig. 114.6, where the laser cavity is of length  $l$  and has mirrors of reflectivity  $R_1$  and  $R_2$  at the two ends. Let  $g$  and  $\gamma$  be the gain and loss coefficients, respectively. The total loss in the system results from a number of processes, which include (1) transmission at the mirrors; (2) absorption, scattering, and diffraction losses at the mirrors; (3) absorption in the cladding regions; and (4) scattering at defects in the medium. The loss coefficient  $\gamma$  includes all the losses except the transmission at the ends.

**Figure 114.6** Schematics of a Fabry-Perot cavity.



We calculate the threshold gain by considering the change in intensity of a beam of light

undergoing a round-trip within the cavity. The light intensity at the center, after traveling the length  $2l$ , is given by

$$\mathcal{I} = \mathcal{I}_0 R_1 R_2 e^{2(g-\gamma)l} \quad (114.10)$$

where  $\mathcal{I}_0$  is the initial intensity. Now if  $g > \gamma$ , the intensity grows and there is net amplification. At threshold, when the round-trip gain exactly equals the losses, the following are true:

$$\mathcal{I} = \mathcal{I}_0 \quad (114.11)$$

$$1 = R_1 R_2 e^{2(g_{th}-\gamma)l} \quad (114.12)$$

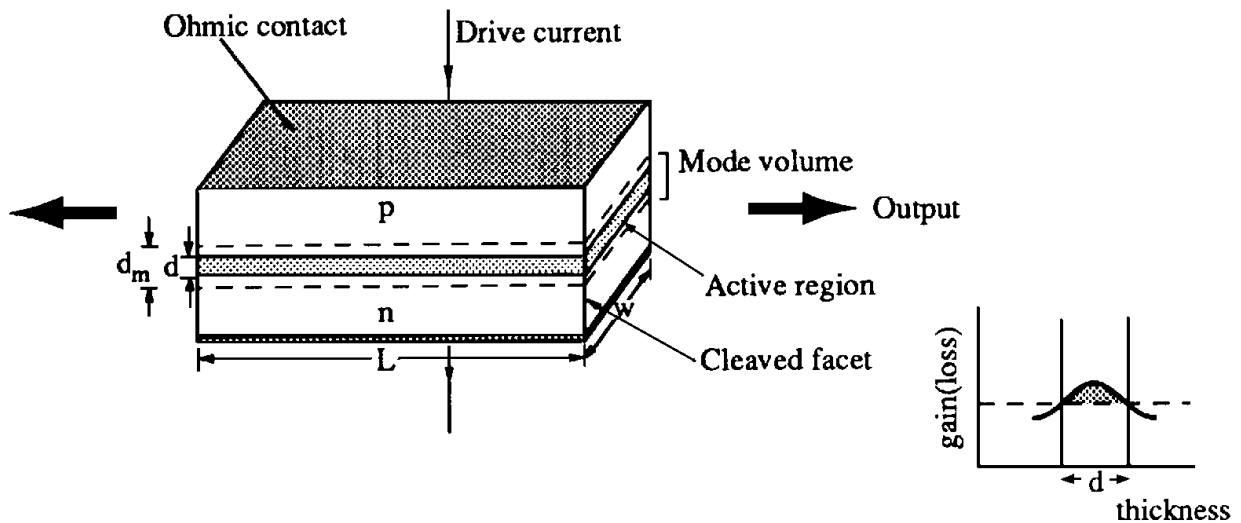
where  $g_{th}$  is the threshold gain. Equation (114.12) can be expressed in the form

$$g_{th} = \gamma + \frac{1}{2l} \ln \left( \frac{1}{R_1 R_2} \right) \quad (114.13)$$

## Operating Principles

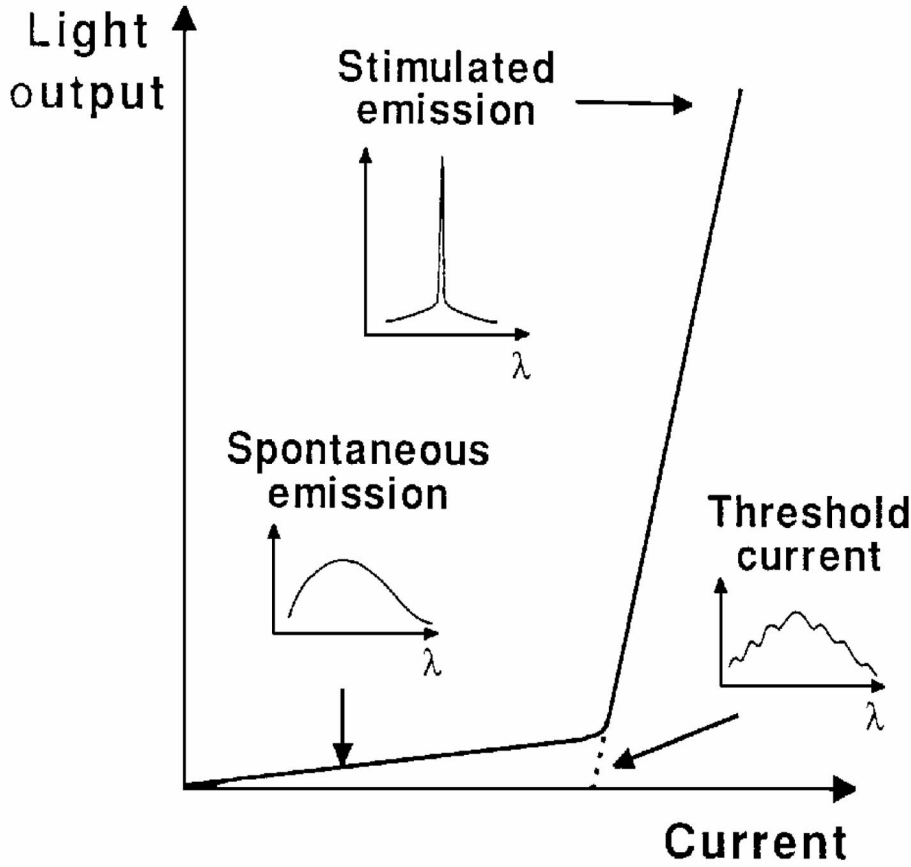
The operating principle of the junction laser is almost identical to that of the LED, in that a large forward bias is applied to inject minority carriers into an active region. In a homojunction laser this active region is defined by the diffusion length of the minority carriers, whereas in a heterojunction laser the active region is defined by the heterojunction boundaries. The heterojunction also helps in confining the optical mode and forming a dielectric waveguide. This device is schematically shown in Fig. 114.7.

**Figure 114.7** Schematics of a broad-area junction laser with cleaved facets.



Within the active region of the junction laser, the gain will eventually exceed the losses with an enhancement in the rate of stimulated emission. Outside the active region the losses will dominate, as depicted in Fig. 114.7. As the rate of stimulated emission increases in the active region, the round-trip gain in the cavity overcomes losses and lasing commences. The injection current at which this occurs is called the *threshold current*, shown schematically in Fig. 114.8. The nature of the spectral output, shown in Fig. 114.8, also changes as incoherent spontaneous emission below threshold is eventually replaced by dominant laser modes above threshold.

**Figure 114.8** The light-current characteristics of a junction laser and the corresponding spectral outputs.



The current flowing in a junction laser is given by

$$J = qdR_{\text{sp}} \text{ (A/cm}^2\text{)} \quad (114.14)$$

where  $d$  is the thickness of the active region and  $R_{\text{sp}}(\text{cm}^{-3} \cdot \text{s}^{-1})$  is the total spontaneous emission rate and is given by

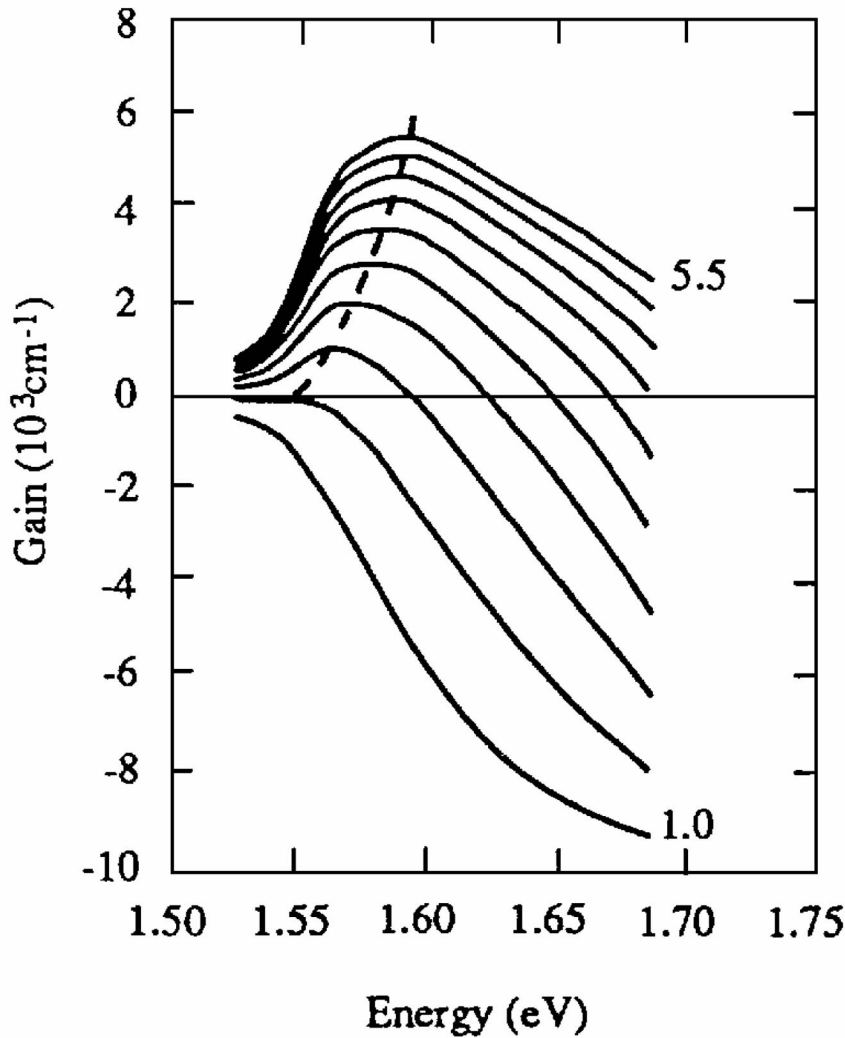
$$R_{\text{sp}} = \frac{(2m_r^*)^{3/2}}{2\pi^2\hbar^3\tau_r} \int_E f_n(E_2)[1 - f_p(E_1)] (E - E_g)^{1/2} dE \quad (114.15)$$

The threshold current density is then given by

$$J_{th} = qdR_{sp}(n_{th}) \quad (114.16)$$

where  $n_{th}$  is the threshold carrier density. This parameter is obtained as follows. The peak gain is plotted against injected carrier density (or injection current) from Fig. 114.5. The plot is shown in Fig. 114.9. The carrier density for which the gain equals the total loss (or  $g_{th}$ ) is  $n_{th}$ .

**Figure 114.9** Calculated TE mode gain coefficient in a 50 Å GaAs/Al<sub>0.3</sub>Ga<sub>0.7</sub>As lattice-matched quantum well laser at 300 K for various carrier injections ( $10^{12}$  carriers/cm<sup>2</sup>) in steps of  $0.5 \cdot 10^{12}/\text{cm}^2$ . The dashed line is the photon energy ( $E_m$ ) at which the gain coefficient is maximum. (Courtesy of J. Singh and J. P. Loehr, University of Michigan.)



## Power Output

With reference to the light-current output of a laser (Fig. 114.8), the optical power output into the modal volume due to a current  $J(> J_{\text{th}})$  can be expressed as

$$P = A(J - J_{\text{th}}) \frac{\eta_i h\nu}{q} \quad (114.17)$$

where  $A$  is the junction area and  $\eta_i$  is the internal quantum efficiency of the semiconductor. A fraction of this power is coupled out through the cleaved facets as useful laser output and the rest is used to overcome the losses,  $\gamma$ , within the cavity. Therefore, the output power of the laser can be expressed as

$$P_0 = A(J - J_{\text{th}}) \left( \frac{\eta_i h\nu}{q} \right) \frac{(1/2l) \ln(1/R_1 R_2)}{[\gamma + (1/2l) \ln(1/R_1 R_2)]} \quad (114.18)$$

## 114.3 Photodetectors

The three main types of detectors are photoconductors, PIN diodes, and avalanche photodiodes. The first and the third types have internal gain. PIN photodiodes have no internal gain but can have very large bandwidths.

Photodetectors are also classified as *intrinsic* or *extrinsic* devices. An intrinsic photodetector detects light of wavelength close to the band gap of the semiconductor. Photoexcitation creates electron-hole pairs, which contribute to the photocurrent. An extrinsic device detects light of energy smaller than the band-gap energy. In these devices the transition corresponding to the absorption of photons involves impurity or defect levels within the band gap.

## Quantum Efficiency and Responsivity

The external quantum efficiency  $\eta_{\text{ext}}$  of a photodetector is given by

$$\eta = \frac{I_{\text{ph}}/q}{P_{\text{inc}}/h\nu} \quad (114.19)$$

where  $I_{\text{ph}}$  is the photocurrent and  $P_{\text{inc}}$  is the incident optical power. The internal quantum efficiency  $\eta_i$  is the number of pairs created divided by the number of photons absorbed and is usually very high. The external quantum efficiency depends on the absorption coefficient,  $\alpha$ , of the material and the thickness of the absorbing region: If this thickness is  $d$ , then the two efficiencies are related by the equation

$$\eta_{\text{ext}} = \eta_i (1 - e^{-\alpha d}) \quad (114.20)$$

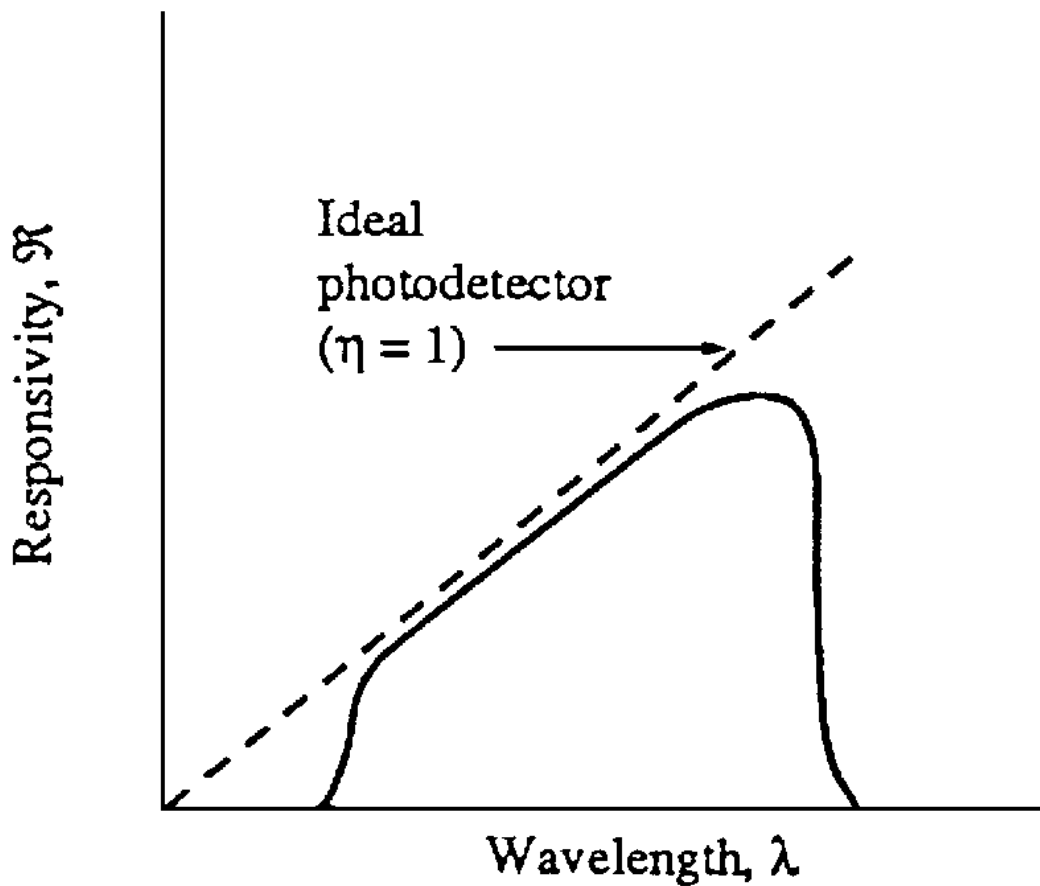
The *responsivity* of a detector,  $R$ , is defined as



$$R = \frac{I_{\text{ph}}}{P_{\text{inc}}} = \frac{\eta q}{h\nu} = \frac{\eta \lambda (\mu\text{m})}{1.24} \text{ (A/W)} \quad (114.21)$$

This relation indicates that for a fixed value of  $\eta$ ,  $R$  should increase linearly with  $\lambda$ . In reality, however,  $\eta$  depends on the absorption coefficient  $\alpha$ , which in turn depends on the incident wavelength  $\lambda$ . Therefore, the long-wavelength cutoff of the spectral response is determined by the absorption edge, or band gap, of the semiconductor. A short-wavelength cutoff also exists because at short wavelengths the value of  $\alpha$  is very large in most semiconductors and all the incident optical energy is absorbed near the surface. The spectral response characteristics of a practical device are illustrated in Fig. 114.10.

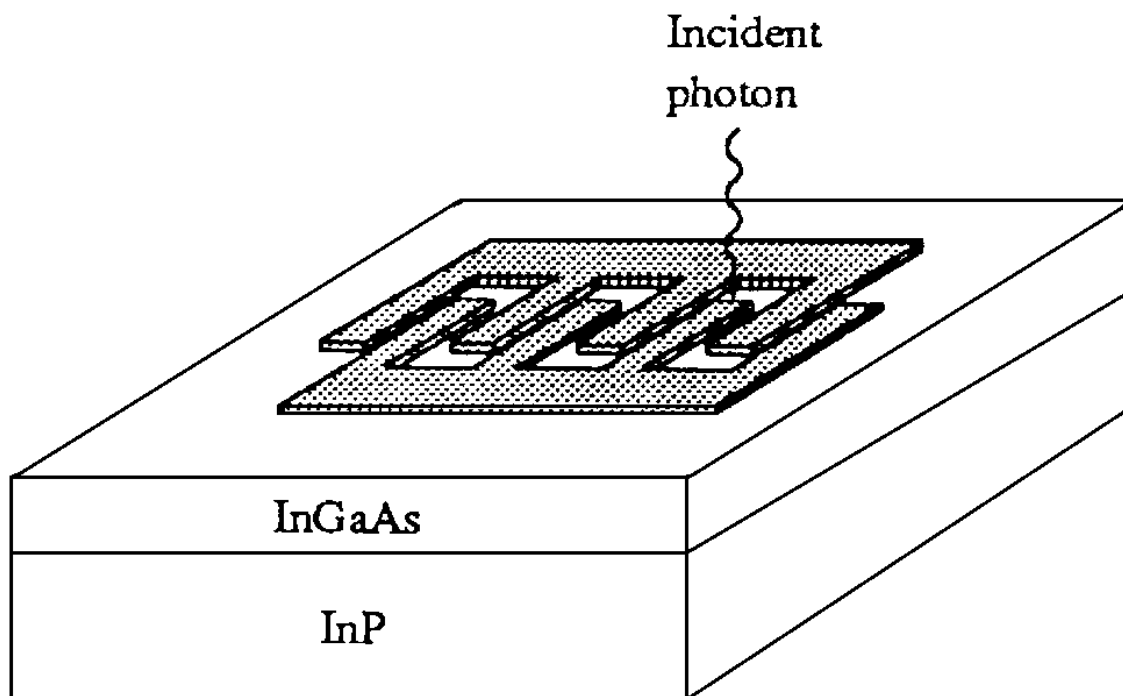
**Figure 114.10** Schematic illustration of the responsivity of ideal and real photodetectors.



## Photoconductor

The photoconductor is a simple photodetection device whose operation is based on the increase in conductivity with photoexcitation. The schematic of a photoconductor is shown in Fig. 114.11. The photogenerated electrons and holes move in opposite directions under the applied bias and produce a photocurrent. The active layer can be formed by diffusion, ion implantation, or epitaxial growth. The thickness of the active layer should be large enough so that it can absorb a significant fraction of the incident light but at the same time small enough so as to minimize the noise current resulting from a low resistance of the semiconductor layer. As we will see later, the separation between the contact pads—either in linear or interdigitated form—is also an important parameter in the operation of the device.

**Figure 114.11** Schematic of an InGaAs/InP photoconductor with interdigitated top contacts.



External quantum efficiencies as high as 80 to 90% can be obtained in photoconductors that are thick enough to absorb most of the incident light if an antireflection coating or a wider band-gap window layer is formed on the surface on which light is incident.

A limitation of photoconductive detectors is the noise performance of the device. The noise is principally generated by the large dark current of the device and is known as *Johnson* or *thermal noise*. The resulting noise current,  $i_J$ , is given by

$$\overline{i_J^2} = \frac{4kTB}{R_c} \quad (114.22)$$

where  $B$  is the bandwidth of the device and  $R_c$  is the resistance of the photoconducting channel. The photogenerated electrons and holes move in opposite directions in the active region under the applied bias. The resulting photocurrent will persist until both carriers are collected at the electrodes or until they recombine in the bulk of the semiconductor before reaching the respective contacts. The parameter of importance is the minority-carrier recombination time,  $\tau$ . The time for detection of the photogenerated current is limited by the transit time between electrodes of the faster carrier—usually the electrons. Therefore, the shortest response time (maximum bandwidth) can be obtained by minimizing the distance between the contacts. Note that the persistence of the slower hole in an n-type channel after the electron is collected will increase the response time, and therefore

$$\text{Bandwidth} \propto \frac{1}{\tau} \quad (114.23)$$

However, the continued persistence of the hole in the channel of an n-type photoconductor will draw more electrons to maintain charge neutrality. This constitutes a photocurrent gain,  $\Gamma_G$ , which is defined as

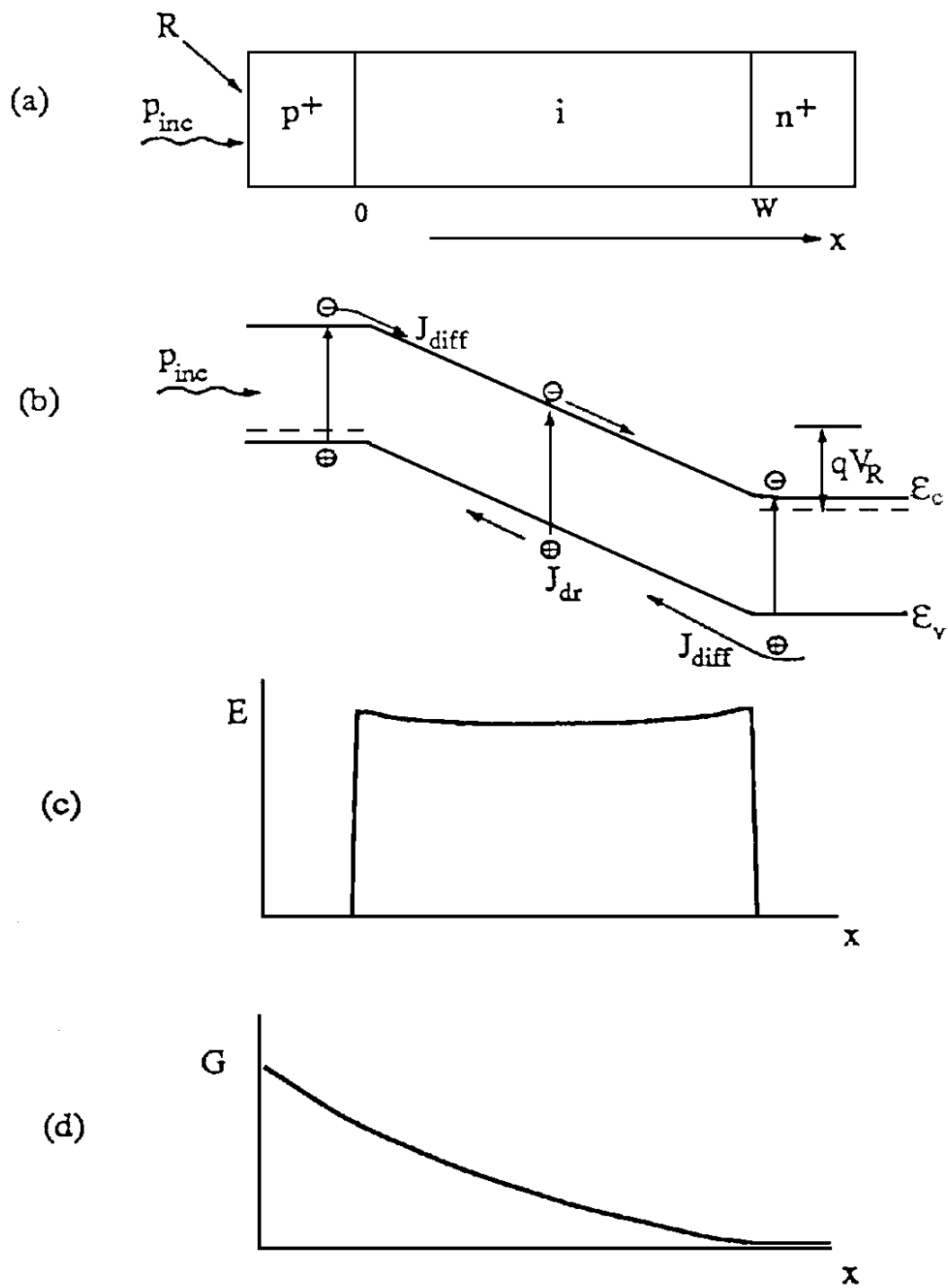
$$\Gamma_G = \frac{\tau}{t_{tr}} \quad (114.24)$$

where  $t_{tr}$  is the transit time of the electrons.

## PIN Photodiode

The p-i-n (or PIN) photodiode is a junction diode that has an undoped i-region ( $p^-$  or  $n^-$ , depending on the method of junction formation) inserted between  $p^+$  and  $n^+$  regions. Because of the very low density of free carriers in the i-region and its high resistivity, any applied bias drops almost entirely across the i-layer, which is fully depleted at zero or a very low value of reverse bias. The depletion layer width must be tailored to meet the requirements of photoresponse and bandwidth. For high response speed, the depletion layer width should be small and for high quantum efficiency, or responsivity, the width should be large. Therefore, a trade-off is necessary. The operation of a PIN photodiode is shown in [Fig. 114.12](#). For practical applications photoexcitation is provided either through an opening in the top contact or through an etched hole in the substrate. The latter reduces the active area of the diode to the size of the incident light beam.

**Figure 114.12** Absorption and carrier generation in a reverse-biased p-i-n diode.



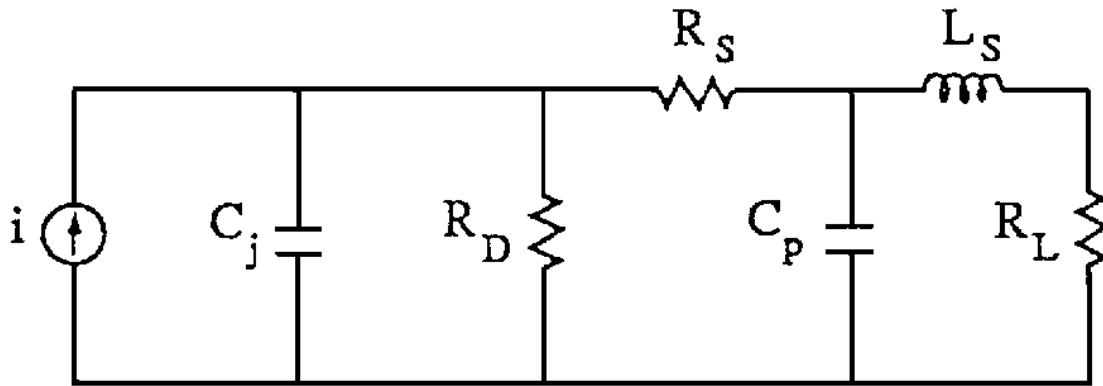
Since there is no internal optical gain in a PIN diode, the maximum internal quantum efficiency  $\eta_i$  is 100% and the gain-bandwidth product is equal to the bandwidth. By careful choice of material parameters and device design, very large bandwidths can be attained. The response speed and bandwidth are ultimately limited either by transit time effects or by circuit parameters.

The equivalent circuit of a PIN diode is shown in Fig. 114.13.  $C_j$  is the junction capacitance originating from the depletion region.  $C_p$  is termed the *parasitic capacitance*, which is external to the wafer.  $L_s$  is the total series inductance, mostly in the leads.  $R_D$  is the diode junction resistance. It has a large value under reverse bias and a very small value under forward bias.  $R_s$  is the series resistance of the diode, which is the sum of the contact resistances and the resistance of the undepleted regions of the diode.  $R_L$  is the load resistance. Typical values of these elements for a reverse-biased diode are listed in Fig. 114.13. The 3-dB bandwidth of a PIN diode can be expressed as

$$f_{3\text{dB}} = \frac{2.8}{2\pi t_r} \quad (114.25)$$

where  $t_r$  is the rise time of the temporal response of a diode to a short optical pulse.

**Figure 114.13** Equivalent circuit of PIN photodiode. Typical values of the circuit elements of a reverse-biased ( $\sim 10\text{V}$ ) InGaAs PIN diode also are listed.



$$R_D = 100 \text{ M}\Omega$$

$$C_j = 80 \text{ fF}$$

$$R_L = 50 \text{ }\Omega$$

$$C_p = 15 \text{ fF}$$

$$R_s = 10 \text{ }\Omega$$

$$L_s = 60 \text{ pH}$$

The external quantum efficiency of the photodiode is given by

$$\eta_{\text{ext}} = \eta_i(1 - R)(1 - e^{-\alpha W}) \quad (114.26)$$

where  $R$  is the reflectivity of the top surface,  $\alpha$  is the absorption coefficient, and  $W$  is the i-region width.

The noise performance of a photodiode is described by the *noise equivalent power* (NEP) and the detectivity,  $D^*$ . The former is given by

$$\text{NEP} = \frac{h\nu}{q\eta} \left[ 2q(I_{\text{ph}} + I_D) + \frac{4kT}{R_{\text{eq}}} \right]^{1/2} (\text{W/Hz}^{1/2}) \quad (114.27)$$

where

$$\frac{1}{R_{\text{eq}}} = \frac{1}{R_D} + \frac{1}{R_L} + \frac{1}{R_i} \quad (114.28)$$

$R_D$  and  $R_L$  are indicated in the diode equivalent circuit of Fig. 114.13, and  $R_i$  is the input resistance of the next stage, usually a preamplifier.  $I_{\text{ph}}$  and  $I_D$  are the photocurrent and dark current of the device, respectively. The detectivity is expressed as

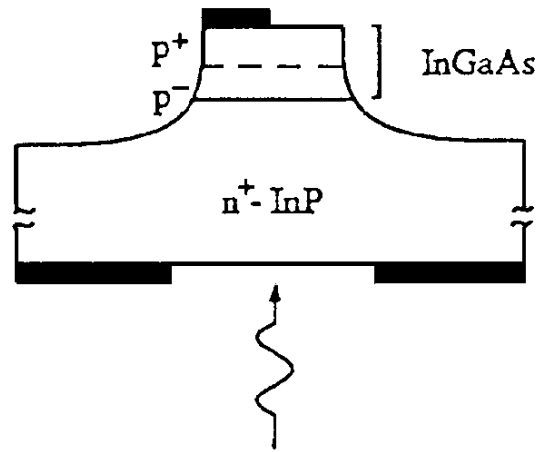
$$D^* = \frac{A^{1/2} B^{1/2}}{(\text{NEP})} (\text{cm} \cdot \text{Hz}^{1/2} / \text{W}) \quad (114.29)$$

where  $A$  is the active area of the device and  $B$  is the bandwidth. The reference bandwidth is usually taken as 1 Hz and  $D^*$  is expressed as  $D^*(\lambda f, 1)$ , where  $\lambda$  is the wavelength and  $f$  is the frequency of modulation.

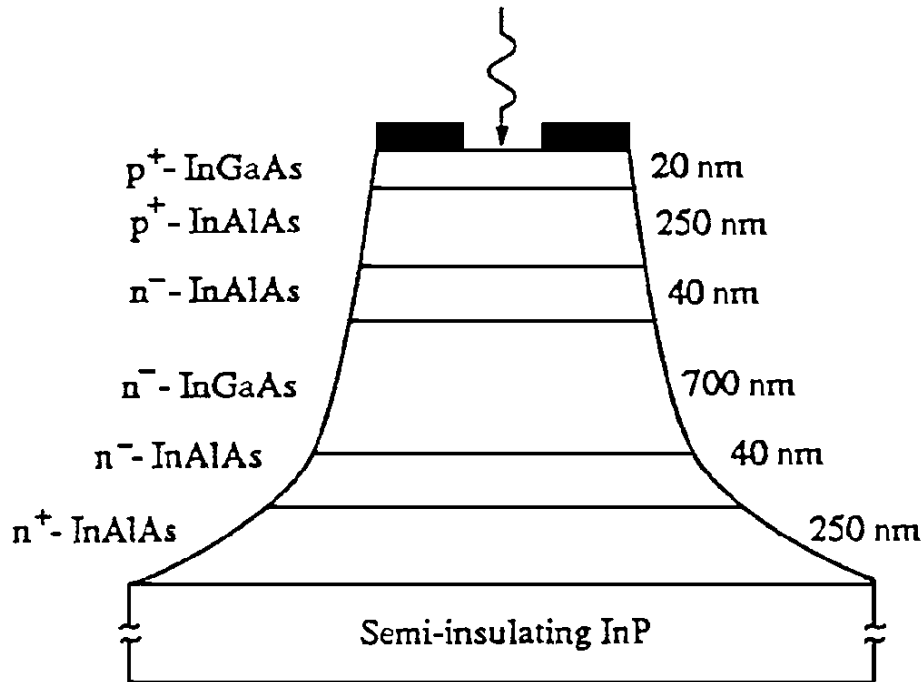
## Heterojunction and Avalanche Photodiodes

One of the essential requirements of a photodiode is a low reverse-bias dark current. In devices made of large-band-gap semiconductors this requirement is easily met. However, for optical communication applications, photodiodes are usually made of  $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ , which has a band gap of 0.75 eV and is lattice matched to InP ( $E_g = 1.35$  eV). Therefore a homojunction diode made of the ternary material will have a large dark current. A solution to this problem is to place the p-n junction in a wider-band-gap, lattice-matched material and absorb the light at  $1.55 \mu\text{m}$  in the lower-band-gap material. The resulting device, with reduced dark current and optimized absorption characteristics, is a heterojunction photodiode, schematically shown in Fig. 114.14.

**Figure 114.14** Two types of heterojunction photodiodes: (a) InGaAs/InP diode made by epitaxy and diffusion and (b) InGaAs/InAlAs/InP diode made by epitaxy.



(a)



(b)

For many applications it is desirable to use a detector with a large sensitivity. Large optical gains can be obtained in an *avalanche photodiode* (APD). The device is essentially a reverse-biased p-n junction that is operated at voltages close to the breakdown voltage. Photogenerated carriers in the depletion region travel at their saturation velocities, and, if they acquire enough energy from the field during such transit, an *ionizing* collision with the lattice can occur. The field necessary to produce an ionization collision is  $\sim 10^5$  V/cm. *Secondary* electron-hole pairs are produced in the process, which again drift in opposite directions, together with the primary carrier, and all or some of them may produce new carriers. The process is known as *impact ionization*, which leads to carrier multiplication and gain. Depending on the semiconductor material and device design, very large avalanche gains ( $\sim 200$  or more) can be achieved, and the avalanche photodiode therefore exhibits very high sensitivity.

## 114.4 Conclusion

---

Some of the most important optoelectronic devices have been briefly described here. More detailed description of their physics and properties are available in the listed references. Another device in which optoelectronic energy conversion takes place is the solar cell, which is also termed a *photovoltaic device*.

### Defining Terms

**Band gap:** The forbidden energy gap in a semiconductor between the top of the valence band and the bottom of the conduction band.

**Effective mass:** Mass of carriers in the lattice of a semiconductor, which is usually different from that of a free carrier due to the presence of the periodic potential in the lattice.

**Heterojunction:** Junction between two semiconductors, usually with different band gaps.

**Lattice matched:** The lattice constants of two semiconductors, usually forming a heterojunction, are equal. They could be a substrate and an epitaxial layer grown on it.

**Stimulated emission:** Emission of light in which all or most of the photons have the same phase, frequency, and direction of propagation—that is, they are *coherent*.

### References

- Bhattacharya, P. 1994. *Semiconductor Optoelectronic Devices*. Prentice Hall, Englewood Cliffs, NJ.
- Saleh, B. E. A. and Teich, M. C. 1991. *Fundamentals of Photonics*. John Wiley & Sons, New York.
- Wilson, J. and Hawkes, J. F. B. 1989. *Optoelectronics: An Introduction*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.

### Further Information

- Agrawal, G. P. and Dutta, N. K. 1993. *Semiconductor Lasers*, 2nd ed. Van Nostrand-Reinhold, New York.
- Forrest, S. R. 1986. Optical detectors: Three contenders. *IEEE Spectrum*. 23:76–84.



Rajashekara, K. S. "Power Electronics"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

### 115.1 Power Semiconductor Devices

Thyristor and Triac • Gate Turnoff Thyristor (GTO) • Power Transistor • Power MOSFETs • Insulated Gate Bipolar Transistor (IGBT) • MOS-Controlled Thyristor (MCT)

### 115.2 Power Conversion

DC-DC Converters • AC-DC Converters • DC-AC Converters • Direct AC-AC Converters

## Kaushik S. Rajashekara

*Delphi Energy & Engine Management Systems*

Power electronic systems deal with the process of converting electrical power from one form to another. The power electronic apparatuses performing the power conversion are called *power converters*. The power conversion is achieved using power semiconductor devices, which are used as switching elements. Power electronic systems cover a wide range of applications, as shown in [Table 115.1](#). The power range of these systems varies from a few watts to several megawatts.

**Table 115.1** Power Electronic Applications

a. Residential	d. Transportation
Refrigeration and freezer	Traction control of electric vehicles
Space heating	Battery chargers for electric vehicles
Air conditioning	Electric locomotives
Cooking	Streetcars, trolleybuses
Lighting	Subways
Electronics (personal computers, other entertainment equipment)	Automotive electronics including engine controls
b. Commercial	e. Utility systems
Heating, ventilating, and air conditioning	High-voltage dc transmission (HVDC)
Central refrigeration	Static var generation (SVG)
Lighting	Supplemental energy sources (wind, photovoltaic)
Computers and office equipment	Energy storage systems
Uninterruptible power supplies (UPS)	Induced-draft fans and boiler feed-water pumps
Elevators	f. Aerospace
c. Industrial	Space shuttle power supply system
Pumps	Satellite power systems
Compressors	Aircraft power systems
Blowers and fans	g. Telecommunications
Machine tools (robots)	Battery chargers

Arc furnaces, induction furnaces  
Lighting  
Industrial lasers  
Induction heating  
Welding

Power supplies (DC and UPS)

---

*Source:* Mohan, N. and Undeland, T. 1989. *Power Electronics: Converters, Applications, and Design*. John Wiley & Sons, New York. With permission.

---

## 115.1 Power Semiconductor Devices

---

The modern age of power electronics began with the introduction of thyristors in the late 1950s. Now there are several types of power devices available for high-power and high-frequency applications. The most notable power devices are gate turnoff thyristors (GTOs), power Darlington transistors, power MOSFETs (metal-oxide silicon field-effect transistors), and insulated gate bipolar transistors (IGBTs). Power semiconductor devices are the most important functional elements in all power conversion applications. The power devices are mainly used as switches to convert power from one form to another.

### Thyristor and Triac

The thyristor, also called the *silicon-controlled rectifier (SCR)*, is basically a four-layer, three-junction PNP device. It has three terminals: anode, cathode, and gate. The device is turned on by applying a short pulse across gate and cathode. Once the device turns on, the gate loses its ability to turn off the device. The turnoff is achieved by applying a reverse voltage across the anode and cathode. There are basically two classifications of thyristors: converter grade and inverter grade. The main difference between a converter-grade and an inverter-grade thyristors is the low turnoff time (in the order of  $\mu s$ ) for the latter. The converter-grade thyristors are slow and are used in natural commutation (or phase-controlled) applications. The inverter-grade thyristors are turned off by forcing the current to zero using external commutation circuit. This requires additional commutating components, thus resulting in additional losses in the inverter. Inverter-grade thyristors have been used in forced commutation applications such as DC-DC choppers and DC-AC inverters. At present, use of thyristors in DC-AC and DC-DC applications is diminishing. But for AC-to-DC power conversion, thyristors are still the best devices. Thyristors are available up to 5000 V, 3000 A.

A triac is functionally a pair of converter-grade thyristors connected in antiparallel. Because of the integration, the triac has poor gate current sensitivity at turn-on and longer turnoff time. A triac is mainly used in phase control application such as in AC regulators for lighting and fan control and in solid-state AC relays.

### Gate Turnoff Thyristor (GTO)

A GTO is a power switching device that can be turned on by a short pulse of gate current and turned off by a reverse gate pulse. This reverse gate current amplitude is dependent on the anode

current to be turned off. Hence there is no need for external commutation circuit to turn it off. Because turnoff is provided by bypassing carriers directly to the gate circuit, its turnoff time is short, thus giving it a greater capacity for high-frequency operation than thyristors. GTOs are used in high-power DC-AC power conversion applications. GTOs are available up to 6000 V, 4000 A.

## Power Transistor

Power transistors are used in applications ranging from a few to several hundred kilowatts and switching frequencies up to about 8 kHz. Power transistors used in power conversion applications are generally NPN type. The device is turned on by supplying sufficient base current, and this base drive has to be maintained throughout its conduction period. It is turned off by removing the base drive and making the base voltage slightly negative [within  $-V_{BE(max)}$ ]. The saturation voltage of the device is normally 0.5 to 2.5 V and increases as the current increases. Hence the on-state losses increase more than proportionately with current. The transistor off-state losses are much lower than the on-state losses because the leakage current of the device when blocking is normally less than a few milliamperes. Because of the relatively larger switching times, the switching loss significantly increases with switching frequency. Power transistors can block only forward voltages. The reverse peak voltage rating of these devices is as low as 5 to 10 volts.

To eliminate high base current requirements, Darlington configurations are commonly used. They are available in monolithic or in isolated packages. The Darlington configuration presents a specific advantage in that it can considerably increase the current switched by the transistor for a given base drive. The  $V_{CE(sat)}$  for the Darlington is generally more than that of a single transistor of similar rating with corresponding increase in on-state power loss. Darlington transistors are available up to 1200 V, 1000 A.

## Power MOSFETs

Power MOSFETs are marketed by various manufacturers with differences in internal geometry and with names such as VMOS, HEXFET, SIPMOS, TMOS, and so on. They have unique features that make them potentially attractive for high-frequency switching applications. They are essentially voltage-driven rather than current-driven devices, unlike bipolar transistors. The gate of a MOSFET is isolated electrically from the source by a layer of silicon oxide. The gate draws only a minute leakage current, of the order of nanoamperes. Hence the gate drive circuit is simple and the power loss in the gate control circuit is practically negligible.

An important feature of the power MOSFET is the absence of secondary breakdown effect, which is present in a bipolar transistor; as a result, the MOSFET has an extremely rugged switching performance. In MOSFETs the drain to source resistance,  $R_{DS(ON)}$ , increases with temperature, and thus current is automatically diverted from the hot spot. Hence it is easy to parallel the MOSFETs. Power MOSFETs are available up to 600 V, 100 A ratings.

## Insulated Gate Bipolar Transistor (IGBT)

The IGBT has the high-input impedance and high-speed characteristics of a MOSFET, with the

conductivity characteristic (low saturation voltage) of a bipolar transistor. The IGBT is turned on by applying a positive voltage between the gate and emitter, and, as in the MOSFET, it is turned off by making the gate signal zero or slightly negative.

Like the power MOSFET, the IGBT does not exhibit the secondary breakdown phenomenon common to bipolar transistors. However, care should be taken not to exceed the maximum power dissipation and specified maximum junction temperature of the device under all conditions for guaranteed reliable operation. The on-state voltage of the IGBT is heavily dependent on the gate voltage. To obtain a low on-state voltage, a sufficiently high gate voltage must be applied. Compared to a MOSFET structure, the IGBT is generally smaller for the same current rating because of its higher current density. The bipolar action in IGBT reduces the speed of the device, so it exhibits a much lower frequency than the MOSFET. The IGBTs cannot be as easily paralleled as MOSFETs. The IGBTs are available up to 1600 V, 1200 A.

## **MOS-Controlled Thyristor (MCT)**

The MCT is basically a thyristor with built-in MOSFETs to turn on and turn off. It is a rugged, high-power, high-frequency, low-conduction-drop device that is more likely to be used in future medium- and high-power applications. The first generation MCTs of 600 V, 75 A are available in the market. Harris Corporation is in the process of releasing second-generation MCTs and high-current modules.

The MCT has three thyristor-type junctions and PNP layers between the anode and the cathode. The p-MCTs are turned on by a negative gate pulse at the gate with respect to the anode and turned off by a positive voltage pulse. Further research is going on to develop n-MCTs. The main advantages of MCTs over IGBTs are the higher current density (thus reduced device size for a given current and voltage rating), low conduction drop, and rugged switching performance. The MCT, because of its superior characteristics, shows a tremendous possibility for applications such as motor drives, high-power uninterrupted power supplies, static VAR compensators, and high-power active power line conditioners.

## **115.2 Power Conversion**

---

The power converters are generally classified as:

1. DC-DC converters (choppers, buck, and boost converters)
2. AC-DC converters (phase-controlled converters)
3. DC-AC converters (inverters)
4. Direct AC-AC converters (cycloconverters)

### **DC-DC Converters**

DC-DC converters are used to convert unregulated DC voltage to regulated or variable DC voltage at the output. They are widely used in switch-mode DC power supplies and in DC motor drive

applications. In DC motor control applications they are called *chopper-controlled drives*. The input voltage source is usually a battery or is derived from AC supply using a diode bridge rectifier. These converters are generally either hard-switched PWM types or soft-switched resonant-link types. There are several DC-DC converter topologies, the most common ones being buck converter, boost converter, and buck-boost converter.

### Buck Converter

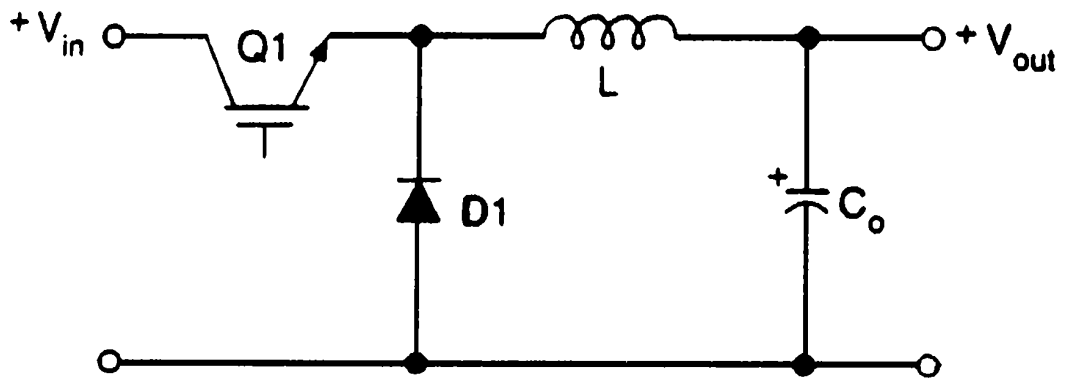
A buck converter is also called a *step-down converter*. Its principle of operation is illustrated by Fig. 115.1(a). The IGBT acts as a high-frequency switch. The IGBT is repetitively closed for a time  $t_{\text{on}}$  and opened for a time  $t_{\text{off}}$ . During  $t_{\text{on}}$  the supply terminals are connected to the load and power flows from supply to the load. During  $t_{\text{off}}$  load current flows through the freewheeling diode  $D_1$ , and the load voltage is ideally zero. The average output voltage is given by

$$V_{\text{out}} = D \cdot V_{\text{in}}$$

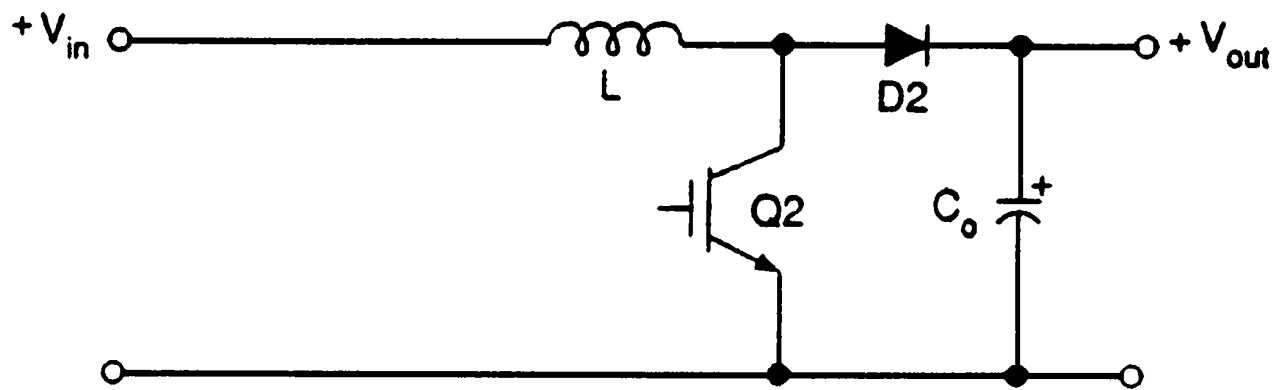
where  $D$  is the **duty cycle** of the switch and is given by  $D = t_{\text{on}}/T$ , where  $T$  is the time for one period. The term  $1/T$  is the **switching frequency** of the power device IGBT.

**Figure 115.1** DC-DC converter configurations: (a) buck converter; (b) boost converter; (c) buck-boost converter. (Source: Rajashekara, K. S., Bhat, A. K. S., and Bose, B. K. 1993. Power electronics. In *The Electrical Engineering Handbook*, ed. R. C. Dorf, p. 709. CRC Press, Boca Raton, FL. With permission.)

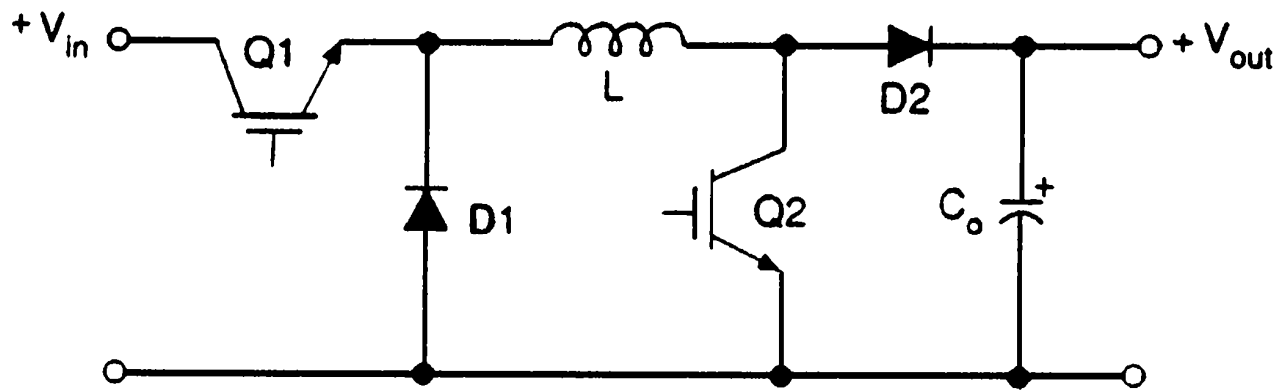
Figure 115.1



(a)



(b)



(c)

### Boost Converter

A boost converter is also called a *step-up converter*. Its principle of operation is illustrated by Fig. 115.1(b). This converter is used to produce higher voltage at the load than the supply voltage. When the power switch is on, the inductor is connected to the DC source and the energy from the supply is stored in it. When the device is off, the inductor current is forced to flow through the diode and the load. The induced voltage across the inductor is negative. The inductor voltage adds to the source voltage to force the inductor current into the load. The output voltage is given by

$$V_{\text{out}} = V_{\text{in}} / (1 - D)$$

Thus, for variation of  $D$  in the range  $0 < D < 1$ , the load voltage  $V_o$  will vary in the range

$$V_{\text{in}} < V_{\text{out}} < \infty$$

### Buck-Boost Converter

A buck-boost converter can be obtained by the cascade connection of the buck and the boost converter. The output voltage  $V_o$ , is given by

$$V_o = V_{\text{in}} \cdot D / (1 - D)$$

The output voltage is higher or lower than the input voltage based on the duty cycle  $D$ . A typical buck-boost converter topology is shown in Fig. 115.1(c). When the power devices are turned on, the input provides energy to the inductor and the diode is reverse biased. When the devices are turned off, the energy stored in the inductor is transferred to the output. No energy is supplied by the input during this interval. In DC power supplies, the output capacitor is assumed to be very large, which results in a constant output voltage. The buck and boost converter topologies enable the four-quadrant operation of a DC motor. In DC drive systems, the chopper is operated in step-down mode during motoring and in step-up mode during regeneration operation.

### Resonant-Link DC-DC Converters

The use of resonant converter topologies in power supplies would help to reduce the switching losses in DC-DC converters and enable the operation at switching frequencies in the megahertz range, which results in reduced size, weight, and cost of the power supplies. Other advantages of resonant converters are that the leakage inductances of the high-frequency transformers and the junction capacitances of semiconductors can be used to further increase the power density of the power supplies. An added advantage is the significant reduction of RFI/EMI. The major disadvantage of resonant converters is increased peak current or voltage stress.

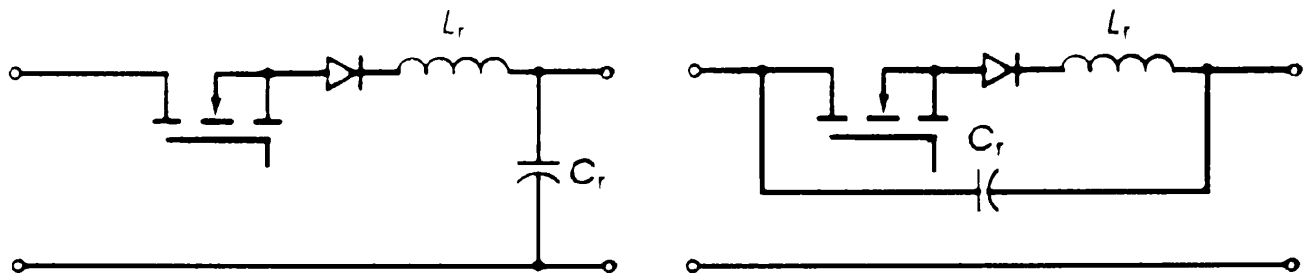
#### *Single-Ended Resonant Converters.*

These type of converters are referred to as *quasi-resonant converters*. Quasi-resonant converters (QRC) do exhibit a resonance in their power section, but, instead of the resonant elements being operated in a continuous fashion, they are operated for only one half of a resonant sine wave at a

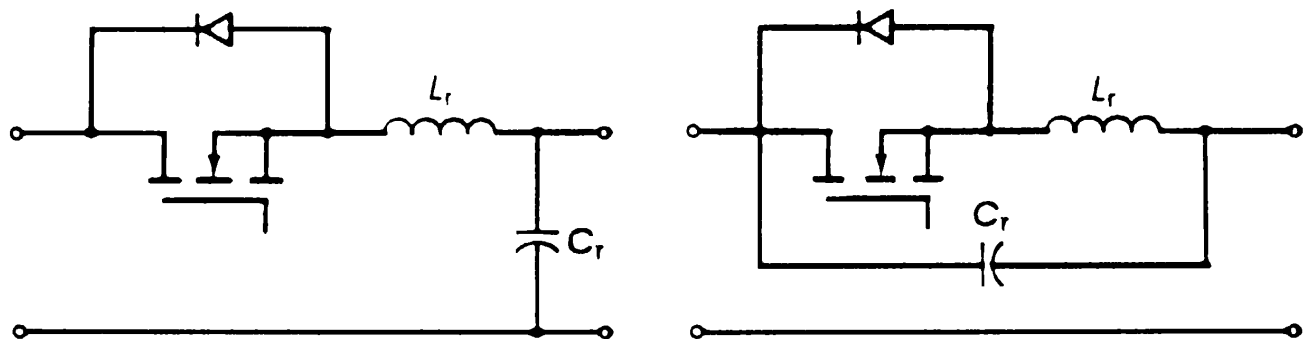


time. The topologies within the quasi-resonant converters are simply resonant elements added to many of the basic PWM topologies as shown in Fig. 115.2 and Fig. 115.3 . The QRCs can operate with zero-current switching or zero-voltage switching or both. The power switch in quasi-resonant converters connects the input voltage source to the tank circuit and is turned on and off in the same step fashion as in PWM switching power supplies. The conduction period of the devices is determined by the resonant frequency of the tank circuit. The power switch turns off after the completion of one half of a resonant period. So, the current at turn-on and turnoff transitions is zero, thus eliminating the switching loss within the switch. In zero-voltage switching QRCs, at the turn-on and turnoff of the power devices, the voltage across the device is zero, thus again reducing the switching loss.

**Figure 115.2.** Zero-current resonant switches.

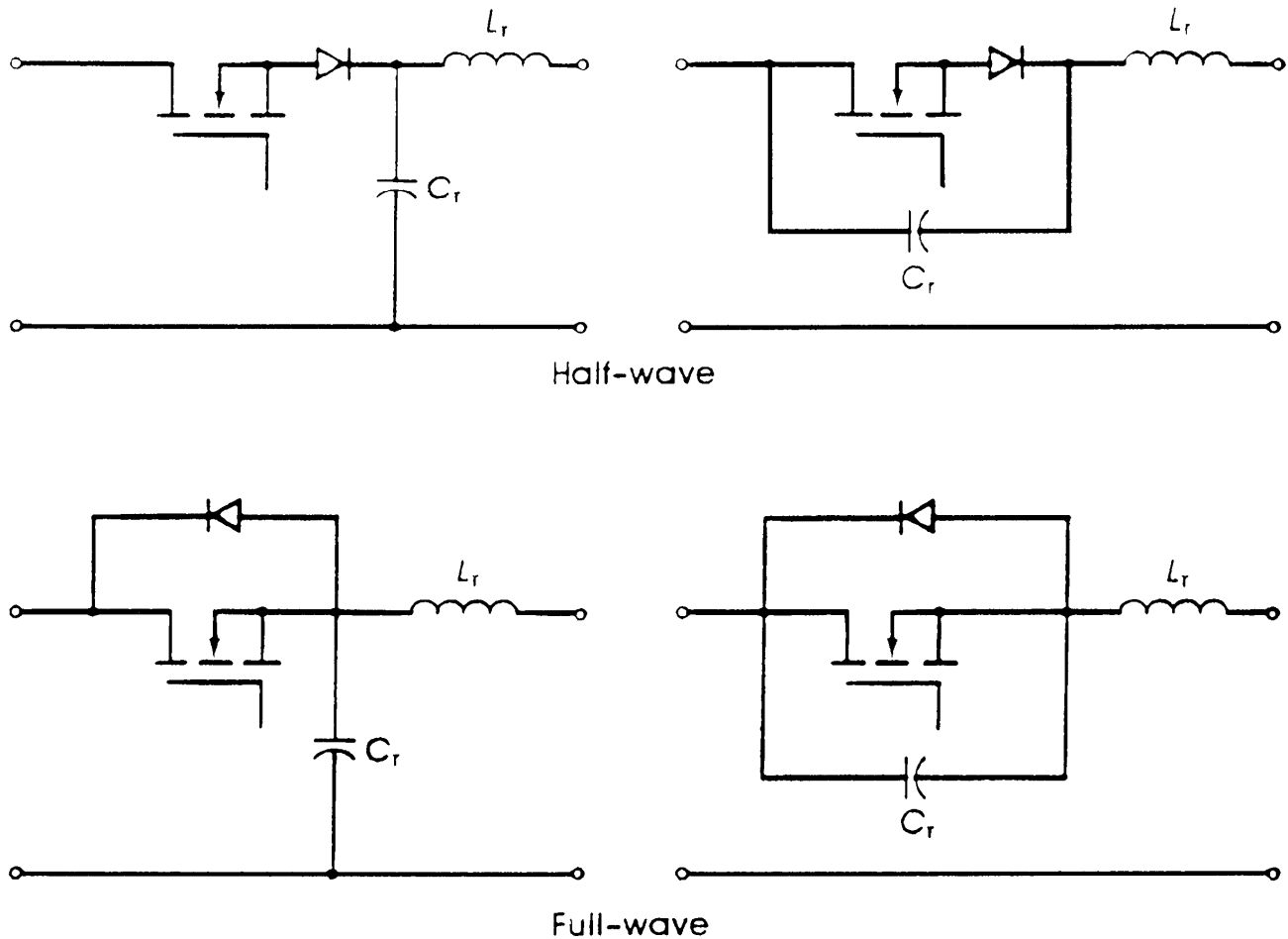


Half-wave, zero-current resonant switches



Full-wave, zero-current resonant switches

**Figure 115.3** Zero-voltage quasi-resonant switches.

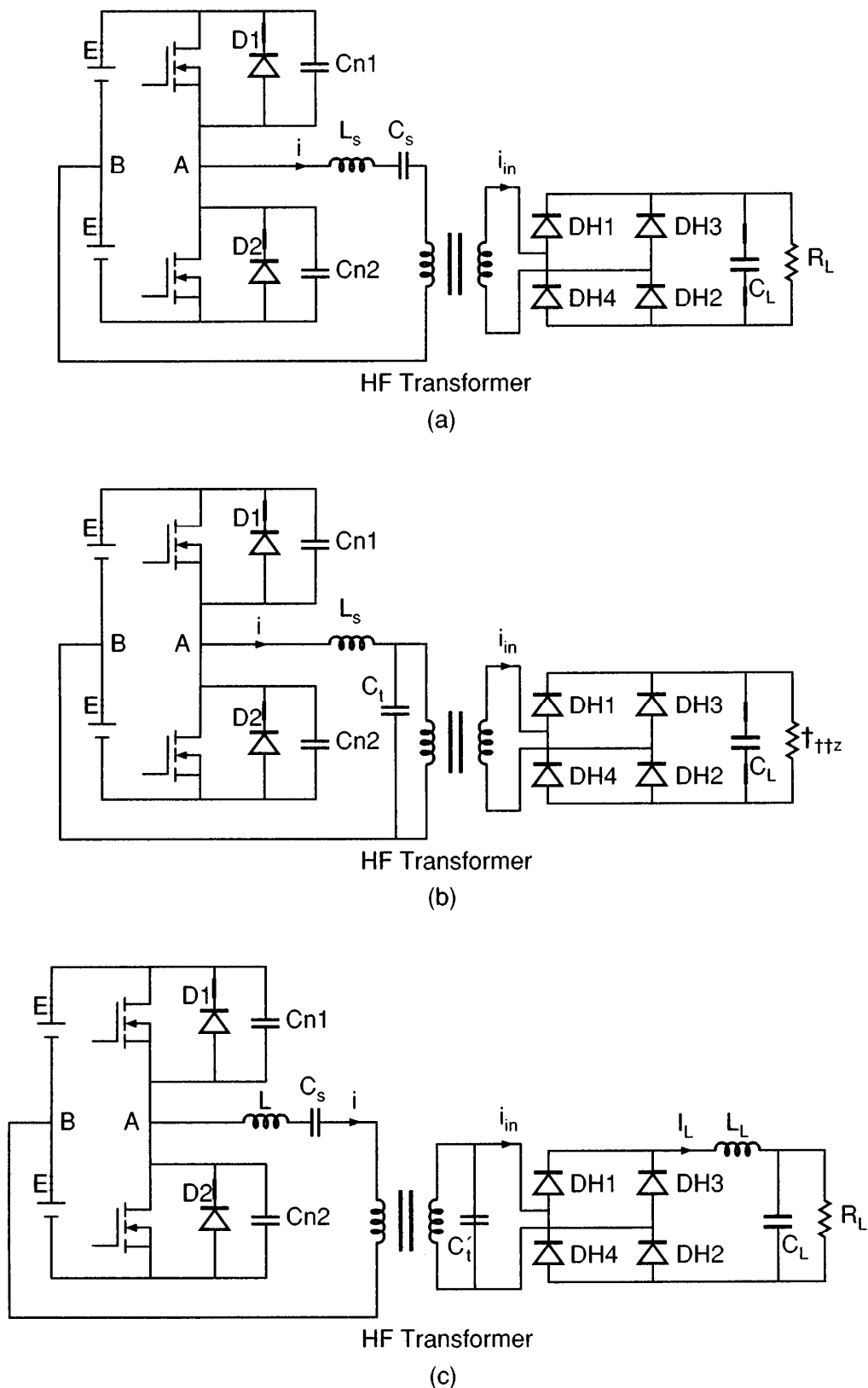


The major problems with the zero-current switching QRC is the high-peak currents through the switch and capacitive turn-on losses. The zero-voltage switching QRCs suffer from increased voltage stress on the power device. The full-wave mode-zero-voltage switching circuit suffers from capacitive turn-on losses.

#### ***Double-Ended Resonant Converters.***

These converters use full-wave rectifiers at the output, and they are generally referred to as resonant converters. A number of resonant converter configurations are realizable by using various resonant tank circuits; the three most popular configurations are the series resonant converter (SRC), the parallel resonant converter (PRC), and the series-parallel resonant converter (SPRC), as shown in [Fig. 115.4](#).

**Figure 115.4** High-frequency resonant converter (half-bridge version) configurations suitable for operation above resonance. (a) Series resonant converter. Leakage inductances of the high-frequency (HF) transformer can be part of resonant inductance. (b) Parallel resonant converter. (c) Series-parallel resonant converter with capacitor  $C_t$  placed on the secondary side of the HF transformer. (Adapted from Rajashekara, K. S., Bhat, A. K. S., and Bose, B. K. 1993. Power electronics. In *The Electrical Engineering Handbook*, ed. R. C. Dorf, p. 720. CRC Press, Boca Raton, FL. With permission.)



Series resonant converters have a high efficiency from full load to part load. Transformer saturation is avoided due to the series-blocking resonating capacitor. The major problems with the SRC are that it requires a very wide change in switching frequency to regulate the load voltage and that the output filter capacitor must carry high-ripple current.

Parallel resonant converters are suitable for low-output voltage, high-output current applications due to the use of filter inductance at the output with low-ripple current requirements for the filter capacitor. The major disadvantage of the PRC is that the device currents do not decrease with the load current, resulting in reduced efficiency at reduced load currents.

The SPRC has the desirable features of both the SRC and the PRC.

Load voltage regulation in resonant converters for input supply variations and load changes is achieved by either varying the switching frequency or using fixed frequency pulse width modulation control.

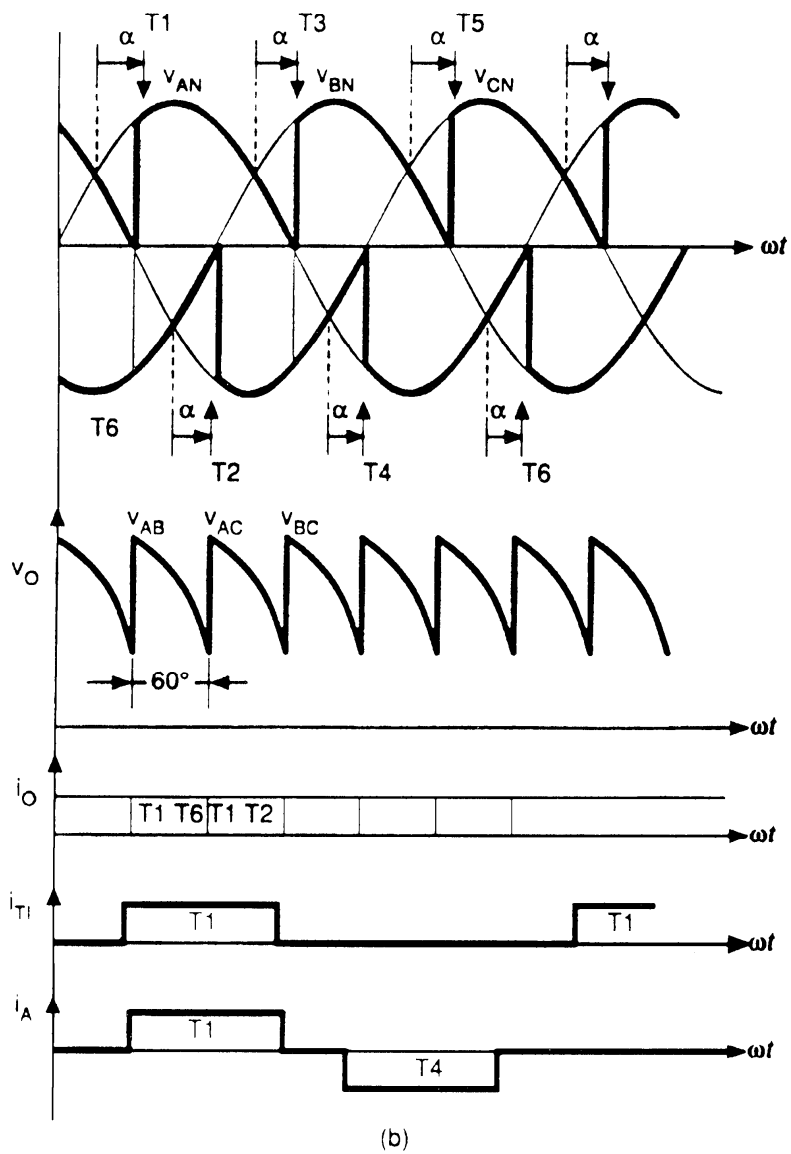
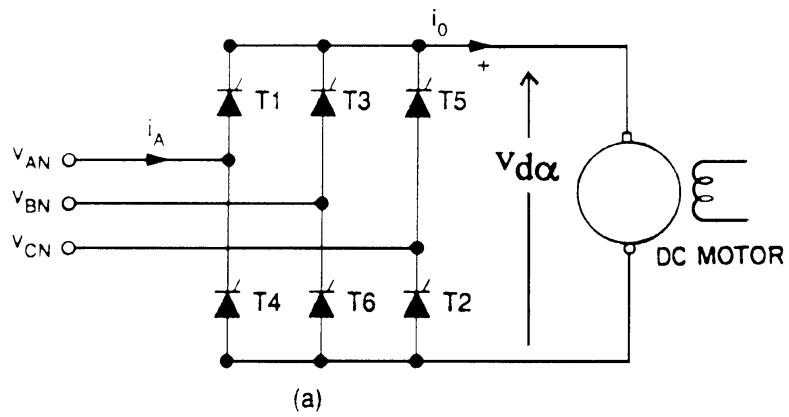
## AC-DC Converters

The basic function of the phase-controlled converter is to convert an alternating voltage of variable amplitude and frequency to a variable DC voltage. The power devices used for this application are generally SCRs. The average value of the output voltage is controlled by varying the conduction time of the SCRs. The turn-on of the SCR is achieved by providing a gate pulse when it is forward biased. The turnoff is achieved by the **commutation** of current from one device to another at the instant the incoming AC voltage has a higher instantaneous potential than that of the outgoing wave. Thus there is a natural tendency for current to be commutated from the outgoing to the incoming SCR without the aid of any external commutation circuitry. This commutation process is often referred to as *natural commutation*.

A three-phase full-wave converter consisting of six thyristor switches is shown in Fig. 115.5(a) . This is the most commonly used three-phase bridge configuration. Thyristors T1, T3, and T5 are turned on during the positive half-cycle of the voltages of the phases to which they are connected, and thyristors T2, T4, and T6 are turned on during the negative cycle of the phase voltages. The reference for the angle in each cycle is at the crossing points of the phase voltages. The ideal output voltage, output current, and input current waveforms are shown in Fig. 115.5(b). The output DC voltage is controlled by controlling the firing angle  $\alpha$ . If the load is a DC motor, the speed of the motor is varied by varying the firing angle  $\alpha$ .

**Figure 115.5** (a) Three-phase thyristor full bridge configuration; (b) output voltage and current waveforms. (Adapted from Rajashekara, K. S., Bhat, A. K. S., and Bose, B. K. 1993. Power electronics. In *The Electrical Engineering Handbook*, ed. R. C. Dorf, p. 704. CRC Press, Boca Raton, FL. With permission.)

**Figure 115.5**



The average output voltage of the converter at a firing angle  $\alpha$  is given by

$$v_{d\alpha} = (3\sqrt{3}/\pi)E_m \cos \alpha$$

where  $E_m$  is the peak value of the phase voltage.

At  $\alpha = 90^\circ$  the output voltage is zero. For  $0^\circ < \alpha < 90^\circ$ ,  $v_{d\alpha}$  is positive and power flows from the AC supply to the DC load. For  $90^\circ < \alpha < 180^\circ$ ,  $v_{d\alpha}$  is negative and the converter operates in the inversion mode. Thus the power can be transferred from the motor to the AC supply, a process known as *regeneration*.

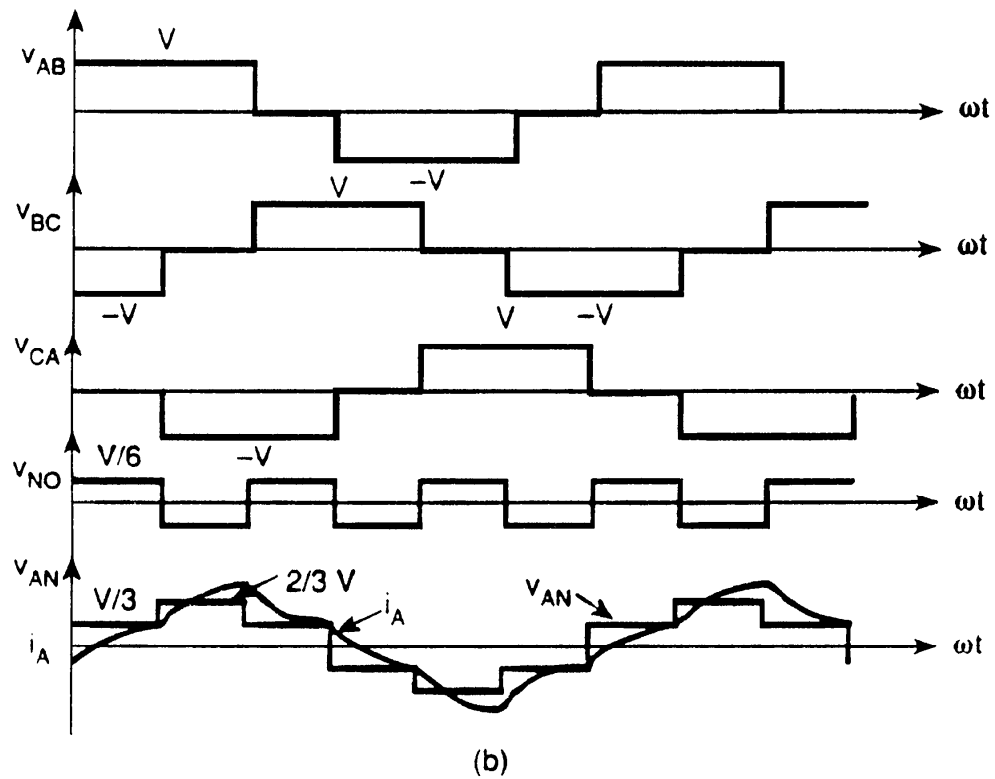
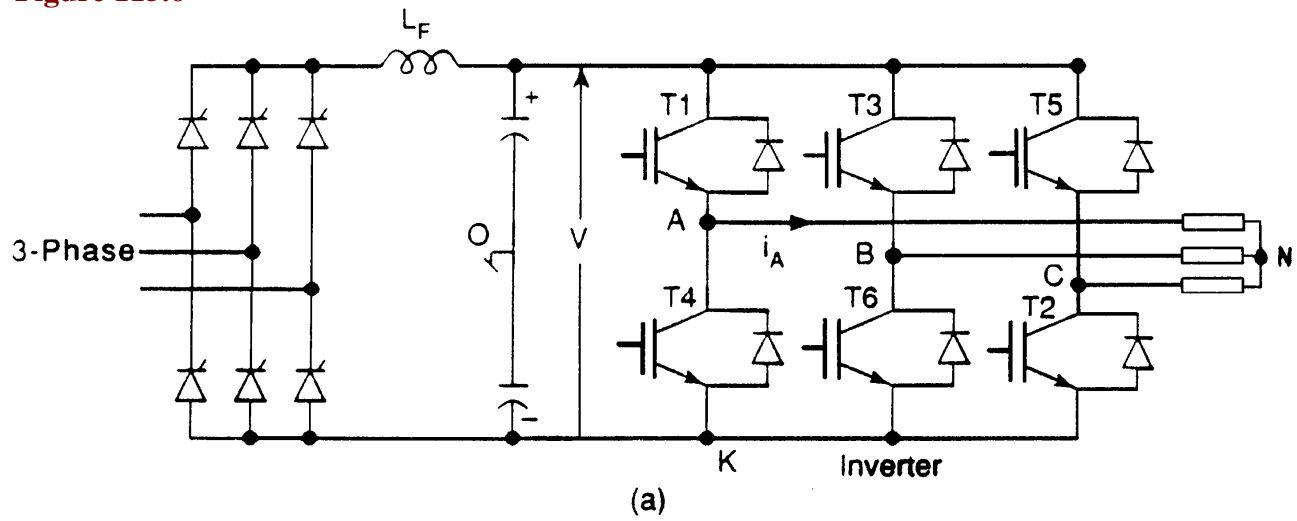
In Fig. 115.5(a), the top or the bottom thyristors could be replaced by diodes. The resulting topology is called *thyristor semiconverter*. With this configuration the input power factor is improved but the regeneration is not possible.

## DC-AC Converters

The DC-AC converters are generally called *inverters*. The AC supply is first converted to DC, which is then converted to a variable-voltage and variable-frequency power. The power conversion stage generally consists of a three-phase bridge converter connected to the AC power source, a DC link with a filter, and the three-phase inverter bridge connected to the load, as shown in Fig. 115.6(a). In the case of battery-operated systems there is no intermediate DC link. An inverter can be classified as a voltage source inverter (VSI) or a current source inverter (CSI). A voltage source inverter is fed by a stiff DC voltage, whereas a current source inverter is fed by a stiff current source. A voltage source can be converted to a current source by connecting a series inductance and then varying the voltage to obtain the desired current. A VSI can also be operated in current-controlled mode and, similarly, a CSI can also be operated in the voltage control mode. These inverters are used in variable-frequency AC motor drives, uninterrupted power supplies, induction heating, static VAR compensators, and so on.

**Figure 115.6** (a) Three-phase converter and voltage source inverter configuration; (b) square-wave inverter waveforms. (Source: Rajashekara, K. S., Bhat, A. K. S., and Bose, B. K. 1993. Power electronics. In *The Electrical Engineering Handbook*, ed. R. C. Dorf, p. 706. CRC Press, Boca Raton, FL. With permission.)

**Figure 115.6**



## Voltage Source Inverter

The inverter configuration shown in Fig. 115.6(a) is the voltage source inverter. The voltage source inverters are controlled either in square-wave mode or in pulse width modulation (PWM) mode. In a square wave mode the frequency of the output is controlled within the inverter; the devices are being used to switch the output circuit between the positive and negative bus. Each device conducts for  $180^\circ$ , and each of the outputs is displaced  $120^\circ$  to generate a six-step waveform as shown in Fig. 115.6(b). The amplitude of the output voltage is controlled by varying DC link voltage. This step is accomplished by varying the firing angle of the thyristors of the three-phase bridge converter at the input. The six-step output is rich in harmonics and hence needs heavy filtering.

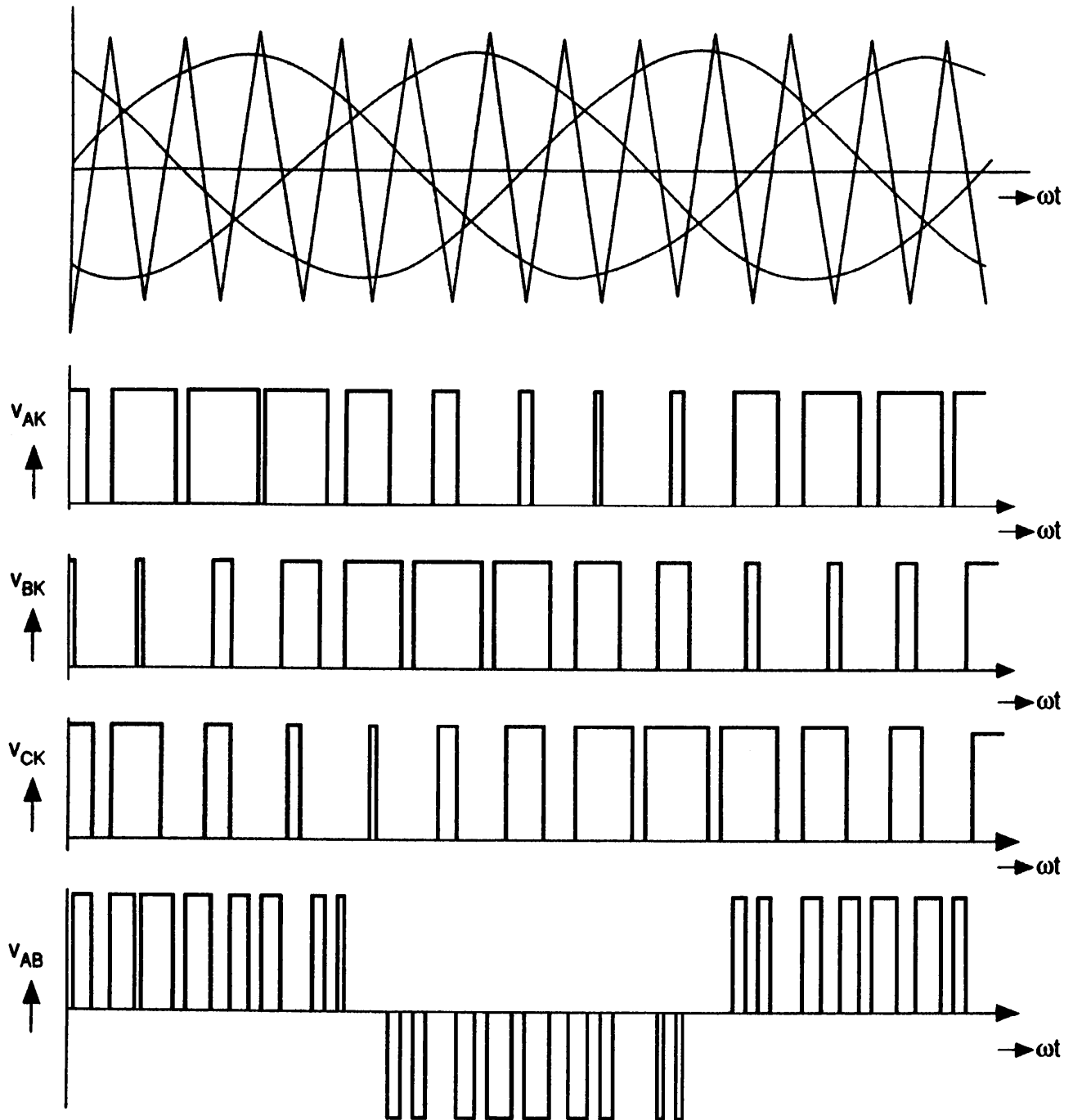
In PWM inverters the output voltage and frequency are controlled within the inverter by varying the width of the output pulses. Hence, at the front end, instead of a phase controlled thyristor converter, a diode bridge rectifier can be used. A very popular method of controlling the output voltage and frequency is by sinusoidal pulse width modulation. In this method a high-frequency triangle carrier wave is compared with a three-phase sinusoidal waveform as shown in Fig. 115.7. The power devices in each phase are switched on at the intersection of the sine and triangle waves. The amplitude and frequency of the output voltage are varied, by varying the amplitude and frequency, respectively, of the reference sine waves. The ratio of the amplitude of the sine wave to the amplitude of the carrier wave is called the *modulation index*. The harmonic components in a PWM wave are easily filtered because they are shifted to a higher-frequency region. It is desirable to have a high ratio of carrier frequency to fundamental frequency to reduce the harmonics of lower-frequency components. There are several other PWM techniques mentioned in the literature. The most notable ones are selected harmonic elimination, **hysteresis control**, and the space vector PWM technique.

## Current Source Inverter

Contrary to the voltage source inverter—in which the voltage of the DC link is imposed on the motor windings—the current source inverter shown in Fig. 115.8 has current that is imposed into the motor. Here, the amplitude and phase angle of the motor voltage depend on the load conditions of the motor. The capacitors and series diodes help commutation of the thyristors. One advantage

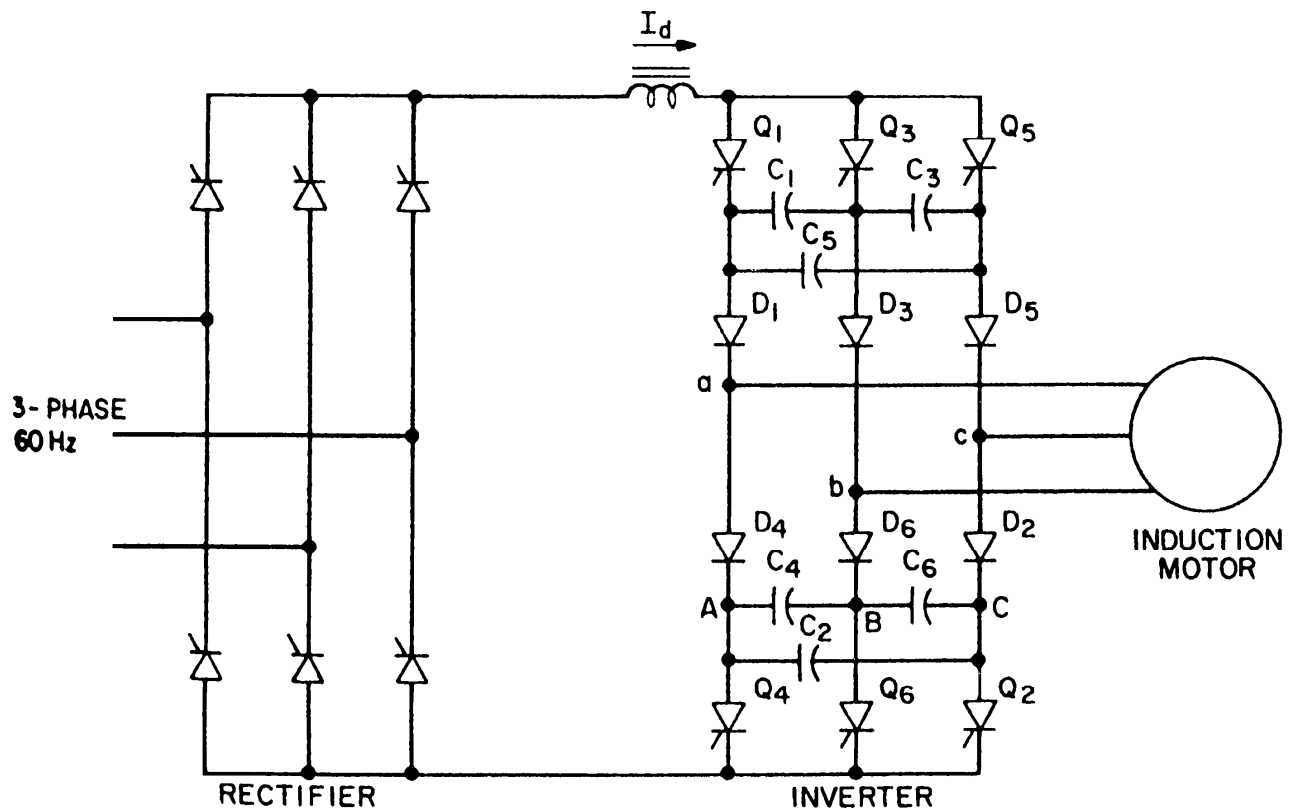


**Figure 115.7** Three-phase sinusoidal PWM inverter waveforms. (Source: Rajashekara, K. S., Bhat, A. K. S., and Bose, B. K. 1993. Power electronics. In *The Electrical Engineering Handbook*, ed. R. C. Dorf, p. 707. CRC Press, Boca Raton, FL. With permission.)



of this drive system is that regenerative braking is easily achieved because the rectifier and inverter can reverse their operation modes. Six-step machine current causes large harmonic heating and torque pulsation, which may be quite harmful at low-speed operation. Another disadvantage is that the converter system cannot be controlled in open loop like in a voltage source inverter.

**Figure 115.8** Force-commutated current-fed inverter control of an induction motor. (Source: Rajashekara, K. S., Bhat, A. K. S., and Bose, B. K. 1993. Power electronics. In *The Electrical Engineering Handbook*, ed. R. C. Dorf, p. 733. CRC Press, Boca Raton, FL. With permission.)

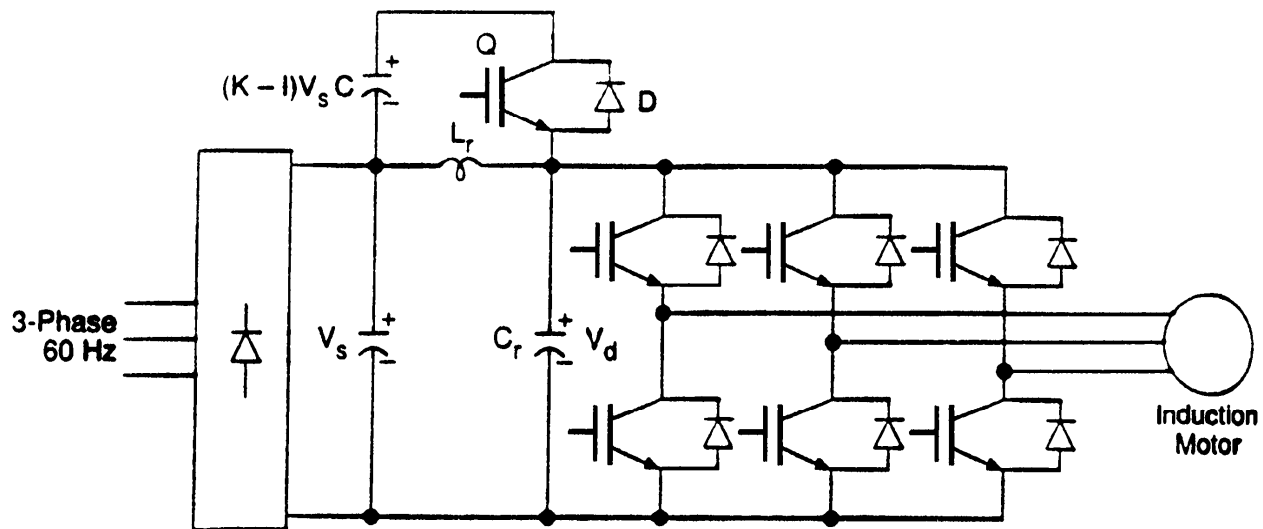


### Resonant-Link Inverter

The use of resonant switching techniques can be applied to inverter topologies to reduce the switching losses in the power devices. They also permit high switching frequency operation to reduce the size of the magnetic components in the inverter unit. In the resonant DC link inverter shown in Fig. 115.9, a resonant circuit is added at the inverter input to convert a fixed DC to a pulsating DC voltage. This resonant circuit enables the devices to be turned on and off during the zero-voltage interval. Zero-voltage switching (ZVS) or zero-current switching (ZCS) is often called *soft switching*. Under soft switching, the switching loss in the power devices is almost eliminated. The EMI problem is less severe because resonant voltage pulses have lower  $dv/dt$  compared to those of hard-switched PWM inverters. Also, the machine insulation is less stretched because of lower  $dv/dt$  resonant voltage pulses. In Fig. 115.9 all the inverter devices are turned on simultaneously to initiate a resonant cycle. The commutation from one device to another is initiated

at the zero-DC link voltage. The inverter output voltage is formed by the integral numbers of quasi-sinusoidal pulses. The circuit consisting of devices Q, D, and the capacitor C acts as an active clamp to limit the DC link voltage to about 1.4 times the diode rectifier voltage  $V_s$ . It is possible to use passive clamp circuits instead of the active clamp.

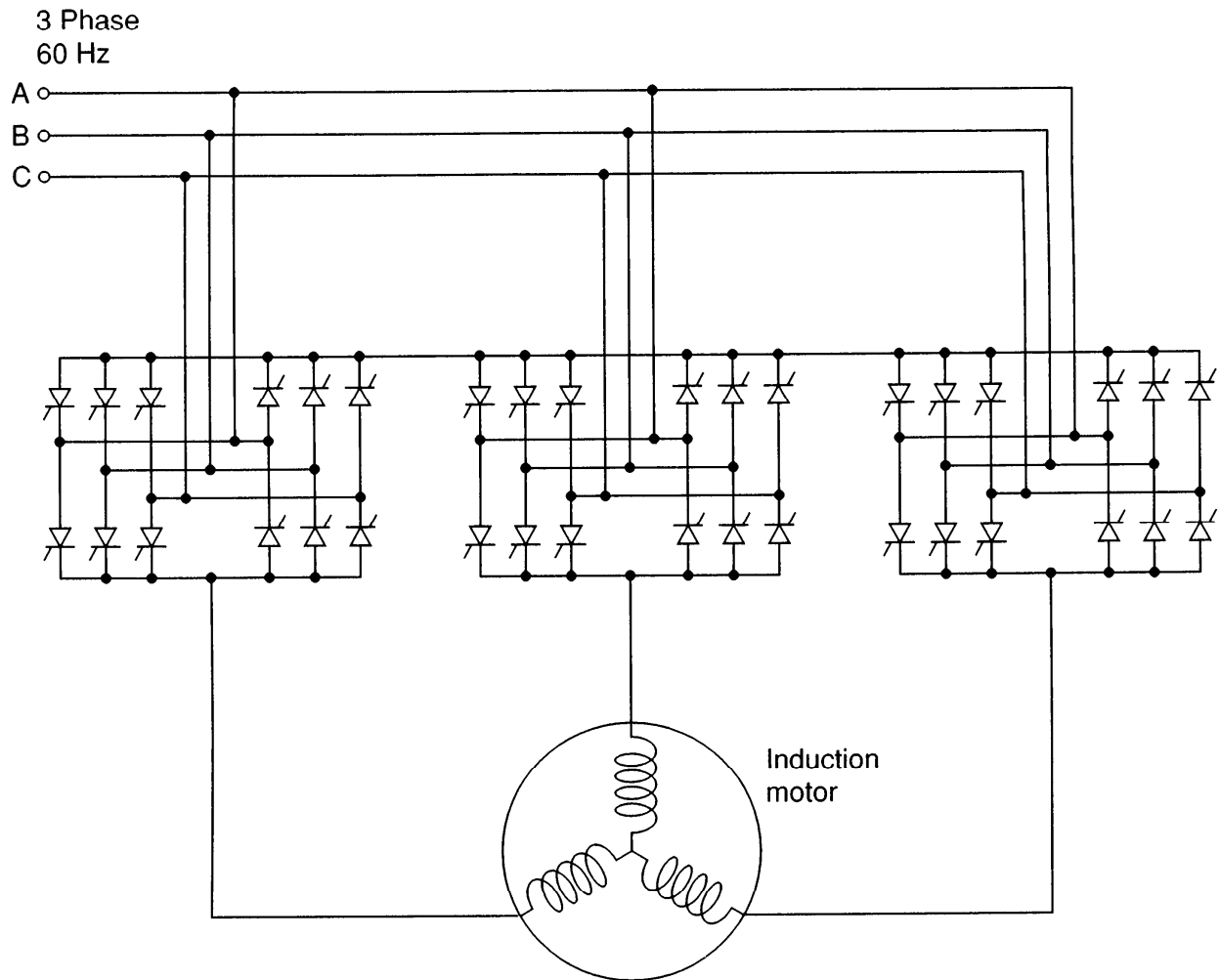
**Figure 115.9** Resonant DC-link inverter system with active voltage clamping. (Source: Rajashekara, K. S., Bhat, A. K. S., and Bose, B. K. 1993. Power electronics. In *The Electrical Engineering Handbook*, ed. R. C. Dorf, p. 708. CRC Press, Boca Raton, FL. With permission.)



## Direct AC-AC Converters

The term *direct conversion* means that the energy does not appear in any form other than the AC input or AC output. The cycloconverters are direct AC-to-AC frequency changers. The three-phase full-wave cycloconverter configuration is shown in Fig. 115.10. Each phase of the three-phase motor is supplied by two antiparallel phase-controlled six-pulse thyristor bridge converters. The output frequency is lower than the input frequency and is generally an integral multiple of the input frequency. The cycloconverter permits energy to be fed back into the utility network without any additional measures. Also, the phase sequence of the output voltage can be easily reversed by the control system. Cycloconverters have found applications in aircraft systems and industrial drives for controlling synchronous and induction motors.

**Figure 115.10** Cycloconverter control of an induction motor.



## Defining Terms

**Commutation:** Process of transferring the current from one power device to another.

**Duty cycle:** Ratio between on-time of a switch and the switching period.

**Forward voltage:** The voltage across the device when the anode is positive with respect to the cathode.

**Full-wave control:** Both the positive and negative half cycle of the waveform is controlled.

**Hysteresis control:** A method of controlling current in which the instantaneous current can vary within a band.

**Isolated:** A power electronic circuit or device having ohmic isolation between the input source

and the load circuit.

**Switching frequency:** The frequency at which the devices are turned on and turned off.

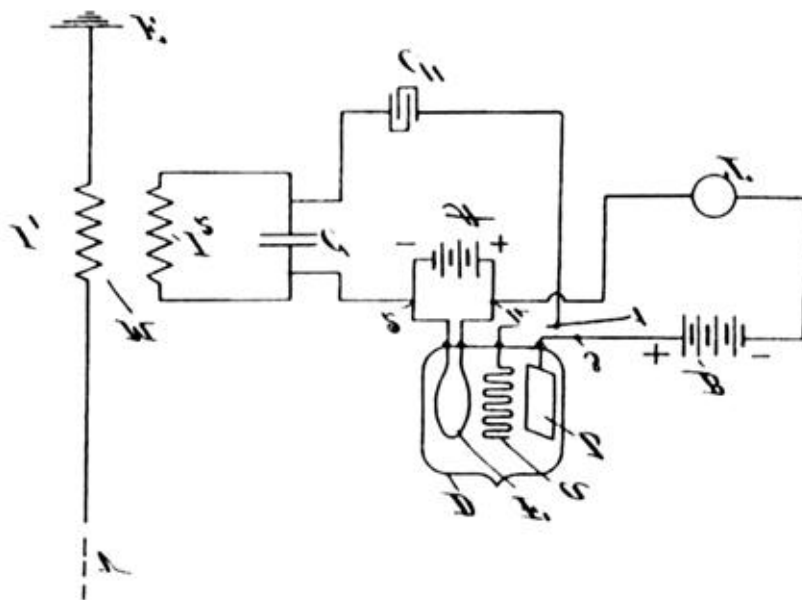
Switching frequency =  $1/(t_{\text{on}} + t_{\text{off}})$ .

## References

- Bhat, A. K. S. 1991. A unified approach for the steady-state analysis of resonant converters. *IEEE Trans. on Industrial Electron.* 38(4):251–259.
- Bose, B. K. 1992. *Modern Power Electronics*. IEEE,
- Brown, M. 1990. *Practical Switching Power Supply Design*. Academic Press,
- Rajashekara, K. S., Bhat, A. K. S., and Bose, B. K. 1993. Power electronics. *The Electrical Engineering Handbook*, ed. R. C. Dorf, pp. 694–737. CRC Press, Boca Raton, FL.
- Rashid, M. H. 1988. *Power Electronics, Circuits, Devices and Applications*. Prentice Hall, Englewood Cliffs, NJ.
- Venkataramanan, G. and Divan, D. 1990. *Pulse Width Modulation with Resonant DC Link Converter*. IEEE IAS annual meeting, pp. 984–990.

## Further Information

- Bose, B. K. 1986. *Power Electronics and AC Drives*. Prentice Hall, Englewood Cliffs, NJ.
- Mohan, N. and Undeland, T. 1989. *Power Electronics: Converters, Applications, and Design*. John Wiley & Sons, New York.
- Murphy, J. M. D. and Turnbull, F. G. 1988. *Power Electronic Control of AC Motors*. Pergamon Press, New York.
- Sen, P. C. 1981. *Thyristor DC Drives*. John Wiley & Sons, New York.
- Sum, K. K. 1988. *Recent Developments in Resonant Power Conversion*. Intertech Communications, Ventura, CA.



## SPACE TELEGRAPHY

*Lee de Forest*

*Patented February 18, 1908*

*#879,532*

An excerpt:

I have determined experimentally that the presence of the conducting member *a*, which as I stated before may be grid-shaped, increases the sensitiveness of the oscillation detector and, inasmuch as the explanation of this phenomenon is exceedingly complex and at best would be merely tentative, I do not deem it necessary to enter into a detailed statement of what I believe to be the probable explanation.

de Forest added a grid-shaped third electrode to a 2-electrode vacuum tube known as a detector, or diode. This one-way electronic "valve" could detect electrical radio signals, but de Forest's third electrode provided a means for increasing the detector's sensitivity or amplifying the signal. de Forest's Audion tube, or triode as it is now known, was the basis of nearly all electronic circuits up to the 1950's and beyond. Today, transistors and integrated circuits have replaced vacuum tubes in all but the highest-power radio and TV transmitter applications. (© 1993, DewRay Products, Inc. Used with permission.)

Hamann, J. C. "A/D and D/A Converters"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## A/D and D/A Converters

### 116.1 The Fundamentals of D/A Converters

### 116.2 The Fundamentals of A/D Converters

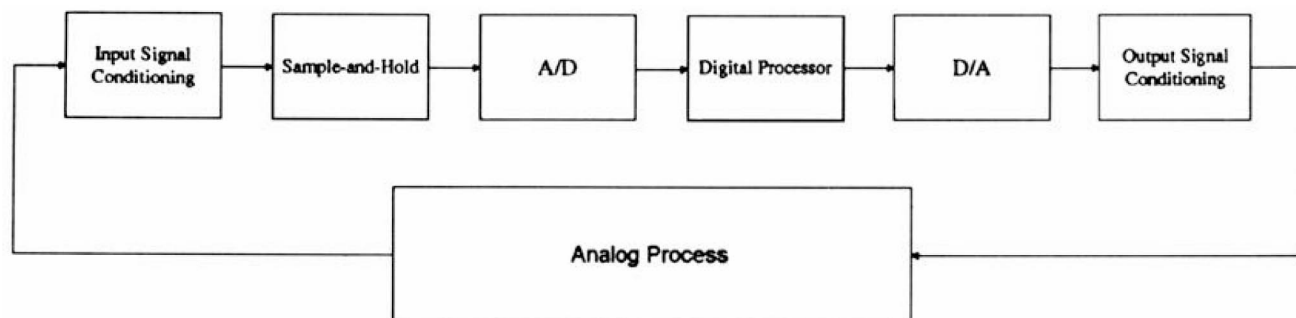
**Jerry C. Hamann**

*University of Wyoming*

*Analog-to-digital* (A/D) and *digital-to-analog* (D/A) converters provide the fundamental interface between continuous-time signals and digital-processing circuitry. These devices find wide application in consumer electronics, instrumentation, data acquisition, signal processing, communication, automatic control, and related areas. The A/D converter provides a discrete-time, discrete-valued **digital encoding** of a real-world signal, typically a voltage. The D/A converter reverses this process by producing a continuous-time, or analog, signal corresponding to a given digital encoding.

The block diagram for an example application including both A/D and D/A functions is given in Fig. 116.1. Here, the digital-processing block might be a dedicated *digital signal processor* (DSP) or a general purpose microprocessor or microcontroller that implements an application-specific signal-processing task. The associated **signal conditioning** blocks provide functionality that might include bandwidth limiting of input signal frequency content, multiplexing of multiple input and output signals, **sample-and-hold** of the input signal, and smoothing of the output analog signals.

**Figure 116.1** Example A/D and D/A application block diagram.



Various methods exist for completing the A/D and D/A conversion tasks. Trade-offs in the specification and selection of a converter typically involve parameters that are closely associated



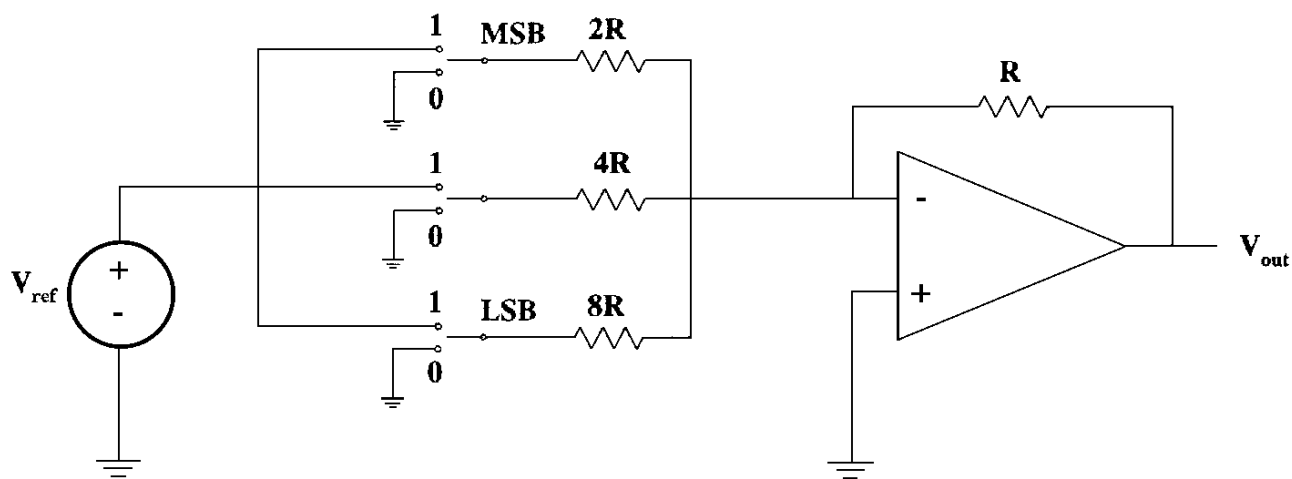
with the conversion method: *resolution* (the number of binary digits  $n$  in the conversion representation) and **conversion rate** (the time required to complete a conversion). Tables 116.1 and 116.2 provide samples of representative converter integrated circuits (ICs) and their associated parameters. Refer to the manufacturer data books listed in the chapter references for further details and additional examples. The interested reader should find in Sheingold [1986] an indispensable reference to the rich field of conversion hardware.

## 116.1 The Fundamentals of D/A Converters

The discussion here is limited to a brief introduction to the most common D/A voltage conversion method. Refer to Loriferne [1982] for a more complete coverage of D/A principles.

The basic architecture of the **multiplying D/A converter** includes a fixed or variable reference voltage source, a network of precision resistors and switches, and an analog summing and buffering stage. For example, the circuit of Fig. 116.2 demonstrates a simple 3-bit D/A. The values of the input bits control the positions of the switches, with all bits equal to 1 yielding an output voltage of  $V_{\text{out}} = -7/8V_{\text{ref}}$ . If the D/A allows the user to supply  $V_{\text{ref}}$ , the multiplying functionality becomes immediately available between input reference voltage and digital encoding.

**Figure 116.2** A simple D/A conversion circuit.



Although the scheme of Fig. 116.2 appears simple and readily generalizable, the required range of precision resistance values for a D/A converter of 8 or more bits of resolution presents an impractical realization for most IC processes. More elaborate resistor/switching arrays can be employed to generate similar functionality with a tighter bound on resistor sizes.

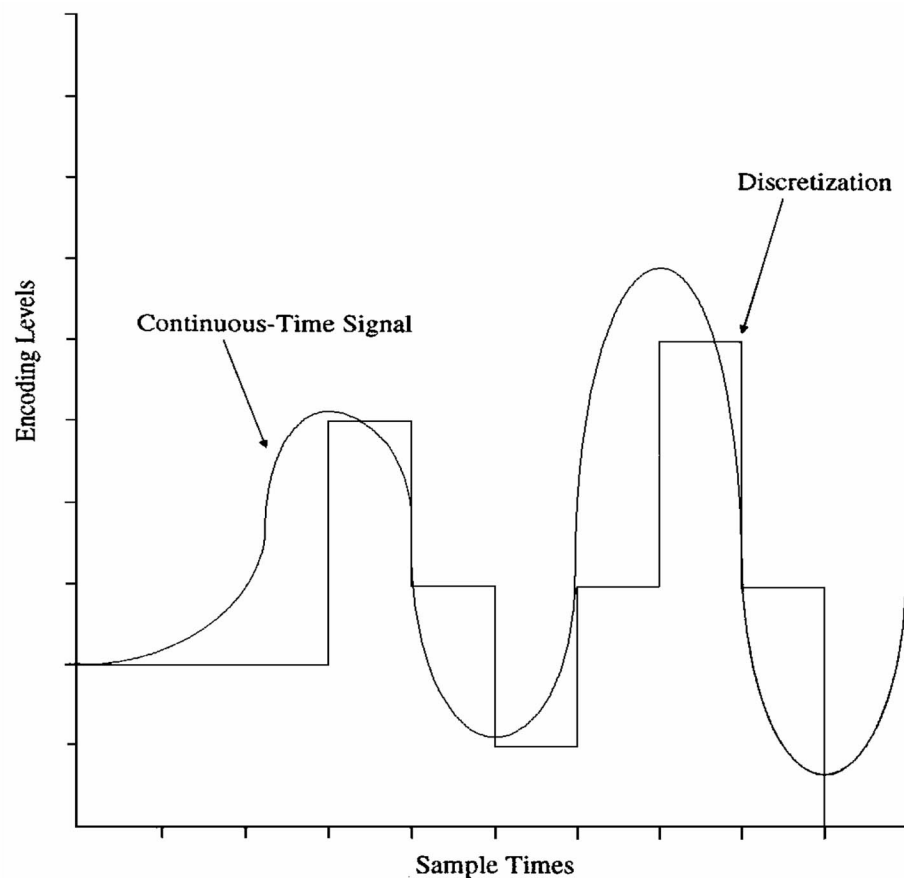
The complexity and potential asynchronous behavior of the D/A switching network introduces a potentially bothersome side effect. For example, if the input digital code switches from 011 to 100, a myriad of possible intermediate switch positions can result (000 being just one possibility). As a result the output voltage may pulse momentarily before settling down to the desired level. This behavior is typically referred to as an output **glitch**. Dedicated signal-conditioning circuitry, either internal to the D/A or supplied by the user externally, may be required to remove such undesirable characteristics.

## 116.2 The Fundamentals of A/D Converters

The brief discussion here is limited to one of the most common A/D voltage conversion methods encountered in contemporary applications. Refer to Seitzer *et al.* [1983] for an in-depth coverage of this and other A/D techniques and Demler [1991] for details of recent high-performance, high-speed technologies.

Many schemes exist for completing the A/D conversion task; **delta-sigma**, **flash**, **integrating**, and **successive approximation** compose the bulk of contemporary implementations. Although these techniques differ in the manner in which they arrive at a digital encoding of the continuous-time signal, many aspects of this quantization can be summarized via the example given in Fig. 116.3. In this plot the continuous-time signal is sampled at well-defined increments of time and represented, at each sample time and until the next sample time, by one of  $2^n$  discrete encoding levels. This finite quantization introduces an unavoidable *quantization error*; see, for example, Oppenheim and Schaffer [1989]. Nonuniformity of the intersample time intervals, as well as the quantum step size between quantization levels, introduces additional errors.

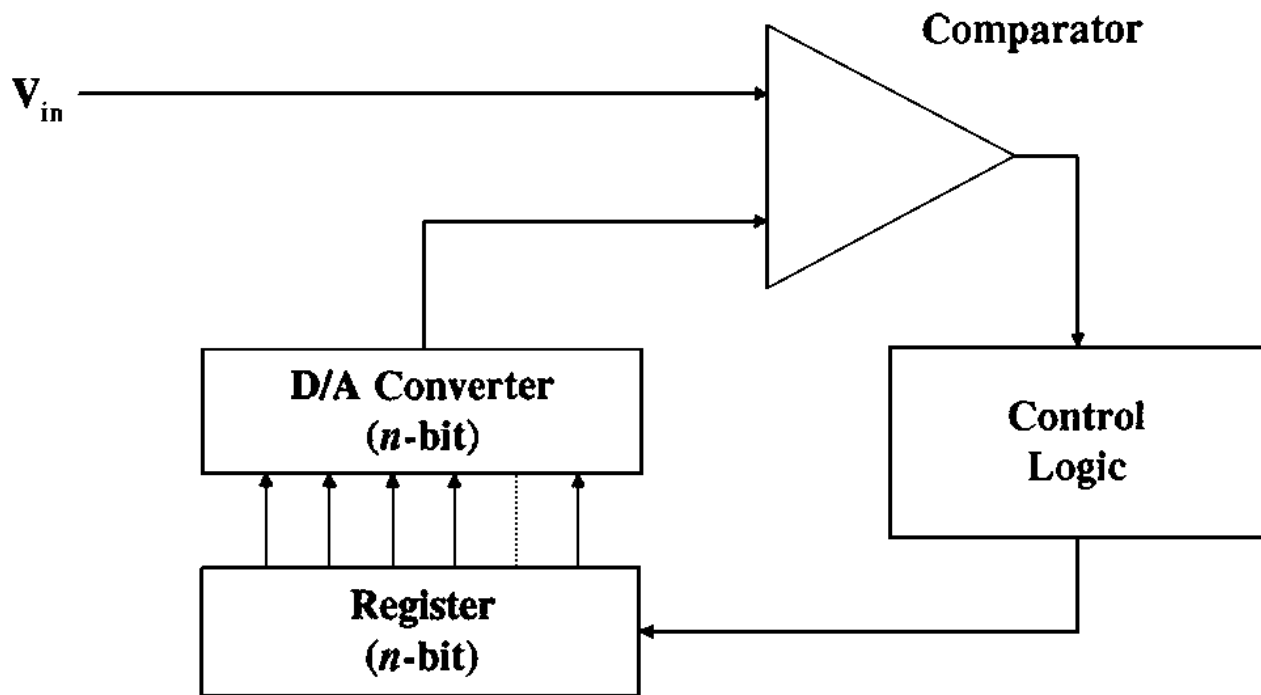
**Figure 116.3** Encoding/discretization consequences of A/D conversion.



To investigate how the A/D quantization might be undertaken, consider the block diagram of a successive approximation architecture given in Fig. 116.4. The role of the control logic is to drive an  $n$ -step iterative algorithm that successively determines the  $n$ -bit quantization from MSB (most significant bit) to LSB (least significant bit). Each new bit is assigned by adding the contribution of the bit to those of the upper bits, which are already assigned. The analog approximation of this sum

is generated by the internal D/A and fed to the input of a comparator along with the signal that is being converted. The decision of the comparator is then utilized by the control logic to either accept the contribution of the new bit or clear the bit before proceeding to the next bit of lesser significance.

**Figure 116.4** Block diagram of a successive approximation A/D.



The conversion process described in the preceding requires a constant input voltage during the approximation steps in order to reduce errors in the converted result. Sample-and-hold circuitry provides precisely this functionality.

## Defining Terms

**Conversion rate:** The time elapsed between the start and completion of the conversion task.

Typically applied in the context of A/D conversion. For D/A conversion, refer to **settling time**.

**Delta-sigma A/D:** Technique for A/D conversion wherein the analog input signal level is compared to an approximate reconstruction of the input at each sampling time. The repeated comparison results in a serial bit stream where the value of each bit denotes whether the input was greater than or less than the approximation at the given sample time. This technique typically involves a very high sampling rate that, when combined with a filtering of the low-resolution comparison results, yields a high-precision, moderate-rate conversion. This technique finds wide application in digital audio.

**Digital encoding:** A variety of binary codes are utilized in conversion hardware. Examples

include *natural binary*, *binary-coded decimal (BCD)*, *Gray code*, *sign-magnitude*, *offset binary*, *twos complement*, and *ones complement*. Some converters allow for use of one or more of these coding schemes.

**Flash A/D:** Technique for A/D conversion that utilizes  $2^n - 1$  comparators, operating in parallel, to provide an  $n$ -bit digitization. This technique provides extremely fast conversion rates, while requiring a large component count in the IC realization.

**Glitch:** When a D/A converter receives a new output request, the asynchronous nature of the switching circuitry can result in momentary spikes or anomalous output values as the converter switches to the new state described by the digital encoding. These spurious changes in the output value are referred to as *glitches*, and the circuitry employed to remove or reduce their effects is said to *deglitch* the output.

**Integrating or ramp A/D:** Technique for A/D conversion wherein the input signal drives an electronic integrator. The height that the integration achieves over a reference time period is utilized to derive the converted value of the input signal. A very popular form of this technique, utilized in many digital multimeters, is the *dual-slope integrating A/D*.

**Multiplying D/A:** A conventional architecture for D/A conversion that utilizes a fixed- or variable-input reference voltage. The resulting output voltage is effectively a scaled product of the input reference voltage and the input digital encoding. In comparison, the *fixed precision reference* conversion architecture utilizes either a precision internal or external reference signal in the multiplication.

**Sample-and-hold:** Circuitry utilized to sample a continuous-time signal at a desired time, then hold the sample constant while an accompanying A/D completes a full conversion. A variant of this idea is the *track-and-hold*, which continuously follows the input signal until instructed to hold.

**Settling time:** With reference to the D/A task, the time elapsed between the change of the input digital encoding and the settling of the output analog signal to within a stated tolerance.

**Signal conditioning:** The general term for circuitry that preconditions the analog signal entering the A/D converter and similarly filters or smooths the output of the D/A.

**Successive approximation A/D:** Technique for A/D conversion wherein a stepwise approximation of the input signal is constructed via a serial hardware comparator algorithm. The algorithm converges to an  $n$ -bit digitization of the input in  $n$  steps by successively assigning the most significant to least significant bits based upon a comparison of the input with the analog equivalent of the bits already assigned.

## References

- Analog Devices Inc. 1989. *Analog Devices Data Conversion Products Data Book*. Analog Devices Inc., Norwood, MA.
- Burr-Brown Inc. 1992. *Burr-Brown Integrated Circuits Data Book and Supplements*. Burr-Brown Inc., Tucson, AZ.
- DATEL Inc. 1991. *DATEL Data Conversion Components Catalog*. DATEL Inc., Mansfield, MA.

Demler, M. J. 1991. *High-Speed Analog-to-Digital Conversion*. Academic Press, San Diego, CA.

Loriferne, B. 1982. *Analog-Digital and Digital-Analog Conversion*. Heyden & Sons, London.

National Semiconductor Corp. 1993. *National Semiconductor Data Acquisition Databook*. National Semiconductor Corp., Santa Clara, CA.

Oppenheim, A. V. and Schaffer, R. W. 1989. *Discrete-Time Signal Processing*. Prentice Hall, Englewood Cliffs, NJ.

Seitzer, D., Pretzl, G., and Hamdy, N. A. 1983. *Electronic Analog-to-Digital Converters*. John Wiley & Sons, New York.

Sheingold, D. H. (Ed.) 1986. *Analog-Digital Conversion Handbook*, 3rd ed. Prentice Hall, Englewood Cliffs, NJ.

Texas Instruments. 1992. *Texas Instruments Linear Circuits Data Conversion, DSP Analog Interface and Video Interface Data Book*. Texas Instruments, Dallas, TX.

## Further Information

Innovative uses of A/D and D/A hardware are broadly disseminated in trade journals, including *Electrical Design News (EDN)*, *Electronic Design*, and *Computer Design*. For application of conversion hardware in digital audio arenas, refer to *AES: Journal of the Audio Engineering Society*.

Manufacturers of conversion hardware often provide detailed application notes and supporting documentation for designing with and utilizing converters. Notable among these are the extensive publications of Analog Devices Inc. and Burr-Brown.

The frontier of conversion hardware technology is growing rapidly. Contemporary techniques for device fabrication and interfacing can often be found in the *IEEE Journal of Solid-State Circuits*, *IEEE Transactions on Instrumentation and Measurement*, and *IEEE Transactions on Consumer Electronics*. For a systems and signal-processing emphasis, refer to the *IEEE Transactions on Acoustics, Sound and Signal Processing* and the *IEEE Transactions on Circuits and Systems*.

Delin, K. A., Orlando, T. P. "Superconductivity"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

This chapter is modified from Delin, K. A. and Orlando, T. P. 1993. Superconductivity. In *The Electrical Engineering Handbook*, ed. R. C. Dorf, pp. 1114–1123. CRC Press, Boca Raton, FL.

### [117.1 Introduction](#)

### [117.2 General Electromagnetic Properties](#)

### [117.3 Superconducting Electronics](#)

### [117.4 Types of Superconductors](#)

**Kevin A. Delin**

*Jet Propulsion Laboratory*

**Terry P. Orlando**

*Massachusetts Institute of Technology*

---

## **117.1 Introduction**

---

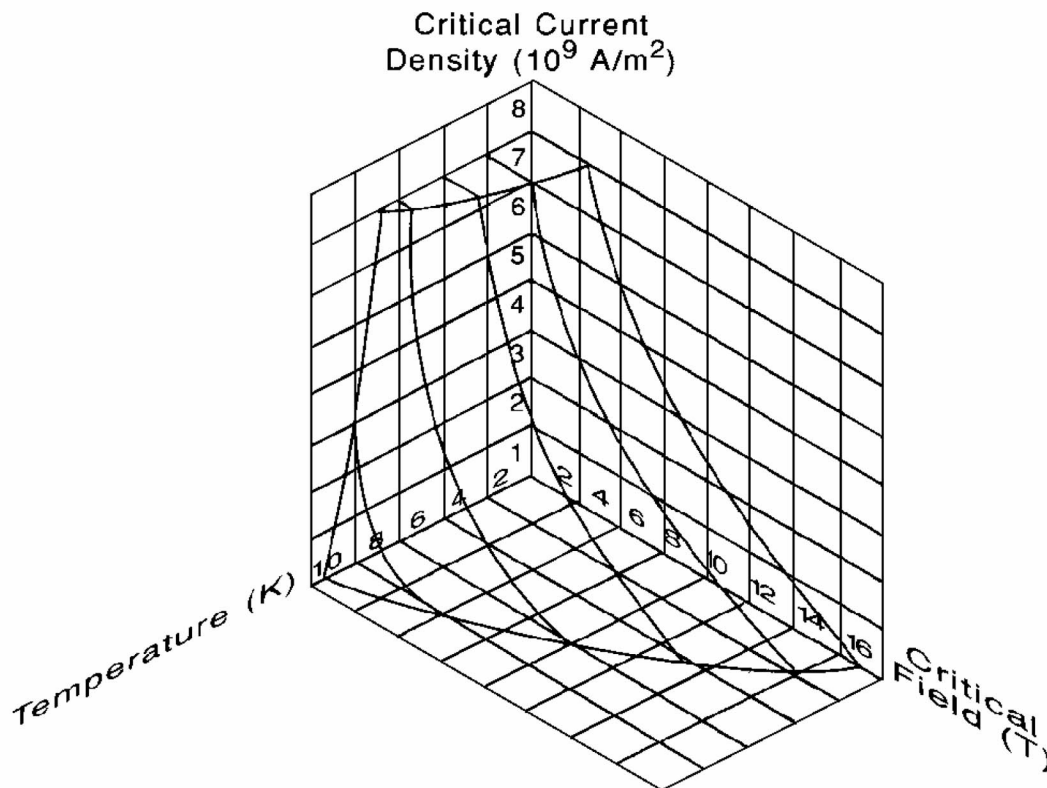
The fundamental idea behind all of a superconductor's unique properties is that **superconductivity** is a quantum mechanical phenomenon on a macroscopic scale created when the motions of individual electrons are correlated. According to the theory developed by John Bardeen, Leon Cooper, and Robert Schrieffer (BCS theory), this correlation takes place when two electrons couple to form a Cooper pair. For our purposes, we may therefore consider the electrical charge carriers in a superconductor to be Cooper pairs (or more colloquially, superelectrons) with a mass  $m^*$  and charge  $q^*$  twice those of normal electrons. The average distance between the two electrons in a Cooper pair is known as the coherence length,  $\xi$ . Both the coherence length and the binding energy of two electrons in a Cooper pair,  $2\Delta$ , depend upon the particular superconducting material. Typically, the coherence length is many times larger than the interatomic spacing of a solid, and so we should not think of Cooper pairs as tightly bound electron molecules. Instead, there are many other electrons between those of a specific Cooper pair allowing for the paired electrons to change partners on a time scale of  $\hbar/(2\Delta)$ , where  $\hbar$  is Planck's constant.

If we prevent the Cooper pairs from forming by ensuring that all the electrons are at an energy greater than the binding energy, we can destroy the superconducting phenomenon. This can be accomplished, for example, with thermal energy. In fact, according to the BCS theory, the critical temperature,  $T_c$ , associated with this energy is

$$\frac{2\Delta}{k_B T_c} \approx 3.5 \quad (117.1)$$

where  $k_B$  is Boltzmann's constant. For low critical temperature (conventional) superconductors,  $2\Delta$  is typically on the order of 1 meV, and we see that these materials must be kept below temperatures of about 10 K to exhibit their unique behavior. Superconductors with high critical temperature, in contrast, will superconduct up to temperatures of about 100 K, which is attractive from a practical view because the materials can be cooled cheaply using liquid nitrogen. A second way of increasing the energy of the electrons is electrically driving them. In other words, if the critical current density,  $J_c$ , of a superconductor is exceeded, the electrons have sufficient kinetic energy to prevent the formation of Cooper pairs. The necessary kinetic energy can also be generated through the induced currents created by an external magnetic field. As a result, if a superconductor is placed in a magnetic field larger than its critical field,  $H_c$ , it will return to its normal metallic state. To summarize, a superconductor must be maintained under the appropriate temperature, electrical current density, and magnetic field conditions to exhibit its special properties. An example of this phase space is shown in Fig. 117.1.

**Figure 117.1** The phase space for the superconducting alloy niobium-titanium. The material is superconducting inside the volume of phase space indicated. [Source: Wilson, M. 1983. *Superconducting Magnets*, Oxford University, New York. With permission.]



## 117.2 General Electromagnetic Properties

The hallmark electromagnetic properties of a superconductor are its ability to carry a static current without any resistance and its ability to exclude a static magnetic flux from its interior. It is this second property, known as the Meissner effect, that distinguishes a superconductor from merely



being a perfect conductor (which conserves the magnetic flux in its interior). Although superconductivity is a manifestly quantum mechanical phenomenon, a useful classical model can be constructed around these two properties. In this section we will outline the rationale for this classical model, which is useful in engineering applications such as waveguides and high-field magnets.

The zero DC resistance criterion implies that the superelectrons move unimpeded. The electromagnetic energy density,  $w$ , stored in a superconductor is therefore

$$w = \frac{1}{2}\varepsilon\mathbf{E}^2 + \frac{1}{2}\mu_o\mathbf{H}^2 + \frac{n^*}{2}m^*\mathbf{v}_s^2 \quad (117.2)$$

where the first two terms are the familiar electric and magnetic energy densities, respectively. (Our electromagnetic notation is standard:  $\varepsilon$  is the permittivity,  $\mu_o$  is the permeability,  $\mathbf{E}$  is the electric field, and the magnetic flux density,  $\mathbf{B}$ , is related to the magnetic field,  $\mathbf{H}$ , via the constitutive law  $\mathbf{B} = \mu_o\mathbf{H}$ .) The last term represents the kinetic energy associated with the undamped superelectrons' motion ( $n^*$  and  $\mathbf{v}_s$  are the superelectrons' density and velocity, respectively). Because the supercurrent density,  $\mathbf{J}_s$ , is related to the superelectron velocity by  $\mathbf{J}_s = n^*q^*\mathbf{v}_s$ , the kinetic energy term can be rewritten

$$n^* \left( \frac{1}{2}m^*\mathbf{v}_s^2 \right) = \frac{1}{2}\Lambda\mathbf{J}_s^2 \quad (117.3)$$

where  $\Lambda$  is defined as

$$\Lambda = \frac{m^*}{n^*(q^*)^2} \quad (117.4)$$

Assuming that all the charge carriers are superelectrons, there is no power dissipation inside the superconductor, and so Poynting's theorem over a volume  $V$  may be written

$$-\int_V \nabla \cdot (\mathbf{E} \times \mathbf{H}) dv = \int_V \frac{\partial w}{\partial t} dv \quad (117.5)$$

where the left side of the expression is the power flowing into the region. By taking the time derivative of the energy density and appealing to Faraday's and Ampère's laws to find the time derivatives of the field quantities, we find that the only way for Poynting's theorem to be satisfied is if

$$\mathbf{E} = \frac{\partial}{\partial t}(\Lambda\mathbf{J}_s) \quad (117.6)$$

This relation, known as the *first London equation* (after the London brothers, Heinz and Fritz), is thus necessary if the superelectrons have no resistance to their motion.

Equation (117.6) also reveals that the superelectrons' inertia creates a lag between their motion and that of the electric field. As a result, a superconductor can support a time-varying voltage drop across itself. The impedance associated with the supercurrent, therefore, is an inductor, and it will be useful to think of  $\Lambda$  as an inductance created by the correlated motion of the superelectrons.

If the first London equation is substituted into Faraday's law,  $\nabla \times \mathbf{E} = -(\partial \mathbf{B} / \partial t)$ , and integrated with respect to time, the *second London equation* results:

$$\nabla \times (\Lambda \mathbf{J}_s) = -\mathbf{B} \quad (117.7)$$

where the constant of integration has been defined to be zero. This choice is made so that the second London equation is consistent with the Meissner effect, as we now demonstrate. Taking the curl of the quasi-static form of Ampère's law,  $\nabla \times \mathbf{H} = \mathbf{J}_s$ , results in the expression  $\nabla^2 \mathbf{B} = -\mu_o \nabla \times \mathbf{J}_s$ , where a vector identity,  $\nabla \times \nabla \times \mathbf{C} = \nabla(\nabla \cdot \mathbf{C}) - \nabla^2 \mathbf{C}$ ; the constitutive relation,  $\mathbf{B} = \mu_o \mathbf{H}$ ; and Gauss's law,  $\nabla \cdot \mathbf{B} = 0$ , have been used. By now appealing to the second London equation, we obtain the vector Helmholtz equation

$$\nabla^2 \mathbf{B} - \frac{1}{\lambda^2} \mathbf{B} = 0 \quad (117.8)$$

where the penetration depth is defined as

$$\lambda \equiv \sqrt{\frac{\Lambda}{\mu_o}} = \sqrt{\frac{m^*}{n^* (q^*)^2 \mu_o}} \quad (117.9)$$

From Eq. (117.8) we find that a flux density applied parallel to the surface of a semi-infinite superconductor will decay away exponentially from the surface on a spatial length scale of order  $\lambda$ . In other words, a bulk superconductor will exclude an applied flux as predicted by the Meissner effect.

The London equations reveal that there is a characteristic length  $\lambda$  over which electromagnetic fields can change inside a superconductor. This penetration depth is different from the more familiar skin depth of electromagnetic theory, the latter being a frequency-dependent quantity. Indeed, the penetration depth at zero temperature is a distinct material property of a particular superconductor.

Notice that  $\lambda$  is sensitive to the number of correlated electrons (the superelectrons) in the material. As previously discussed, this number is a function of temperature, and so only at  $T = 0$  do *all* the electrons that usually conduct ohmically participate in the Cooper pairing. For intermediate temperatures,  $0 < T < T_c$ , there are actually two sets of interpenetrating electron fluids: the uncorrelated electrons providing ohmic conduction and the correlated ones creating supercurrents. This two-fluid model is a useful way to build temperature effects into the London relations.

Under the two-fluid model, the electrical current density,  $\mathbf{J}$ , is carried by both the uncorrelated (normal) electrons and the superelectrons:  $\mathbf{J} = \mathbf{J}_n + \mathbf{J}_s$ , where  $\mathbf{J}_n$  is the normal current density. The two channels are modeled in a circuit, as shown in [Fig. 117.2](#), by a parallel combination of a resistor (representing the ohmic channel) and an inductor (representing the superconducting channel). To a good approximation, the respective temperature dependences of the conductor and inductor are

$$\tilde{\sigma}_o(T) = \sigma_o(T_c) \left( \frac{T}{T_c} \right)^4 \quad \text{for } T \leq T_c \quad (117.10)$$

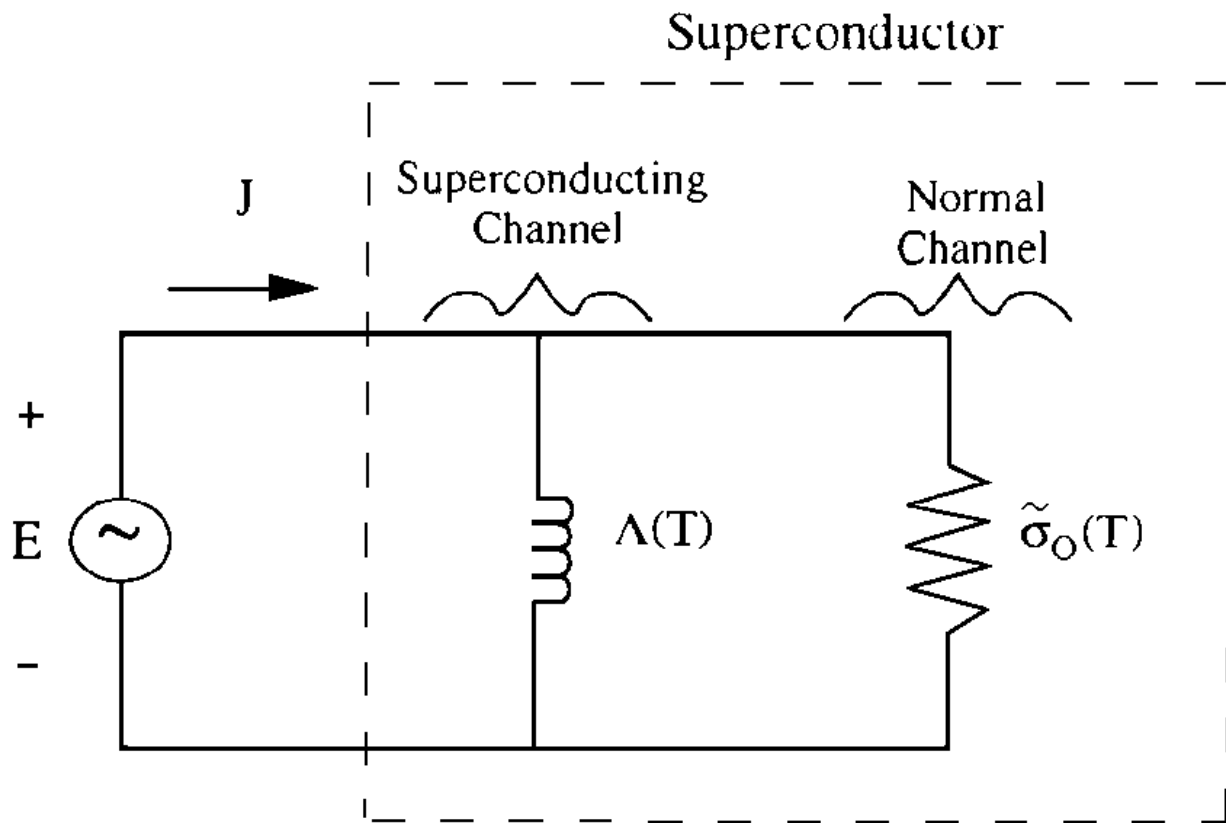
$$\Lambda(T) = \Lambda(0) \left( \frac{1}{1 - (T/T_c)^4} \right) \quad \text{for } T \leq T_c \quad (117.11)$$

where  $\sigma_o$  is the DC conductance of the normal channel. (Strictly speaking, the normal channel should also contain an inductance representing the inertia of the normal electrons, but typically such an inductor contributes negligibly to the overall electrical response.) Since the temperature-dependent penetration depth is defined as  $\lambda(T) = \sqrt{\Lambda(T)/\mu_o}$ , the effective conductance of a superconductor in the sinusoidal steady state is

$$\sigma = \tilde{\sigma}_o + \frac{1}{j\omega\mu_o\lambda^2} \quad (117.12)$$

where the explicit temperature dependence notation has been suppressed.

**Figure 117.2** A lumped element model of a superconductor.

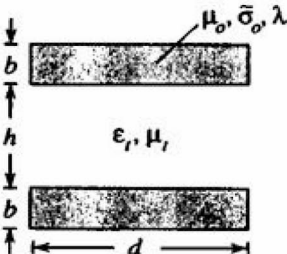
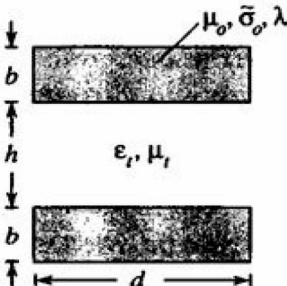
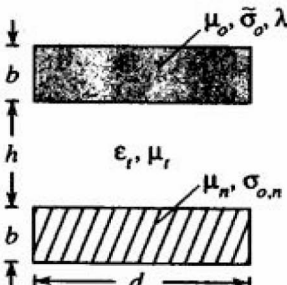


Most of the important physics associated with the classical model is embedded in Eq. (117.12). As is clear from the lumped element model, the relative importance of the normal and superconducting channels is a function not only of temperature but also of frequency. The familiar  $L/R$  time constant, here equal to  $\Lambda\tilde{\sigma}_o$ , delineates the frequency regimes where most of the total current is carried by  $J_n$  (if  $\omega\Lambda\tilde{\sigma}_o \gg 1$ ) or  $J_s$  (if  $\omega\Lambda\tilde{\sigma}_o \ll 1$ ). This same result can also be obtained by comparing the skin depth associated with the normal channel,  $\delta = \sqrt{2/(\omega\mu_o\tilde{\sigma}_o)}$ , to the penetration depth to see which channel provides more field screening. In addition, it is straightforward to use Eq. (117.12) to rederive Poynting's theorem for systems that involve superconducting materials:

$$\begin{aligned}
-\int_V \nabla \cdot (\mathbf{E} \times \mathbf{H}) dv &= \frac{d}{dt} \int_V \left( \frac{1}{2} \epsilon \mathbf{E}^2 + \frac{1}{2} \mu_o \mathbf{H}^2 + \frac{1}{2} \Lambda(T) \mathbf{J}_s^2 \right) dv \\
&+ \int_V \frac{1}{\tilde{\sigma}_o(T)} \mathbf{J}_n^2 dv
\end{aligned} \tag{117.13}$$

Using this expression, it is possible to apply the usual electromagnetic analysis to find the inductance ( $L_o$ ), capacitance ( $C_o$ ), and resistance ( $R_o$ ) per unit length along a parallel plate transmission line. The results of such analysis for typical cases are summarized in [Table 117.1](#).

**Table 117.1** Lumped Circuit Element Parameters Per Unit Length for Typical Transverse Electromagnetic Parallel Plate Waveguides\*

Transmission Line Geometry	$L_o$	$C_o$	$R_o$
 <p>Two identical, thin (<math>\lambda \gg b</math>) superconducting plates</p>	$\frac{\mu_t h}{d} + \frac{2\mu_o \lambda^2}{db}$	$\frac{\epsilon_t d}{h}$	$\frac{8}{db\tilde{\sigma}_o} \left( \frac{\lambda}{\delta} \right)^4$
 <p>Two identical, thick (<math>\lambda \ll b</math>) superconducting plates</p>	$\frac{\mu_t h}{d} + \frac{2\mu_o \lambda}{d}$	$\frac{\epsilon_t d}{h}$	$\frac{4}{d\delta\tilde{\sigma}_o} \left( \frac{\lambda}{\delta} \right)^3$
 <p>One thick (<math>\lambda \ll b</math>) superconducting plate and one thick (<math>\delta_n \ll b</math>) ohmic plate</p>	$\frac{\mu_t h}{d} + \frac{\mu_o \lambda}{d} + \frac{\mu_n \delta_n}{2d}$	$\frac{\epsilon_t d}{h}$	$\frac{1}{d\delta_n \sigma_{o,n}}$

\*The subscript  $n$  refers to parameters associated with a normal (ohmic) plate. Using these expressions, line input impedance, attenuation, and wave velocity can be calculated.

Source: Orlando, T. P. and Delin, K. A. *Foundations of Applied Superconductivity*, p. 171. Addison-Wesley, Reading, MA. With permission.

## 117.3 Superconducting Electronics

The macroscopic quantum nature of superconductivity can be usefully exploited to create a new type of electronic device. Because all the superelectrons exhibit correlated motion, the usual wave-particle duality normally associated with a single quantum particle can now be applied to the entire ensemble of superelectrons. Thus, there is a spatiotemporal phase associated with the ensemble that characterizes the supercurrent flowing in the material.

Naturally, if the overall electron correlation is broken, this phase is lost and the material is no longer a superconductor. There is a broad class of structures, however, known as *weak links*, where the correlation is merely perturbed locally in space rather than outright destroyed. Colloquially, we say that the phase "slips" across the weak link to acknowledge the perturbation.

The unusual properties of this phase slippage were first investigated by Brian Josephson and constitute the central principles behind superconducting electronics. Josephson found that the phase slippage could be defined as the difference between the macroscopic phases on either side of the weak link. This phase difference, denoted as  $\phi$ , determined the supercurrent,  $i_s$ , through and voltage,  $v$ , across the weak link according to the Josephson equations,

$$i_s = I_c \sin \phi \quad (117.14)$$

$$v = \frac{\Phi_o}{2\pi} \frac{\partial \phi}{\partial t} \quad (117.15)$$

where  $I_c$  is the critical (maximum) current of the junction and  $\Phi_o$  is the quantum unit of flux. (The flux quantum has a precise definition in terms of Planck's constant,  $h$ , and the electron charge,  $e$ :  $\Phi_o \equiv h/(2e) \approx 2.068 \times 10^{-15}$  Wb). As in the previous section, the correlated motion of the electrons, here represented by the superelectron phase, manifests itself through an inductance. This is straightforwardly demonstrated by taking the time derivative of Eq. (117.14) and combining this expression with Eq. (117.15). Although the resulting inductance is nonlinear (it depends on  $\cos \phi$ ), its relative scale is determined by

$$L_j = \frac{\Phi_o}{2\pi I_c} \quad (117.16)$$

a useful quantity for making engineering estimates.

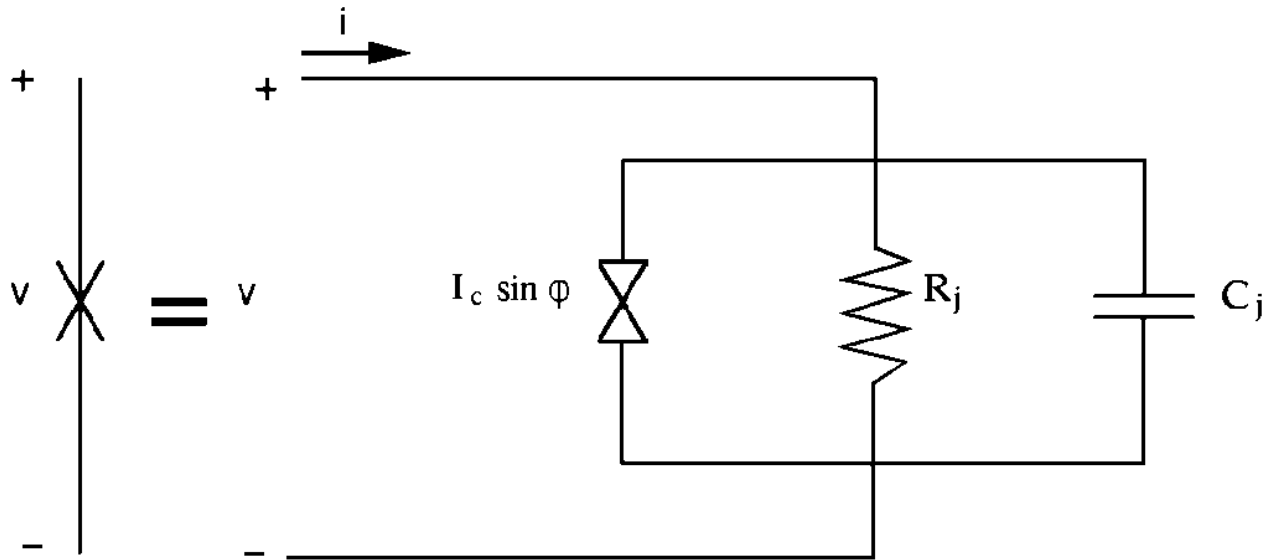
A common weak link, known as the Josephson tunnel junction, is made by separating two superconducting films with a very thin (typically 20 Å) insulating layer. Such a structure is conveniently analyzed using the resistively and capacitively shunted junction (RCSJ) model shown in Fig. 117.3. Under the RCSJ model an ideal lumped junction [described by Eqs. (117.14) and

(117.15)] and a resistor  $R_j$  represent how the weak link structure influences the respective phases of the super and normal electrons, and a capacitor  $C_j$  represents the physical capacitance of the sandwich structure. If the ideal lumped junction portion of the circuit is treated as an inductor-like element, many Josephson tunnel junction properties can be calculated with the familiar circuit time constants associated with the model. For example, the quality factor  $Q$  of the RCSJ circuit can be expressed as

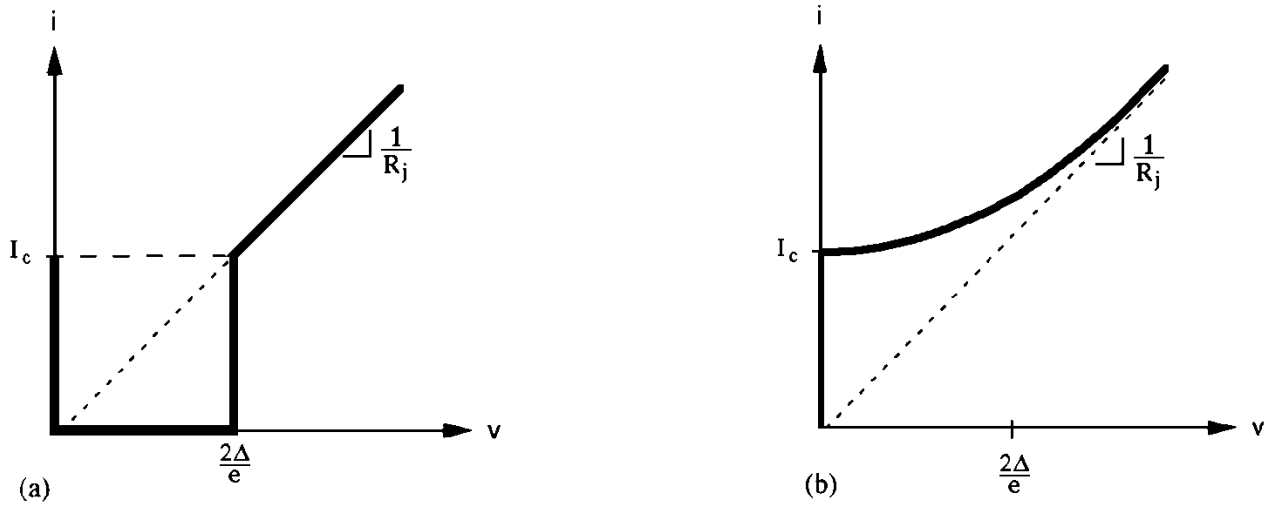
$$Q^2 = \frac{R_j C_j}{L_j / R_j} = \frac{2\pi I_c R_j^2 C_j}{\Phi_o} \equiv \beta \quad (117.17)$$

where  $\beta$  is known as the Stewart-McCumber parameter. Clearly, if  $\beta \gg 1$ , the ideal lumped junction element is underdamped in that the capacitor readily charges up, dominates the overall response of the circuit, and therefore creates a hysteretic  $i$ - $v$  curve as shown in Fig. 117.4(a). In the case when the bias current is raised from zero, no time-averaged voltage is created until the critical current is exceeded. At this point the junction switches to the voltage  $2\Delta/e$  with a time constant  $\sqrt{L_j C_j}$ . Once the junction has latched into the voltage state, however, the bias current must be lowered to zero before it can again be steered through the superconducting path. Conversely,  $\beta \ll 1$  implies that the  $L_j / R_j$  time constant dominates the circuit response, so that the capacitor does not charge up and the  $i$ - $v$  curve is not hysteretic [Fig. 117.4(b)].

**Figure 117.3** A real Josephson tunnel junction can be modeled using ideal lumped circuit elements.



**Figure 117.4** The  $i$ - $v$  curves for a Josephson junction: (a)  $\beta \gg 1$ , (b)  $\beta \ll 1$ .



Just as the correlated motion of the superelectrons creates the frequency-independent Meissner effect in a bulk superconductor through Faraday's law, so too the macroscopic quantum nature of superconductivity allows the possibility of a device whose output voltage is a function of a static magnetic field. If two weak links are connected in parallel, the lumped version of Faraday's law gives the voltage across the second weak link as  $v_2 = v_1 + (d\Phi/dt)$ , where  $\Phi$  is the total flux threading the loop between the links. Substituting Eq. (117.15) and integrating with respect to time yields

$$\phi_2 - \phi_1 = (2\pi\Phi)/\Phi_o \quad (117.18)$$

showing that the spatial change in the phase of the macroscopic wavefunction is proportional to the local magnetic flux. The structure described is known as a *superconducting quantum interference device (SQUID)* and can be used as a highly sensitive magnetometer by biasing it with current and measuring the resulting voltage as a function of magnetic flux. From this discussion, it is apparent that a duality exists in how fields interact with the macroscopic phase: electric fields are coupled to its rate of change in time and magnetic fields are coupled to its rate of change in space.

## 117.4 Types of Superconductors

The macroscopic quantum nature of superconductivity also affects the general electromagnetic properties previously discussed. This is most clearly illustrated by the interplay of the characteristic lengths  $\xi$ , representing the scale of quantum correlations, and  $\lambda$ , representing the scale of electromagnetic screening. Consider the scenario where a magnetic field,  $H$ , is applied parallel to the surface of a semi-infinite superconductor. The correlations of the electrons in the superconductor must lower the overall energy of the system or else the material would not be superconducting in the first place. Because the critical magnetic field  $H_c$  destroys all the correlations, it is convenient to define the energy density gained by the system in the superconducting state as  $(\frac{1}{2})\mu_o H_c^2$ . The electrons in a Cooper pair are separated on a length scale

of  $\xi$ , however, and so the correlations cannot be fully achieved until a distance roughly  $\xi$  from the boundary of the superconductor. There is thus an energy per unit area,  $(\frac{1}{2})\mu_o H_c^2 \xi$ , that is lost because of the presence of the boundary. Now consider the effects of the applied magnetic field on this system. It costs the superconductor energy to maintain the Meissner effect,  $B = 0$ , in its bulk; in fact, the energy density required is  $(\frac{1}{2})\mu_o H^2$ . However, since the field can penetrate the superconductor a distance roughly  $\lambda$ , the system need not expend an energy per unit area of  $(\frac{1}{2})\mu_o H^2 \lambda$  to screen over this volume. To summarize, more than a distance  $\xi$  from the boundary, the energy of the material is lowered (because it is superconducting), and more than a distance  $\lambda$  from the boundary the energy of the material is raised (to shield the applied field).

Now, if  $\lambda < \xi$ , the region of superconducting material greater than  $\lambda$  from the boundary but less than  $\xi$  will be higher in energy than that in the bulk of the material. Thus, the surface energy of the boundary is positive and so costs the total system some energy. This class of superconductors is known as type I. Most elemental superconductors, such as aluminum, tin, and lead, are type I. In addition to having  $\lambda < \xi$ , type I superconductors are generally characterized by low critical temperatures ( $\sim 5$  K) and critical fields ( $\sim 0.05$  T). Typical type I superconductors and their properties are listed in [Table 117.2](#).

**Table 117.2** Material Parameters for Type I Superconductors\*

Material	$T_c$ (K)	$\lambda_o$ (nm)	$\xi_o$ (nm)	$\Delta_o$ (meV)	$\mu_o H_{co}$ (mT)
Al	1.18	50	1600	0.18	10.5
In	3.41	65	360	0.54	23.0
Sn	3.72	50	230	0.59	30.5
Pb	7.20	40	90	1.35	80.0
Nb	9.25	85	40	1.50	198.0

\*The penetration depth  $\lambda_o$  is given at zero temperature, as are the coherence length  $\xi_o$ , the thermodynamic critical field  $H_{co}$ , and the energy gap  $\Delta_o$ .

Source: Donnelly, R. J. 1981. Cryogenics. In *Physics Vade Mecum*, ed. H. L. Anderson. American Institute of Physics, New York. With permission.

Conversely, if  $\lambda > \xi$ , the surface energy associated with the boundary is negative and lowers the total system energy. It is therefore thermodynamically favorable for a normal–superconducting interface to form inside these type II materials. Consequently, this class of superconductors does not exhibit the simple Meissner effect as do type I materials. Instead, there are now two critical fields: for applied fields below the lower critical field,  $H_{c1}$ , a type II superconductor is in the Meissner state, and for applied fields greater than the upper critical field,  $H_{c2}$ , superconductivity is destroyed. The three critical fields are related to each other by  $H_c \approx \sqrt{H_{c1} H_{c2}}$ .

In the range  $H_{c1} < H < H_{c2}$ , a type II superconductor is said to be in the vortex state because now the applied field can enter the bulk superconductor. Because flux exists in the material, however, the superconductivity is destroyed locally, creating normal regions. Recall that for type II materials the boundary between the normal and superconducting regions lowers the overall energy of the system. Therefore, the flux in the superconductor creates as many normal–superconducting interfaces as possible without violating quantum criteria. The net result is that flux enters a type II superconductor in quantized bundles of magnitude  $\Phi_o$  known as *vortices* or *fluxons* (the former



name derives from the fact that current flows around each quantized bundle in the same manner as a fluid vortex circulates around a drain). The central portion of a vortex, known as the core, is a normal region with an approximate radius of  $\xi$ . If a defect-free superconductor is placed in a magnetic field, the individual vortices, whose cores essentially follow the local average field lines, form an ordered triangular array, or flux lattice. As the applied field is raised beyond  $H_{c1}$  (where the first vortex enters the superconductor), the distance between adjacent vortex cores decreases to maintain the appropriate flux density in the material. Finally, the upper critical field is reached when the normal cores overlap and the material is no longer superconducting. Indeed, a precise calculation of  $H_{c2}$  using the phenomenological theory developed by Vitaly Ginzburg and Lev Landau yields

$$H_{c2} = \frac{\Phi_o}{2\pi\mu_o\xi^2} \quad (117.19)$$

which verifies our simple picture. The values of typical type II material parameters are listed in [Tables 117.3](#) and [117.4](#).

**Table 117.3** Material Parameters for Conventional Type II Superconductors\*

Material	$T_c$ (K)	$\lambda_{GL}(0)$ (nm)	$\xi_{GL}(0)$ (nm)	$\Delta_o$ (meV)	$\mu_o H_{c2,o}$ (T)
Pb-In	7.0	150	30	1.2	0.2
Pb-Bi	8.3	200	20	1.7	0.5
Nb-Ti	9.5	300	4	1.5	13
Nb-N	16	200	5	2.4	15
PbMo <sub>6</sub> S <sub>8</sub>	15	200	2	2.4	60
V <sub>3</sub> Ga	15	90	2–3	2.3	23
V <sub>3</sub> Si	16	60	3	2.3	20
Nb <sub>3</sub> Sn	18	65	3	3.4	23
Nb <sub>3</sub> Ge	23	90	3	3.7	38

\*The values are only representative because the parameters for alloys and compounds depend on how the material is fabricated. The penetration depth  $\lambda_{GL}(0)$  is given as the coefficient of the Ginzburg-Landau temperature dependence as  $\lambda_{GL}(T) = \lambda_{GL}(0)(1 - T/T_c)^{-1/2}$ ; likewise for the coherence length where  $\xi_{GL}(T) = \xi_{GL}(0)(1 - T/T_c)^{-1/2}$ . The upper critical field  $H_{c2,o}$  is given at zero temperature as well as the energy gap  $\Delta_o$ .

Source: Donnelly, R. J. 1981. Cryogenics. In *Physics Vade Mecum*, ed. H. L. Anderson. American Institute of Physics, New York. With permission.

**Table 117.4** Type II (High-Temperature Superconductors)

Material	$T_c$ (K)	$\lambda_{a,b}$ (nm)	$\lambda_c$ (nm)	$\xi_{a,b}$ (nm)	$\xi_c$ (nm)
LuNi <sub>2</sub> B <sub>2</sub> C	17	71		6	
Rb <sub>3</sub> C <sub>60</sub>	33	300		3	
YBa <sub>2</sub> Cu <sub>3</sub> O <sub>7</sub>	95	150	1500	3	0.2
Bi <sub>2</sub> Sr <sub>2</sub> CaCu <sub>2</sub> O <sub>8</sub>	85	25	500	4	0.2
Bi <sub>2</sub> Sr <sub>2</sub> Ca <sub>2</sub> Cu <sub>3</sub> O <sub>10</sub>	110				
TlBa <sub>2</sub> Ca <sub>2</sub> Cu <sub>3</sub> O <sub>10</sub>	125				
HgBaCaCu <sub>2</sub> O <sub>6</sub>	115	150		2.5	
HgBa <sub>2</sub> Ca <sub>2</sub> Cu <sub>3</sub> O <sub>8</sub>	135				

Type II superconductors are of great technical importance because typical  $H_{c2}$  values are at least an order of magnitude greater than the typical  $H_c$  values of type I materials. It is therefore possible to use type II materials to make high-field magnet wire. Unfortunately, when current is applied to the wire, there is a Lorentz-like force on the vortices, causing them to move. Because the moving vortices carry flux, their motion creates a static voltage drop along the superconducting wire by Faraday's law. As a result, the wire no longer has a zero DC resistance, even though the material is still superconducting. To fix this problem, type II superconductors are usually fabricated with intentional defects, such as impurities or grain boundaries, in their crystalline structure to pin the vortices and prevent vortex motion. The pinning is created because the defect locally weakens the superconductivity in the material, and it is thus energetically favorable for the normal core of the vortex to overlap the nonsuperconducting region in the material. Critical current densities usually quoted for practical type II materials, therefore, really represent the depinning critical current density where the Lorentz-like force can overcome the pinning force. (The depinning critical current density should not be confused with the depairing critical current density, which represents the current when the Cooper pairs have enough kinetic energy to overcome their correlation. The depinning critical current density is typically an order of magnitude less than the depairing critical current density, the latter of which represents the theoretical maximum for  $J_c$ .)

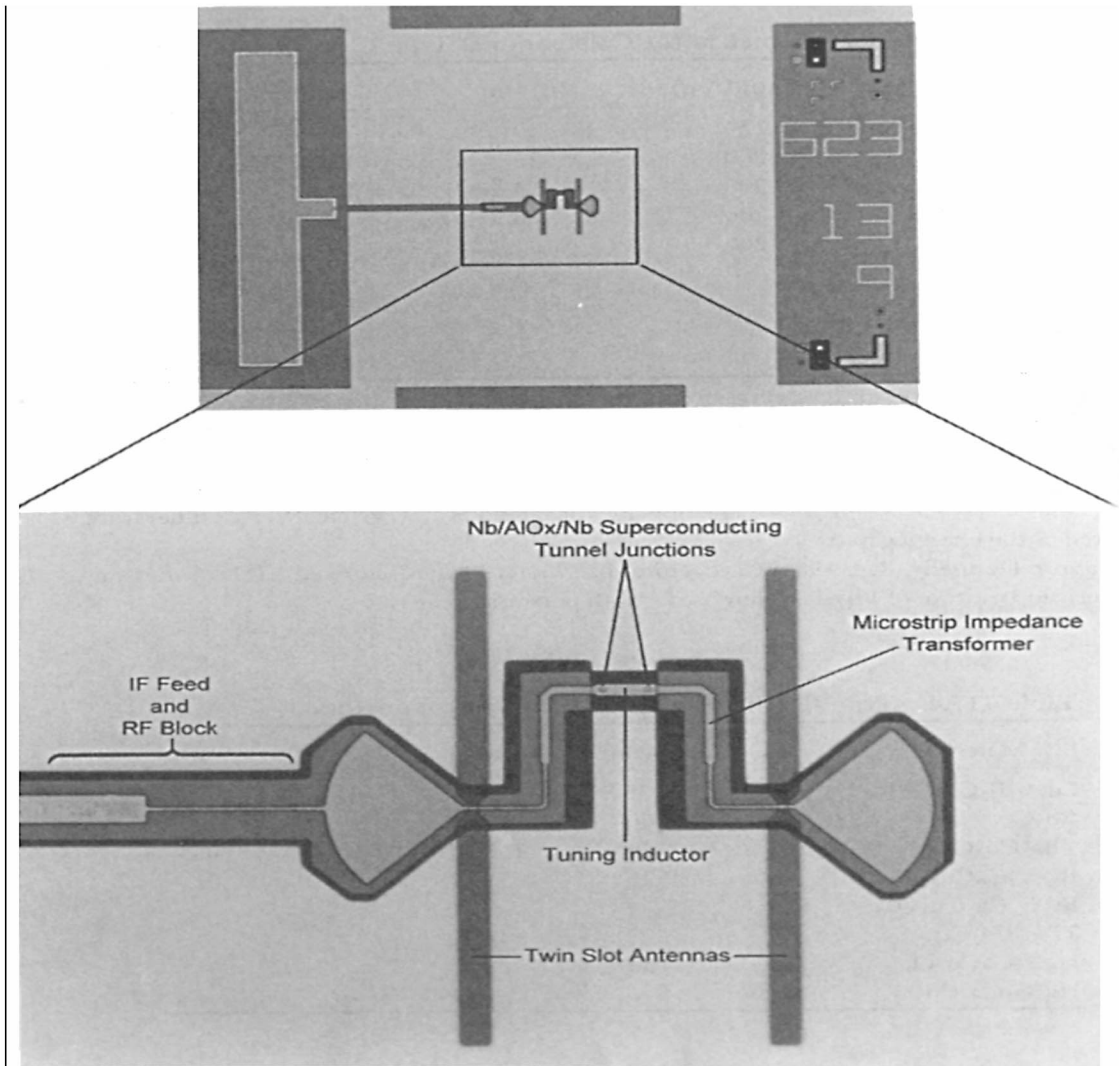
By careful manufacturing, it is possible to make superconducting wire with tremendous amounts of current-carrying capacity. For example, standard copper wire used in homes will carry about  $10^7$  A/m<sup>2</sup>, whereas a practical type II superconductor like niobium-titanium can carry current densities of  $10^{10}$  A/m<sup>2</sup> or higher even in fields of several teslas. This property, more than a zero DC resistance, is what makes superconducting wire so desirable.

## Defining Terms

**Superconductivity:** A state of matter whereby the correlation of conduction electrons allows a static current to pass without resistance and a static magnetic flux to be excluded from the bulk of the material.

## References

- Donnelly, R. J. 1981. Cryogenics. In *Physics Vade Mecum*, ed. H. L. Anderson. American Institute of Physics, New York.
- Foner, S. and Schwartz, B. B. 1974. *Superconducting Machines and Devices*. Plenum Press, New York.
- Foner, S. and Schwartz, B. B. 1981. *Superconducting Materials Science*. Plenum Press, New York.
- Orlando, T. P. and Delin, K. A. 1991. *Foundations of Applied Superconductivity*. Addison-Wesley, Reading, MA.
- Ruggiero, S. T. and Rudman, D. A. 1990. *Superconducting Devices*. Academic Press, Boston, MA.
- Schwartz, B. B. and Foner, S. 1977. *Superconducting Applications: SQUIDS and Machines*. Plenum Press, New York.
- Van Duzer, T. and Turner, C. W. 1981. *Principles of Superconductive Devices and Circuits*. Elsevier North Holland, New York.
- Wilson, M. N. 1983. *Superconducting Magnets*. Oxford University Press, Oxford, UK.



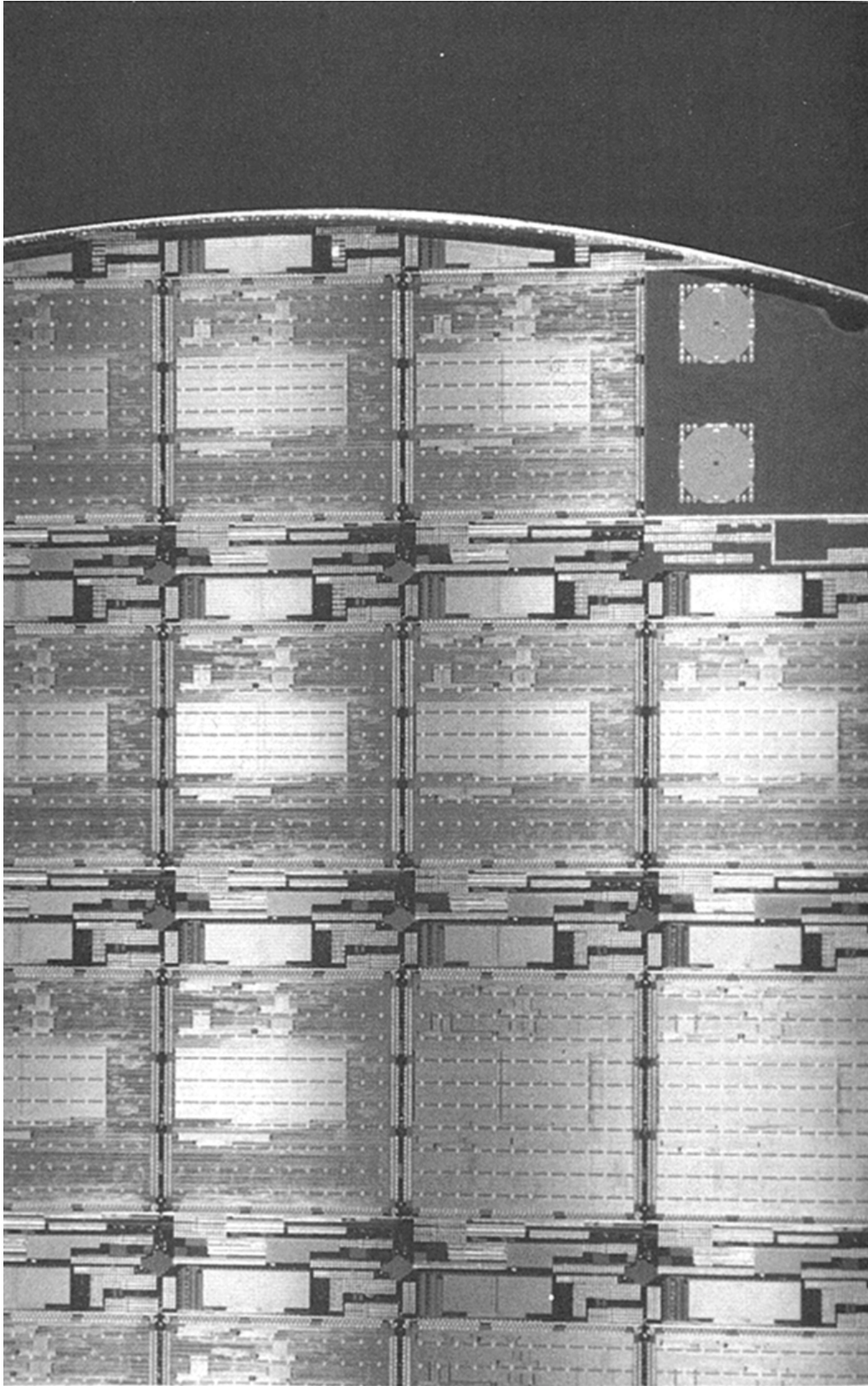
Sub-millimeter-wave heterodyne mixer elements based on superconductor-insulator-superconductor (SIS) tunnel junctions are currently the most sensitive mixer elements available from 100–800 GHz (3–0.4 mm) with noise levels of a few times the limiting value set by quantum mechanics. The primary application of SIS mixers is radio astronomy, where the observation of cold interstellar objects can be aided by ultrahigh-performance heterodyne receiver systems. The device shown was designed by Prof. J. X. Zmuidzinas at the California Institute of Technology (Caltech) and fabricated by the Low Temperature Superconductivity group at Caltech's Jet Propulsion Laboratory, Center for Space Microelectronics Technology under a contract from the National Aeronautics and Space Administration. This particular quasioptically coupled design was fabricated using Niobium superconductor technology and represents the state of the art in the 800 GHz band. The incoming RF radiation is coupled from free space into a pair of slot antennas using a series of lenses. The power absorbed by the antennas is coupled through a superconducting microstrip line impedance transformer into two Nb/AlOx/Nb SIS tunnel junctions. The placement of the microstrip "fans" near the center of the slot antennas causes the junctions to be fed asymmetrically by the RF signal so that the short section of microstrip connecting the two junctions behaves like a shunt inductor. This shunt inductance is designed to tune out the parasitic capacitance of the SIS tunnel junctions in a band about the design frequency. Other superconducting microstrip components serve as a low-pass filter to couple the intermediate frequency (IF) mixing products out of the mixer while confining RF signals to the active device area. (Courtesy of the Jet Propulsion Laboratory, center for Space Microelectronics Technology, and the California Institute of Technology.)

## Further Information

Every two years an Applied Superconductivity Conference is held devoted to practical technological issues. The proceedings of these conferences have been published every other year from 1977 to 1991 in the *IEEE Transactions on Magnetics*.

In 1991 the *IEEE Transactions on Applied Superconductivity* began publication. This quarterly journal focuses on both the science and the technology of superconductors and their applications, including materials issues, analog and digital circuits, and power systems. The proceedings of the Applied Superconductivity Conference now appear in this journal.

Vojin G. Oklobdzija, V. G. "Digital Systems"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



At present, most of the digital systems described in this section of the handbook are implemented in CMOS (complementary metal oxide semiconductor) technology. This technology allows for the fabrication of chips with wiring that is several times thinner than a human hair. Each system is packaged into an IC (integrated circuit). The different layers of silicon and metal defining the circuit are then sent for fabrication. Copies of the same circuit, and sometimes different circuits, are fabricated on the same wafer, as shown above. (Photo courtesy of IBM.)

# XIX

## Digital Systems

---

**Vojin G. Oklobdzija**  
*University of California, Davis*

**118 Logic Devices** *R. S. Sandige*

AND Gates • OR Gates • INVERTER Circuit • NAND Gates • NOR Gates

**119 Counters and State Machines (Sequencers)** *B. Wilkinson*

Binary Counters • Arbitrary Code Counters • Counter Design • State Machines • State Diagrams • State Diagrams Using Transition Expressions

**120 Microprocessors and Microcontrollers** *F. J. Hill*

Digital Hardware Systems • Processors • Assembly Language • Some Real Microprocessors and Microcontrollers

**121 Memory Systems** *R. S. Sandige*

CPU, Memory, and I/O Interface Connections • CPU Memory Systems Overview • Common and Separate I/O Data Buses • Single-Port RAM Devices • Additional Types of Memory Devices • Design Examples

**122 Computer-Aided Design and Simulation** *M. D. Ciletti*

Design Flow • Schematic Entry • Hardware Description Languages • Trade-offs between HDLs and Schematic Entry • HDLs and Synthesis • Transistor-Level Design and Simulation

**123 Logic Analyzers** *S. Mourad and M. S. Haydt*

Nature of Digital Signals • Signal Sampling • Timing Analysis • State Analysis • Components of a Logic Analyzer • Advanced Features of Logic Analyzers • Applications of Logic Analyzers

DIGITAL LOGIC DESIGN emerged as a discipline in the early days of computer systems when the need arose for a higher level of abstraction than that provided by electronic circuits. This marked the first time that electrical engineers designed using symbols and blocks that contained a number of sometimes quite complex electronic circuits. In addition, a number of nonengineers (logicians and computer scientists) entered the design process without an intimate knowledge of electronic circuits and electrical engineering. They were able to do so because they were working on a higher level of abstraction which did not require detailed knowledge. As a result, the productivity of designers has increased, as they have been relieved of design details such as speed, signal levels, noise, and power. Digital design was partitioned into *digital circuits*, *digital logic*, and *digital systems* on the highest level.

The digital circuit level concerns the design of digital components, often referred to as a "family" or "library." A set of gates and components is designed to be used by the logic designer working on the next level in the design hierarchy. These circuits must conform to certain electrical requirements and satisfy certain design criteria, such as speed and power consumption. In the beginning, logic components were built as modules which were later integrated on a single chip containing one or several logic gates or even blocks. This development was initiated by the birth of

the first integrated circuit at Texas Instruments in 1959, attributed to Jack Kilby. Today we can integrate millions of logic components on a single VLSI chip.

The consequences of this development have altered the way we design today. Managing such complexity would not be possible without extensive use of computer tools, known as CAD (computer-aided design). Not only is the logic designed with the use of CAD, but the design process itself is done by computers through extensive use of "logic synthesis." What this means is that the design behavior and specifications are described in some sort of language—HDL (hardware description language)—which serves as an entry point. From the specifications given in the HDL description, logic is automatically synthesized by a computer. In this way, it is possible to manage increasingly complex and large designs on a system level, dealing with functional specification and building blocks of high complexity. Not only is it important today to be able to manage the increased complexity of logic, but this task has to be done in an increasingly short time interval, known as the design "turn-around time." Introduction of hierarchy and use of CAD tools not only facilitates the design process, but it is the only way to manage designs consisting of millions of logic transistors.

Besides turn-around time, other issues are starting to dominate the design process. These are design correctness and testability. Testing and ensuring the testability of such increased complexity has become a problem and a discipline in itself. It is not only necessary to meet the increasingly short design time and to design for ever-increasing speed, but it is also necessary that the logic be "testable"; that is, the correctness of its functionality after the manufacturing process has to be guaranteed and detection of any possible faults ensured with a very high degree of confidence. For those reasons, logic design today is increasingly dependent on computer tools, to the point that it is difficult to separate one from the other.

Another very important development in logic design has been the introduction of field programmable logic (FPGA). Its importance is attributed to its ability to be configured very rapidly and often in the field—meaning in the product itself. The configuration process can be as rapid as several tens of milliseconds, allowing virtually flexible and configurable hardware to be used. Being generic and being able to take any shape and form, FPGA is very rapidly replacing logic previously found on the boards, known as "glue logic," usually surrounding large VLSI chips and connecting them in a system. FPGA is very closely integrated with CAD tools, making the FPGA design process automated to the point that it is possible to have "hardware libraries," files containing various designs that can be "loaded" into an FPGA chip during the computation, allowing the best fit to the particular algorithm or software routine. The notion of "reconfigurable hardware" and the abilities that FPGA offers are opening an entirely new realm in the way logic design is done and perceived.



Sandige, R. S. "Logic Devices"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

118.1 AND Gates

118.2 OR Gates

118.3 INVERTER Circuit

118.4 NAND Gates

118.5 NOR Gates

**Richard S. Sandige**

*University of Wyoming*

Logic devices as discussed in this chapter are of the electronic type and are also called **digital circuits and switching circuits**. Electronic logic devices that are manufactured today are primarily transistor-transistor logic (TTL) circuits and complementary metal oxide semiconductor (CMOS) circuits. Specialized electronic design engineers called *digital circuit designers* design TTL and CMOS switching circuits. *Logic* and *system designers* gain familiarity with available logic devices via manufacturer's catalogs.

The AND gate, the OR gate, the INVERTER circuit, the NAND gate, and the NOR gate are the fundamental building blocks of practically all electronic manufactured switching circuits. From these fundamental logic devices, larger blocks such as counters, state machines, microprocessors, and computers are constructed. Construction of larger logic devices is accomplished using the mathematics introduced in 1854 by an English mathematician named George Boole [Boole, 1954]. The algebra he invented is appropriately called *Boolean algebra*. It was not until 1938 that Claude Shannon [Shannon, 1938], a research assistant at Massachusetts Institute of Technology (MIT), showed the world how to use Boolean algebra to design switching circuits.

In the following sections the building blocks for the fundamental logic device are discussed, and their truth table functions and Boolean equations are presented.

---

## 118.1 AND Gates

The AND logic device is referred to as an *AND gate*. The function it performs is the AND function. The function is most easily represented by a table of truth called a *truth table*, as illustrated in Table 118.1.

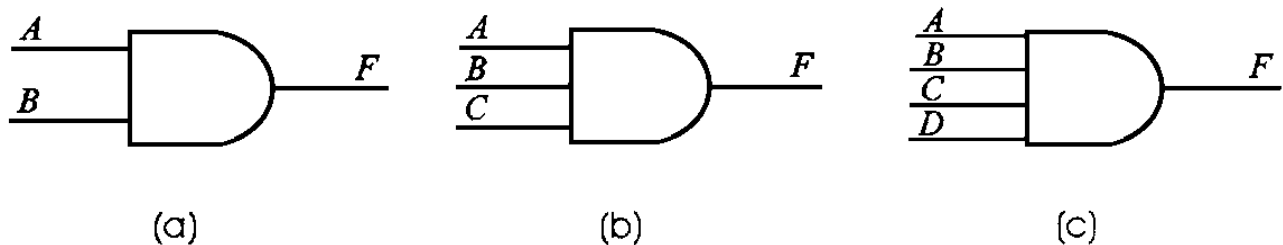
**Table 118.1** AND Gate Truth Table Function

$A$	$B$	$F$
0	0	0
0	1	0
1	0	0
1	1	1

The values in the truth table are ones and zeroes, but they could just as easily be trues and falses, highs and lows, ups and downs, or any other symbolic entries that are easily distinguished. The output  $F$  is 1 only when both inputs  $A$  and  $B$  are 1, as illustrated in Table 118.1. For all other input values of  $A$  and  $B$  the output values of  $F$  are 0. Figure 118.1(a) shows the logic symbol for an AND gate with two inputs [ANSI/IEEE Std 91-1984]. The Boolean equation that represents the function performed by the AND gate is written as

$$F = A \text{ AND } B \quad \text{or} \quad F = A \cdot B \quad (118.1)$$

**Figure 118.1** (a) AND gate logic symbol with two inputs; (b) logic symbol for a 3-input AND gate; (c) logic symbol for a 4-input AND gate.



The raised dot represents the AND operator and signifies the operation to be performed on the input variables  $A$  and  $B$ , with the result assigned to  $F$ . For  $n$  input **switching variables** there are  $2^n$  rows in the truth table and output values for the switching variable  $F$ . An AND gate can consist of 2 or more inputs. Standard off-the-shelf TTL AND gate devices [Texas Instruments, 1984] have as many as 4 inputs. **Programmable logic devices** (PLDs) that are used to configure larger circuits [Advanced Micro Devices, 1988] either by fuses or by stored charge movement electrically have AND gates with 32 or more inputs. Figures 118.1(b) and 118.1(c) show the logic symbols for a 3-input AND gate and a 4-input AND gate, respectively.

## 118.2 OR Gates

The OR logic device is referred to as an *OR gate*. The function it performs is the OR function. The truth table function for a 2-input OR gate is shown in Table 118.2.

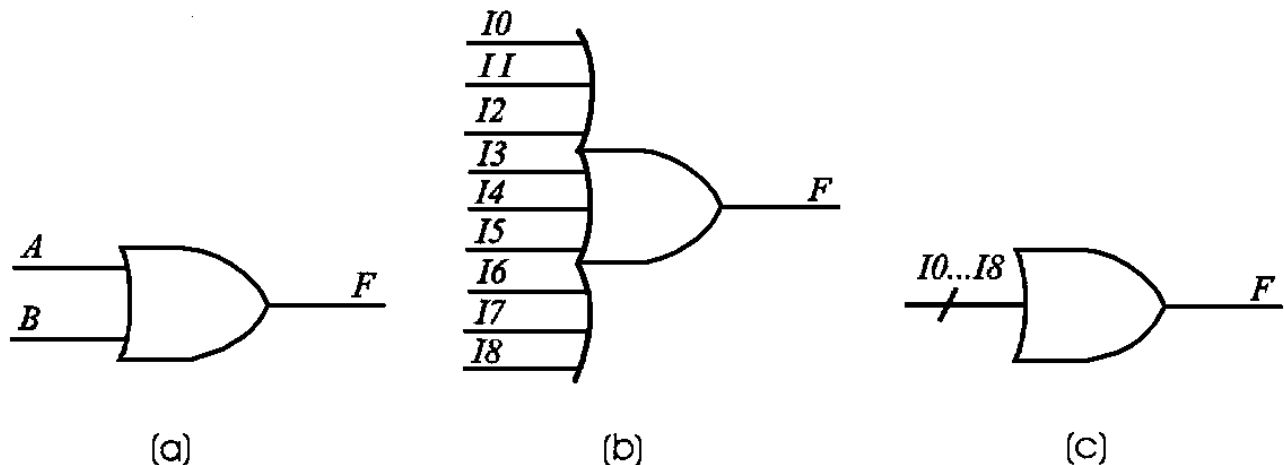
**Table 118.2** OR Gate Truth Table Function

<i>A B</i>	<i>F</i>
0 0	0
0 1	1
1 0	1
1 1	1

Like the AND gate, the OR gate can have many inputs. The output  $F$  for the OR gate is 1 any time one of the inputs  $A$  or  $B$  is 1, as illustrated in [Table 118.2](#). When the input values of  $A$  and  $B$  are both 0, the output value of  $F$  is 0. [Figure 118.2\(a\)](#) shows the logic symbol for an OR gate with two inputs. The Boolean equation that represents the function performed by the OR gate is written as

$$F = A \text{ OR } B \quad \text{or} \quad F = A + B \quad (118.2)$$

**Figure 118.2** (a) OR gate logic symbol with two inputs; (b) logic symbol for a 9-input OR gate; (c) simplified logic symbol for a 9-input OR gate.



The plus sign represents the OR operator and signifies the operation to be performed on the input variables  $A$  and  $B$ , with the result assigned to  $F$ . Standard off-the-shelf TTL AND gate devices are generally limited to 2 inputs. Programmable logic devices (PLDs) that are used to configure larger circuits either by fuses or by stored charge movement electrically have OR gates with as many as 9 or more inputs. [Figure 118.2\(b\)](#) shows a logic symbol for a 9-input OR gate. [Figure 118.2\(c\)](#) shows a simplified logic symbol for a 9-input OR gate.

## 118.3 INVERTER Circuit

The INVERTER circuit, sometimes called the NOT circuit, is the simplest circuit of the fundamental logic devices because it only has one input. Its truth table function is shown in [Table 118.3](#).

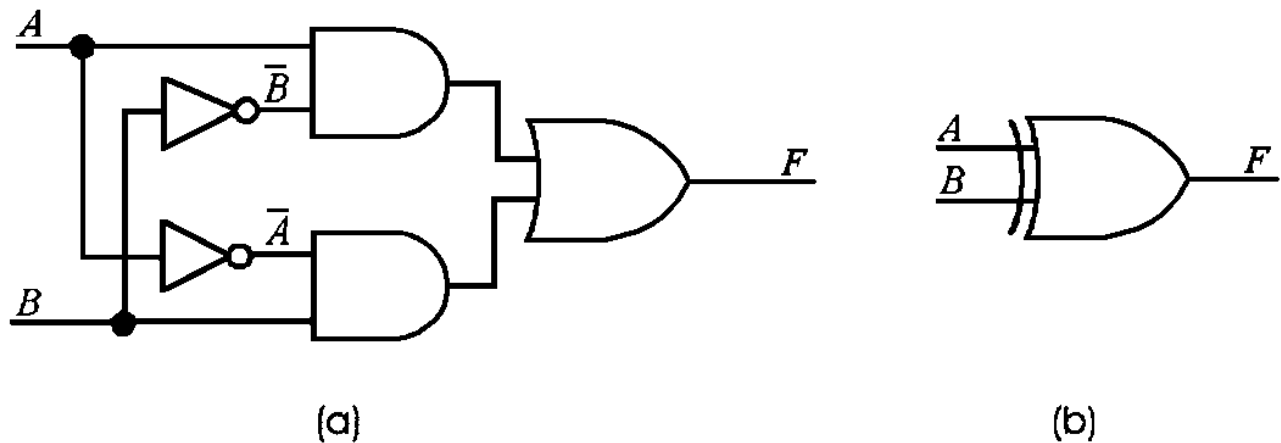
**Table 118.3** INVERTER Circuit Truth Table Function

$A$	$F$
0	1
1	0

Notice in [Table 118.3](#) that each output value of  $F$  is just the opposite of each input value for  $A$ . This inverting property provides a simple means of establishing a 1 output when a 0 input is supplied and a 0 output when a 1 input is supplied.

Larger circuits are often formed by connecting the outputs of two or more AND gates to the inputs of an OR gate. This form of configuration is called a **sum-of-products (SOP) form** since the AND gates perform a logic product and the OR gate performs a logic sum—hence the name *sum of products*. [Figure 118.3\(a\)](#) shows the sum-of-products form for a very common logic circuit that performs **modulo 2 addition** called an *exclusive OR*. Observe that the logic symbol for the INVERTER circuit—that is, the triangular symbol with the small circle at the output—is used in two places in the circuit for the exclusive OR. The switching variable  $A$  is inverted to obtain  $\bar{A}$  by one INVERTER circuit, and  $B$  is inverted to obtain  $\bar{B}$  by the second INVERTER circuit. The logic symbol for the exclusive OR circuit is shown in [Fig. 118.3\(b\)](#). The truth table function for the Exclusive OR circuit is illustrated in [Table 118.4](#).

**Figure 118.3** (a) Exclusive OR circuit in SOP form; (b) logic symbol for the exclusive OR circuit.



**Table 118.4** Exclusive OR Circuit Truth Table Function

$A$ $B$	$F$
0 0	0
0 1	1
1 0	1
1 1	0

The Boolean equation for the exclusive OR circuit is

$$F = A \cdot \overline{B} + \overline{A} \cdot B \quad \text{or} \quad F = A \oplus B \quad (118.3)$$

The circled + symbol is the exclusive OR operator, which allows the function to be expressed in a simpler form than the first form, which uses the AND operator (the raised dot), the OR operator (the plus symbol), and the INVERTER operator (the overbar).

## 118.4 NAND Gates

The NAND gate—where NAND stands for NOT AND—is the basic gate perhaps most frequently used. It consists of an AND gate followed by an INVERTER circuit. An off-the-shelf NAND gate is simpler to design, costs less to manufacture, and provides less **propagation delay time** than the equivalent connection of an AND gate followed by an INVERTER circuit. The NAND gate performs the NAND function. The truth table function for a NAND gate with two inputs is shown in [Table 118.5](#).

**Table 118.5** NAND Gate Truth Table Function

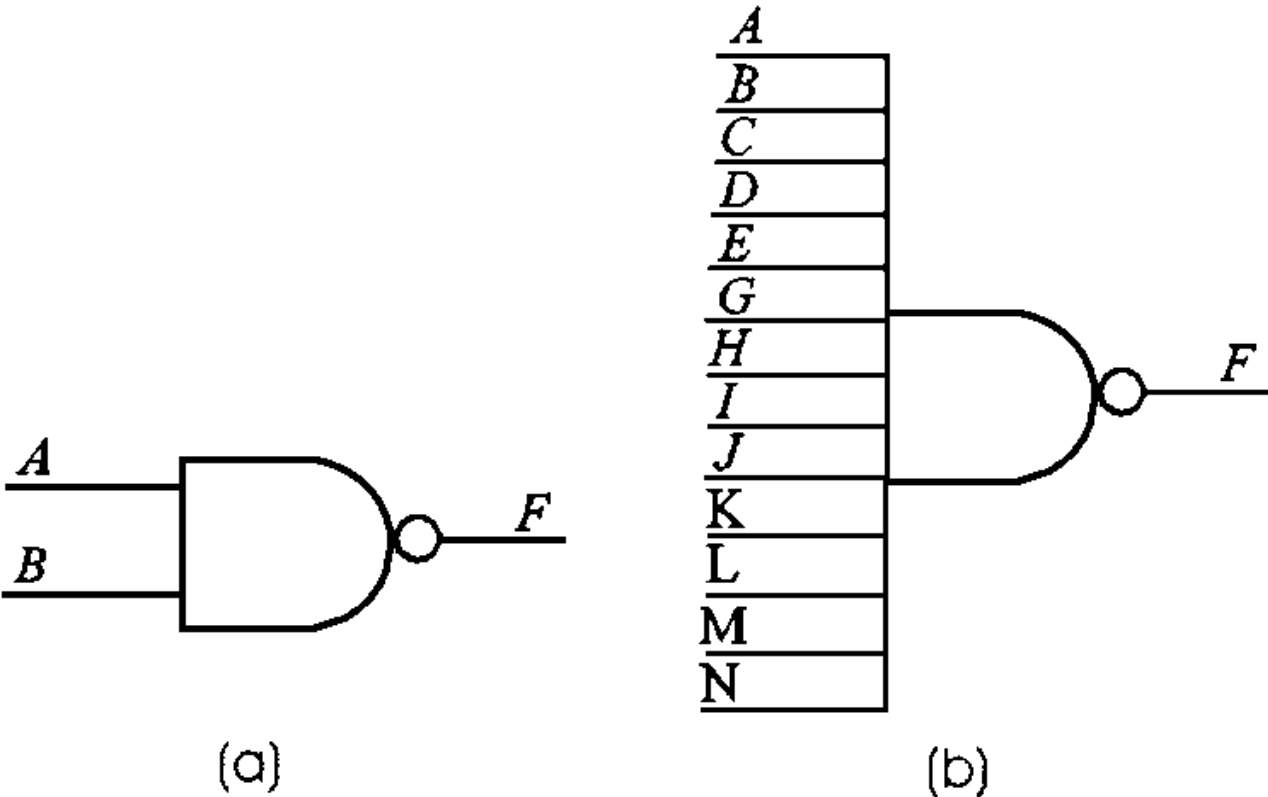
$A$ $B$	$F$
0 0	1
0 1	1
1 0	1
1 1	0

The output  $F$  is 0 only when both inputs  $A$  and  $B$  are 1, and for all other input values of  $A$  and  $B$  the output values of  $F$  are 1. As might be expected, the output values of  $F$  in the truth table function for the NAND gate, [Table 118.5](#), are each inverted from those in the truth table function for the AND gate, [Table 118.1](#). The Boolean equation for the NAND gate, Eq. (118.4), is the same as the Boolean equation for the AND gate, Eq. (118.1), with an added overbar to indicate that the output is inverted:

$$F = A \text{ NAND } B \quad \text{or} \quad F = \overline{A \cdot B} \quad (118.4)$$

[Figure 118.4\(a\)](#) illustrates the logic symbol for a NAND gate with two inputs. Notice that the NAND gate logic symbol simply consists of the AND gate logic symbol with a small circle attached to its output. The small circle indicates inversion like the small circle used in the logic symbol for the INVERTER circuit.

**Figure 118.4** (a) NAND gate logic symbol with two inputs; (b) logic symbol for a 13-input NAND gate.



Off-the-shelf TTL NAND gates are available with up to 13 inputs, as illustrated by the logic symbol shown in [Fig. 118.4\(b\)](#) for a 13-input NAND gate.

# 118.5 NOR Gates

The NOR gate consists of an OR gate followed by an INVERTER circuit. NOR is a contraction of NOT OR. Like the NAND gate, an off-the-shelf NOR gate is simpler to design, costs less to manufacture, and provides less propagation delay time than the equivalent connection of an OR gate followed by an INVERTER circuit. The NOR gate truth table function for two inputs is represented in [Table 118.6](#).

**Table 118.6** NOR Gate Truth Table Function

<i>A</i>	<i>B</i>	<i>F</i>
0	0	1
0	1	0
1	0	0
1	1	0

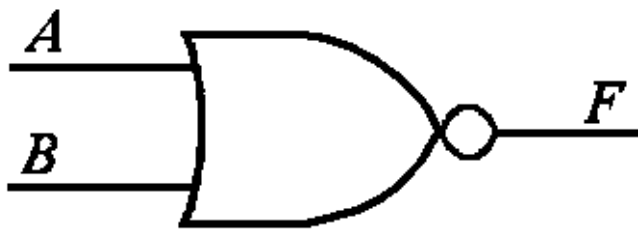
The output *F* is 1 only when both inputs *A* and *B* are 0, and for all other input values of *A* and *B* the output values of *F* are 0. The output values of *F* in the truth table function for the NOR gate,

Table 118.6, are each inverted from those in the truth table function for the OR gate, Table 118.2. The Boolean equation for the NOR gate, Eq. (118.5), is the same as the Boolean equation for the OR gate, Eq. (118.2), with an added overbar to indicate that the output is inverted.

$$F = A \text{ NOR } B \quad \text{or} \quad F = \overline{A + B} \quad (118.5)$$

The logic symbol for a NOR gate with two inputs is shown in Fig. 118.5. The NOR gate logic symbol is represented by the OR gate logic symbol with a small circle attached to its output. As mentioned earlier, the small circle indicates inversion. Off-the-shelf TTL NOR gates are available with up to 5 inputs.

**Figure 118.5** (a) NOR gate logic symbol with two inputs.



**Example.** Suppose a three-person digital voting circuit is needed to allow three judges to make a decision on a proposition, and their decision must be indicated as soon as all votes are made. For three judges there must be  $2^3 = 8$  rows in the truth table such that each row represents a different combination of judges' votes. The entry 101 would represent judge 1 voting yes, judge 2 voting no, and judge 3 voting yes; the decision for this row would be 1, showing a majority ruling, and the proposition would pass. For the case of 001 the decision for the row would be a 0, showing that a majority was not achieved, and the proposition would fail. The complete truth table function for a three-person voting circuit is shown in Table 118.7.

**Table 118.7** Truth Table Function for a Three-Person Voting Circuit

Judge 1	Judge 2	Judge 3	Decision
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

The truth table function for the three-person voting circuit cannot be represented by one of the fundamental logic devices. The circuit requires a combination of the fundamental logic devices. By



closely observing the truth table entries, we can write the following Boolean equation for the decision in terms of the three judges' inputs:

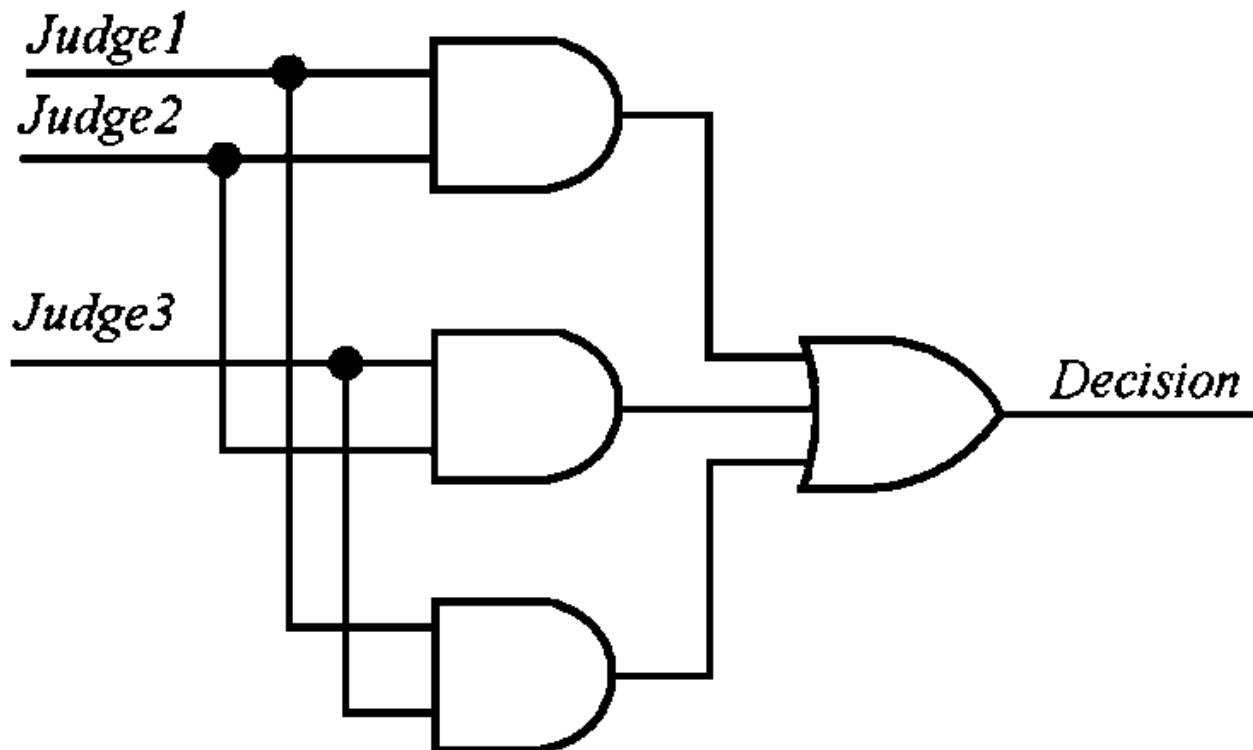
$$\begin{aligned} \text{Decision} = & (\text{judge 2 AND judge 3}) \text{ OR } (\text{judge 1 AND judge 3}) \\ & \text{OR } (\text{judge 1 AND judge 2}) \end{aligned} \quad (118.6)$$

which can be written more simply as

$$\text{Decision} = (\text{judge 2} \cdot \text{judge 3}) + (\text{judge 1} \cdot \text{judge 3}) + (\text{judge 1} \cdot \text{judge 2}) \quad (118.7)$$

The circuit for this Boolean equation in SOP form is shown in Fig. 118.6. Each input to the circuit can be supplied by an on/off switch that represent a yes/no vote respectively, and the output can be set up to drive a light *on* (lighted) or *off* (not lighted) to indicate the decision of the voting—where *on* represents a proposition has passed and *off* represents a proposition has failed. Other circuit designs are carried out in a similar manner by logic designers and system designers to build larger circuits and systems.

**Figure 118.6** (a) Three-person voting circuit in SOP form.



Logic devices are operated with voltages, so voltage values are sometimes used for the two distinct values, that is, the 1 and 0 discussed earlier. Each symbolic entry in a truth table thus

represents a range of voltages. Logic circuits of the same family, such as TTL or CMOS, are designed to be used in combination with other devices of the same family without the user worrying about actual voltages required other than  $V_{cc}$  and GND. Most switching circuits today are operated from a 5-volt supply source; however, newer switching circuits are rapidly appearing on the market requiring a 3-volt supply source. Logic devices utilizing a smaller voltage require less power to operate, and portable equipment manufacturers are the first to benefit from this trend in device technology.

## Defining Terms

**Digital circuits and switching circuits:** A class of electronic circuits that operates with two distinct levels.

**Modulo 2 addition:** Binary addition of two single binary digits, as expressed in [Table 118.4](#).

**Programmable logic devices:** Logic devices that are either one time programmable by blowing or leaving fuses intact or many times programmable by moving stored charges electrically.

**Propagation delay time:** Delay time through a logic device from input to output.

**Sum-of-products (SOP form):** Very popular form of Boolean equation representation for a switching circuit, also called AND/OR form, since AND gates feed into an OR gate.

**Switching variable:** Another name for an input variable, output variable, or signal name that can take on only two distinct values.

**$V_{cc}$  and GND:** Power supply connections required of all electronic devices. For a 5-volt logic device, the 5-volt power supply terminal is connected to the  $V_{cc}$  terminal of the device and the power supply ground terminal is connected to the GND terminal of the device.

## References

- Advanced Micro Devices. 1988. *PAL Device Data Book*. Advanced Micro Devices, Sunnyvale, CA.
- ANSI/IEEE Std 91-1984. *IEEE Standard Graphic Symbols for Logic Functions*. Institute of Electrical and Electronic Engineers. New York.
- ANSI/IEEE Std 991-1986. *IEEE Standard for Logic Circuit Diagrams*. Institute of Electrical and Electronic Engineers. New York.
- Boole, G. 1954. *An Investigation of the Laws of Thought*. Dover, New York.
- Roth, C. H., Jr. 1985. *Fundamentals of Logic Design*, 3rd ed. West, St. Paul, MN.
- Sandige, R. S. 1990. *Modern Digital Design*. McGraw-Hill, New York.
- Shannon, C. E. 1938. A symbolic analysis of relay and switching circuits. *Trans. AIEE*. 57:713–23.
- Texas Instruments. 1984. *The TTL Data Book Volume 3 (Advanced Low-Power Schottky, Advanced Schottky)*. Texas Instruments, Dallas, TX.
- Wakerly, J. F. 1994. *Digital Design Principles and Practices*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.

## **Further Information**

*IEEE Transactions on Education* published quarterly by the Institute of Electrical and Electronic Engineers.

Dorf, R. C. (Ed.) 1993. *The Electrical Engineering Handbook*. CRC Press, Boca Raton, FL.

Barry Wilkinson. "Counters and State Machines (Sequencers)"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Counters and State Machines (Sequencers)

---

- 119.1 Binary Counters
- 119.2 Arbitrary Code Counters
- 119.3 Counter Design
- 119.4 State Machines
- 119.5 State Diagrams
- 119.6 State Diagrams Using Transition Expressions

**Barry Wilkinson**

*University of North Carolina, Charlotte*

A *counter* is a logic circuit whose outputs follow a defined repeating sequence. After the final number in the sequence is reached, the counter outputs return to the first number. To cause the outputs to change from one number in the sequence to the next, a clock signal is applied to the circuit. A *clock signal* is a logic signal that changes from a logic 0 to a logic 1 and from a logic 1 to a logic 0 at regular intervals. The counter outputs change at one of the transitions of the clock signal, either a 0-to-1 transition or a 1-to-0 transition, depending on the design of the counter and logic components used. Counters are widely used in logic systems to generate control signal sequences and to count events.

### 119.1 Binary Counters

---

A common counter is the *binary counter*, whose outputs follows a linearly increasing or decreasing binary number sequence. A *binary-up counter* has outputs that follow a linearly increasing sequence. For example, a 4-bit binary-up counter (a binary-up counter with four outputs) has outputs that follow the sequence: 0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001, 1010, 1011, 1100, 1101, 1110, 1111, 0000, (i.e., 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 0, in decimal). A *binary-down counter* is a counter whose outputs follow a counting sequence in reverse order, for example, 1111, 1110, 1101, 1100, 1011, 1010, 1001, 1000, 0111, 0110, 0101, 0100, 0011, 0010, 0001, 0000, 1111, (i.e., 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0, 15, in decimal). A *bidirectional counter* is a counter that can count upwards or downwards depending on control signals applied to the circuit. The control signals specify whether the counter is to count up or down.

## 119.2 Arbitrary Code Counters

Counters can also be designed to follow other sequences. Examples include binary counters that count up to a specific number. For example, a 4-bit binary counter could be designed to count from 0 to 4 repeatedly, rather than from 0 to 15 repeatedly. Such counters can be binary counter designs with additional circuitry to reset the counter after the maximum value required has been reached. A *ring counter* generates a sequence in which a 1 moves from one position to the next position in the output pattern. For example, the outputs of a 5-bit ring counter follow the sequence: 10000, 01000, 00100, 00010, 00001, 10000,. A *Johnson counter* or *twisted ring counter* is a ring counter whose final output is "twisted" before being fed back to the first stage, so that when the final output is a 1 the next value of the first output is a 0, and when the final output is a 0 the next value of the first output is a 1. The sequence for a 5-bit Johnson counter is 10000, 11000, 11100, 11110, 11111, 01111, 00111, 00011, 00001, 10000,. Johnson counters have the characteristic that only one digit changes from one number in the sequence to the next number. This characteristic is particularly convenient in eliminating *logic glitches* that can occur in logic circuits using counter outputs. Logic glitches are unwanted logic pulses of very short duration.

## 119.3 Counter Design

Most counters are based upon *D*-type or *J-K* **flip-flops**. *J-K* flops are particularly convenient as they can be made to "toggle" (i.e., change from 0 to 1, or from 1 to 0) if a logic 1 is applied permanently to the *J* output and to the *K* input of the flip-flop. Hence a single *J-K* flip-flop can behave as a one-bit binary counter whose output follows the sequence 0, 1, 0, as shown in Fig. 119.1. We will describe counter designs using the toggle action of *J-K* flip-flops. In Fig. 119.1 the flip-flop output changes on a 1-to-0 transition of the clock signal—that is, a negative **edge-triggered** flip-flop is being used. We shall assume such flip-flops in our designs.

**Figure 119.1** One-bit counter using a *J-K* flip-flop: (a) circuit, (b) output waveform.

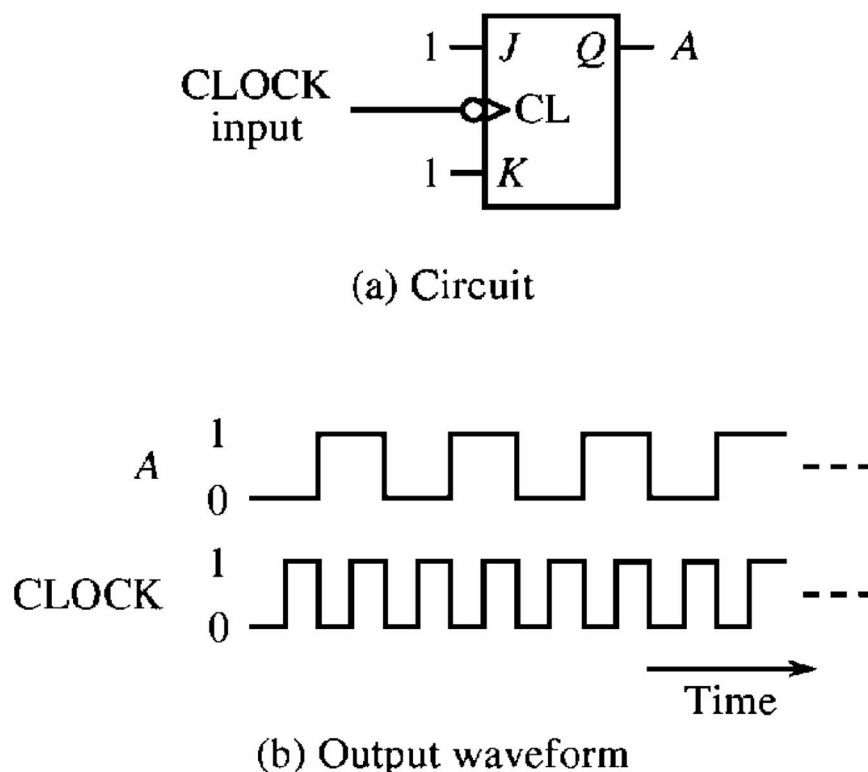
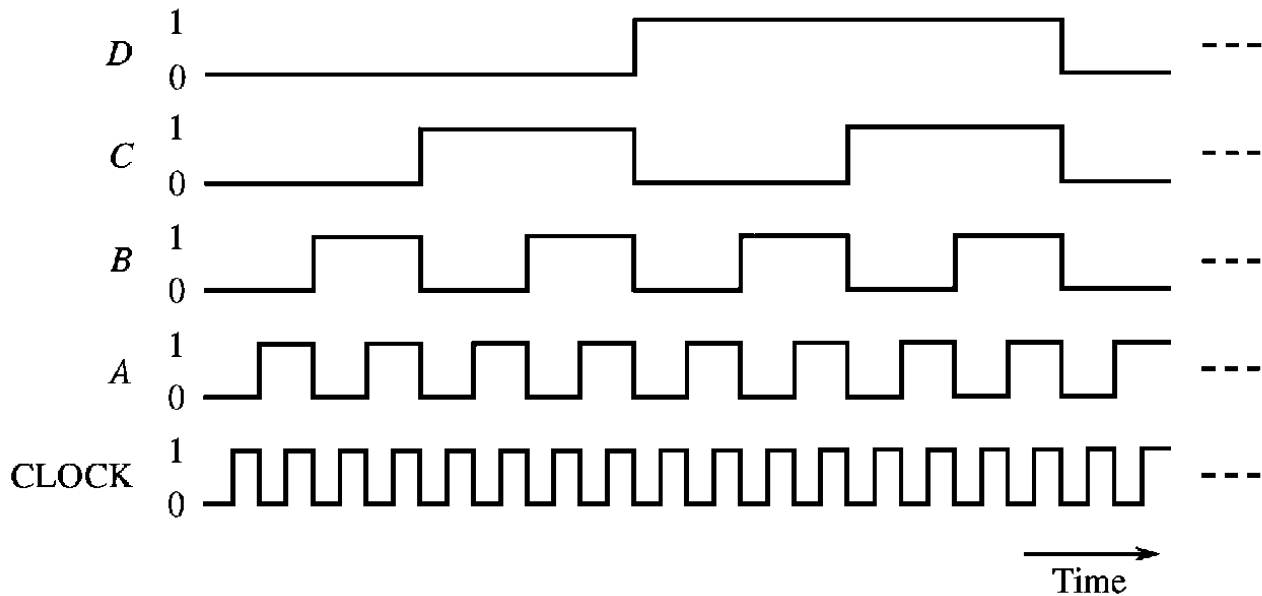


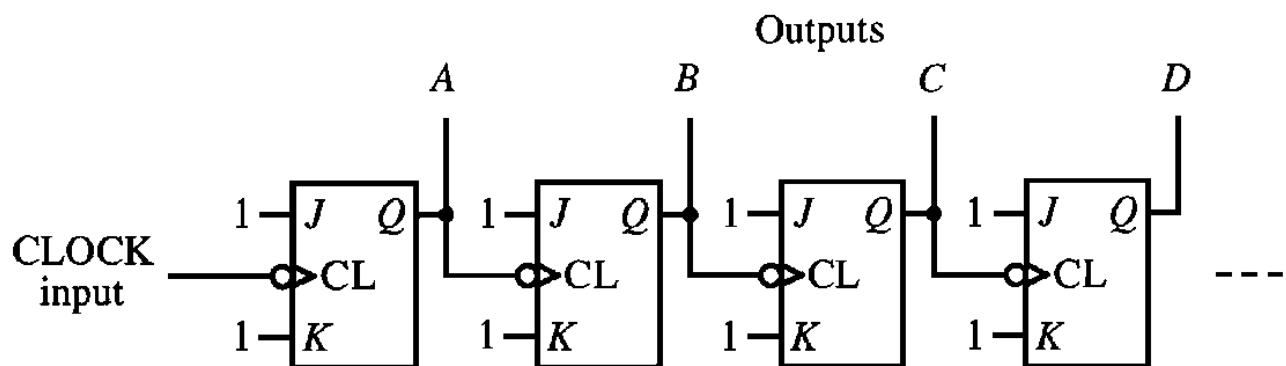
Figure 119.2 shows idealized waveforms of a binary-up counter whose outputs are  $A$ ,  $B$ ,  $C$ , and  $D$ . Such counter outputs can be created in one of several ways. There will be one flip-flop for each output and four flip-flops for a 4-bit counter.

**Figure 119.2** Binary-up counter outputs.



Counter designs can be classified as synchronous or asynchronous. The key characteristic of an *asynchronous counter* is that the output of one stage is used to activate the next stage. A 4-bit asynchronous binary (up) counter is shown in Fig. 119.3. Output  $A$  is the least significant bit of the sequence, and output  $D$  is the most significant bit. We can understand the operation of this counter by referring to the required counter waveforms shown in Fig. 119.2. Output  $A$  is to change when there is a 1-to-0 transition on the clock input signal. Output  $B$  is to change when there is a 1-to-0 transition on output  $A$ . Output  $C$  is to change when there is a 1-to-0 transition on output  $B$ . Output  $D$  is to change when there is a 1-to-0 transition on output  $C$ . These transitions are achieved with the connections shown.

**Figure 119.3** Asynchronous binary counter.

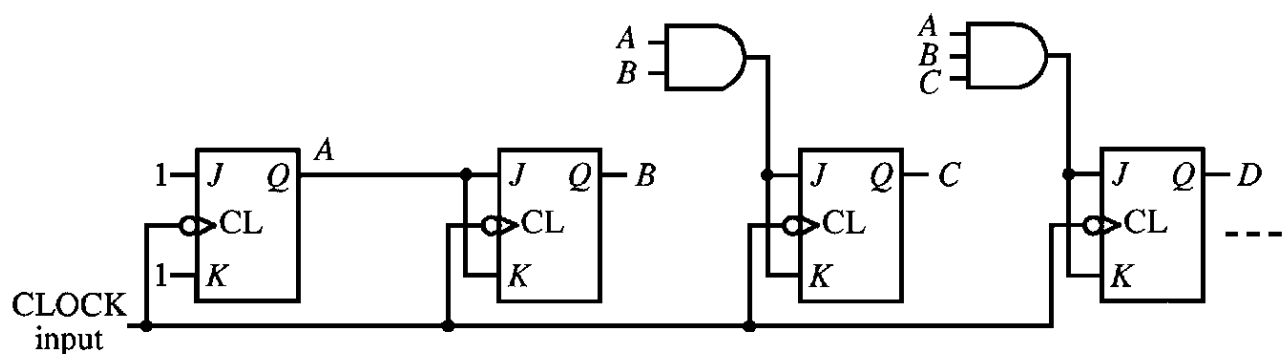


The circuit arrangement can be extended to any number of stages. The counter is asynchronous in nature because each output depends on a change in the previous output. Consequently, there will be a small delay between changes in successive outputs. The delays are cumulative, and the overall

delay between the first output change and the last output change could be significant, for example, from 1111 to 0000. This delay will limit the rate at which clock pulses can be applied (the maximum frequency of operation). Asynchronous binary counters are sometimes called *ripple counters* because the clock "ripples" through the circuit.

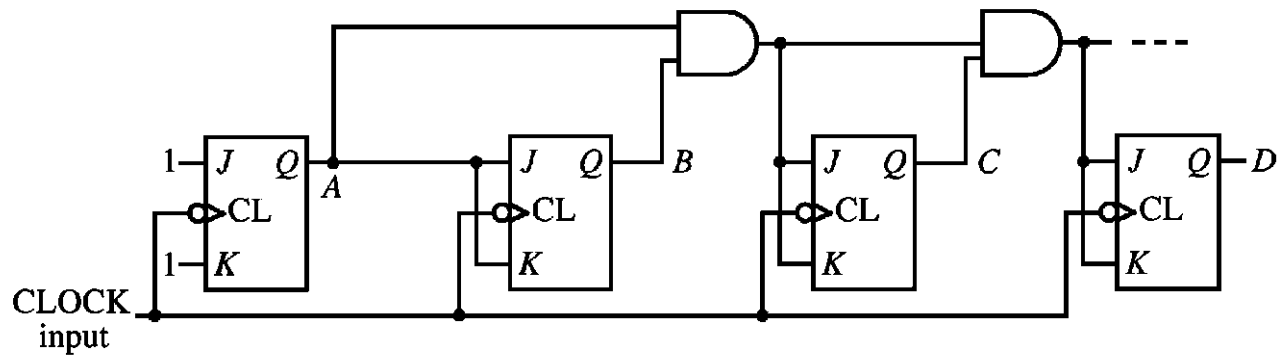
In the *synchronous counter* we try to avoid delays between successive outputs by synchronizing flip-flop actions. This step is done by connecting the clock signal to all the flip-flop clock inputs. Consequently, when outputs of flip-flops change, the changes occur simultaneously (ignoring variations between flip-flops). Logic is attached to the flip-flop inputs to produce the required sequence. Figure 119.4 shows one design of a synchronous binary (up) counter. We can understand this design also by examining the counter waveforms of Fig. 119.2. The first stage is the same as the asynchronous counter, and output changes occur on every activating clock transition. The second stage should toggle if, at the time of the activating clock transition, the *A* output is a 1. The third stage should toggle if, at the time of the activating clock transition, the *A* output and the *B* output are both 1. The fourth stage should toggle if, at the time of the activating clock transition, the *A* output, the *B* output, and the *C* output are all 1. These transitions are achieved by applying *A* to the *J-K* inputs of the second stage, the logic function  $A \cdot B$  (i.e., *A* AND *B*) to the third stage, and the logic function  $A \cdot B \cdot C$  (i.e., *A* AND *B* AND *C*) to the fourth stage. In each case an AND gate is used in Fig. 119.4. The toggle action for a fifth stage occurs when all of *A*, *B*, *C*, and *D* are a 1, and hence requires a four-input AND gate. Similarly, the sixth stage requires a five-input AND gate, and so on. An alternative implementation that reduces the number of inputs of the gates to two inputs is shown in Fig. 119.5. However, this particular implementation has the disadvantage that signals may have to pass through several gates to reach the inputs of flip-flops, which will limit the speed of operation of the counter.

**Figure 119.4** Synchronous binary counter.





**Figure 119.5** Synchronous binary counter using 2-input AND gates.



## 119.4 State Machines

Counters are in the logic classification called **sequential circuits**. The outputs of sequential circuits depend on present inputs and past output values, as opposed to **combinational circuits**, in which the outputs depend only on a particular combination of the present input values. Sequential circuits exist in defined *states*. The state of the circuit changes to another state on application of new input values. Such circuits are called *state machines*. Since in all practical circuits there will be a finite number of states, practical state machines are more accurately called *finite state machines*.

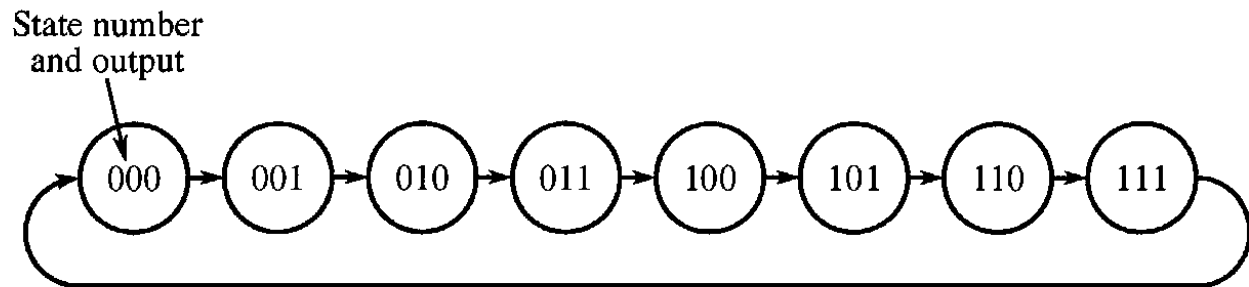
A counter is clearly a circuit that exists in defined states; each state corresponds to one output pattern. Other sequential circuits include circuits that detect particular sequences of input patterns (*sequence detectors*). As parts of the required input pattern are received, the circuit state changes from one state to the next, finally reaching the state corresponding to the complete input pattern being received.

Using *state variables*, each state is assigned a unique binary number, which is stored internally. An  $n$ -bit state variable can be used to represent  $2^n$  states. In *one-hot assignment* each state is assigned one unique state variable, which is a 1 when the system is in that state. All other state variables are zero for that state. Hence, with  $n$  states,  $n$  state variables would be needed. Though the one-hot assignment leads to more state variables, it usually requires less complicated logic.

## 119.5 State Diagrams

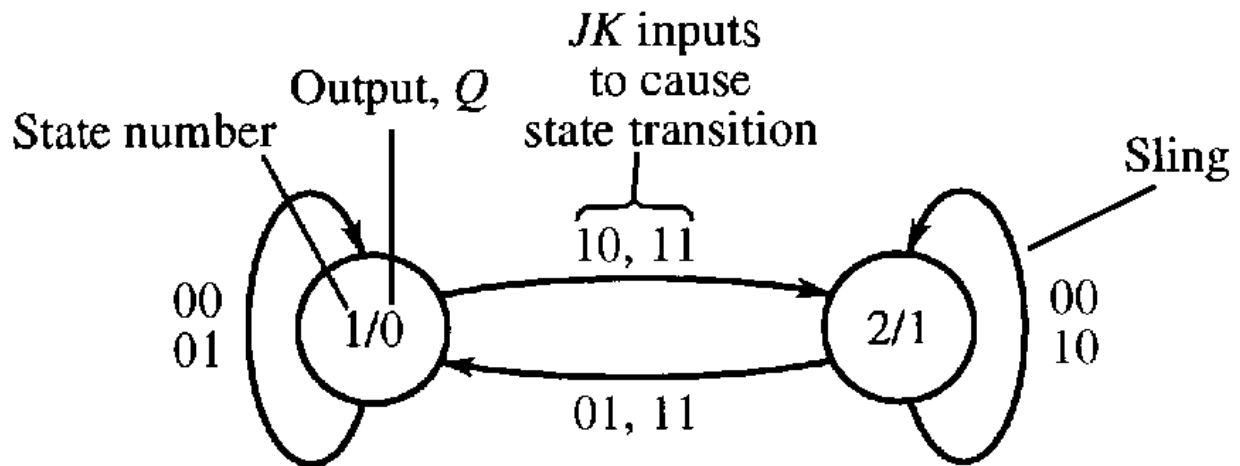
A *state diagram* is a graphical representation of the logical operation of circuits, such as counters, and sequential circuits in general. A state diagram indicates the various states that the circuit can enter and the required input/output conditions necessary to enter the next state. A state diagram of a 3-bit binary (up) counter is shown in Fig. 119.6. States are shown by circles. The state number for the state, as given by the state variables, is shown inside each circle. In a counter the state variables and the counter outputs are the same. Lines between the state circles indicate transitions from one state to the next state, which occurs for a counter circuit on application of the activating clock transition.

**Figure 119.6** State diagram of a binary counter.

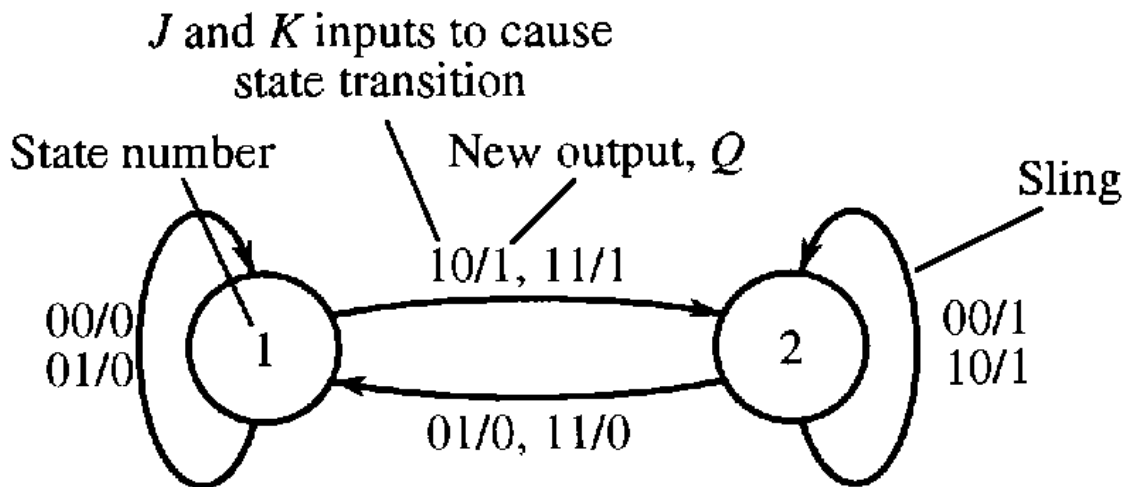


Other sequential circuits may have data inputs. For example, a  $J$ - $K$  flip-flop is a sequential circuit in its own right and has  $J$  and  $K$  inputs as well as a clock input. In that case we mark the lines between the states in the state diagram with the required input values to cause the state transition. There are two common forms of state diagram for circuits incorporating data inputs—the *Moore model state diagram* and the *Mealy model state diagram*. In both models the inputs that cause a new state are shown next to the lines. The effects of all combinations of input values must be considered in each state. A line that returns to the same state forms a "sling" and indicates that the state does not change with the particular input values. In the Moore model the outputs are associated with the states and are shown inside the circles. In the Mealy model the outputs are associated with input values and resultant states and are shown next to the transition lines between states. [Figure 119.7](#) shows the Moore model state diagram and the Mealy model state diagram for a  $J$ - $K$  flip-flop. The fundamental difference between the Moore model state diagram and the Mealy model state diagram is that, in the Moore model state diagram, each state always generates a defined output, whereas, in the Mealy model state diagram, states can generate various outputs depending upon present input values. Each model can represent any sequential circuit, though the number of states may be greater in the Moore model state diagram than in the Mealy model state diagram. The information given in the state diagram can be produced in a table called a *state table*.

**Figure 119.7** State diagrams of a *J-K* flip-flop: (a) Moore model state diagram, (b) Mealy model state diagram.



(a) Moore model state diagram



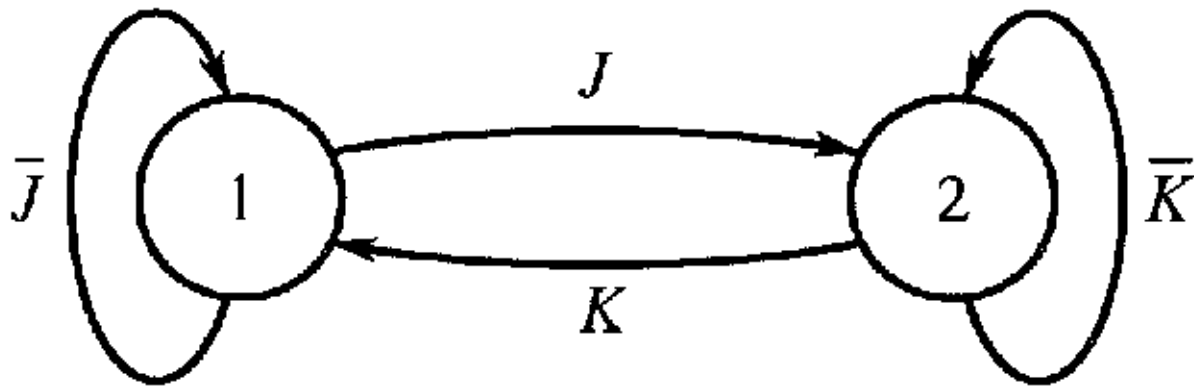
(b) Mealy model state diagram

## 119.6 State Diagrams Using Transition Expressions

An alternative form of state diagram that is particularly suitable for a state machine with a large number of states marks each line with a Boolean expression that must be true for the transition to occur. All expressions associated with each node must be exclusive of each other (i.e., no combinations of values must make more than one expression true). Also, the set of expressions must 0

cover all combinations of variables. This form of state diagram for a  $J$ - $K$  flip-flop is shown in Fig. 119.8. In this diagram we can see clearly that the transition from state 1 to state 2 depends only on  $J$  being a 1, and the transition from state 2 to state 1 depends only on  $K$  being a 1.

**Figure 119.8** State diagram of a  $J$ - $K$  flip-flop, using transition expressions.



## Defining Terms

**Asynchronous sequential logic circuit:** A sequential circuit in which a clock signal does not synchronize changes. In an asynchronous sequential logic circuit, changes in more than one output do not necessarily occur simultaneously. Changes in outputs may depend on other output changes.

**Combinational logic circuit:** A logic circuit whose outputs depend only on present input values.

**Edge triggering:** A mechanism for activating a flip-flop. In positive edge triggering the output changes after a 0-to-1 logic transition on the clock input. In negative edge triggering the output changes after a 1-to-0 logic transition on the clock input.

**Flip-flop:** A basic logic circuit that can maintain its output at either 0 or 1 permanently, but whose output can "flip" from 0 to 1 or "flop" from 1 to 0 on application of specific input values. Common flip-flops are the  $S$ - $R$  flip-flop, the  $J$ - $K$  flip-flop, and the  $D$ -type flip-flop. The names are derived from the letters used for the inputs.

**Programmable logic sequencers:** A logic circuit in the family of programmable logic devices (PLDs) containing the components to implement a synchronous sequential circuit. A programmable logic sequencer has gates and flip-flops with user-selectable feedback connections. See *PAL@ Device Data Book Bipolar and CMOS*, Advanced Micro Devices, Inc., Sunnyvale, CA, 1990; *Digital System Design Using Programmable Logic Devices* by P. Lala, Prentice Hall, Englewood Cliffs, NJ, 1990; or *Programmable Logic*, Intel Corp., Mt. Prospect, IL, 1994, for further details.

**Propagation delay:** The very short time period in a basic logic circuit between the application of a new input value and the generation of the resultant output.

**Sequential logic circuit:** A logic circuit whose outputs depend on previous output values and

present input values. Sequential circuits have memory to maintain the output values.

**Synchronous sequential logic circuit:** A sequential circuit in which a clock signal initiates all changes in state.

## References

Katz, R. H. 1994. *Contemporary Logic Design*. Benjamin/Cummings, Redwood City, CA.

Wakerly, J. F. 1990. *Digital Design Principles and Practices*. Prentice Hall, Englewood Cliffs, NJ.

Wilkinson, B. 1992. *Digital System Design*, 2nd ed. Prentice Hall, Hemel Hempstead, England.

## Further Information

The topic of counters and state machines can be found in most logic design books, a sample of which is listed in the reference section. *Contemporary Logic Design* by Katz is very readable and thorough for the electrical engineer. Other logic design books include:

Prosser, F. P. and Winkel, D. E. 1987. *The Art of Digital Design*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.

McCalla, T. R. 1992. *Digital Logic and Computer Design*. Macmillan, New York.

Some books, such as the book by McCalla, integrate logic design with computer design because logic design is used in the design of computers.

Hill, F. J. "Microprocessors and Microcontrollers"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Microprocessors and Microcontrollers

---

120.1 Digital Hardware Systems

120.2 Processors

120.3 Assembly Language

120.4 Some Real Microprocessors and Microcontrollers

**Fredrick J. Hill**

*University of Arizona*

A *processor* executes coded instructions on electronically stored data. The stream of coded instructions has usually been translated (compiled and assembled) from an alphanumeric representation to vectors of 1s and 0s and is stored in the same electronic medium as the data. A *microprocessor* is a processor fabricated on a single silicon chip of components whose sizes are measured in microns. As such, the microprocessor is an example of a very-large-scale *integrated* (VLSI) *circuit*. It is mounted in a package less than an inch wide and a few inches in length with a row of between 20 and 64 electrical connecting pins projecting from each side or, for more complex processors, a square package with a larger number of pins extending from the bottom. The most sophisticated of the 16-bit and 32-bit microprocessors are suitable for use as principal components of a personal computer. A *microcontroller* is a special kind of a microprocessor.

A microcontroller supplied by a particular vendor will be similar in many respects to the microprocessors intended by that vendor for use in general purpose computers. The microcontrollers will often have a simplified instruction set and will be designed to work with a smaller address space than other microprocessors. The space on the VLSI chip made available by these simplifications is typically used in the realization of *input-output* support functions that would otherwise be realized on separate chips. These actions have the effect of reducing the total parts cost in a microcontroller system, an important consideration if the processor is to be designed into a consumer product for which cost containment is paramount.

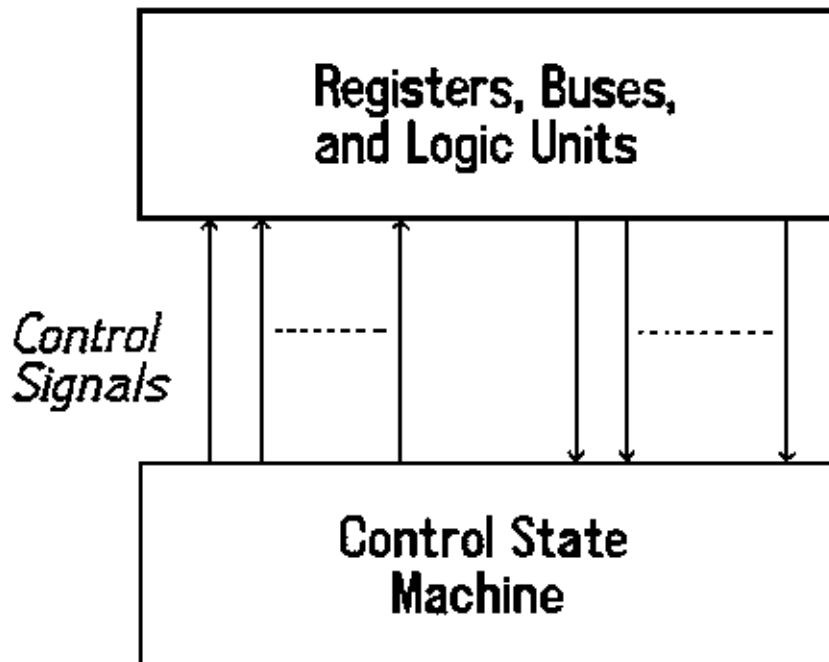
## 120.1 Digital Hardware Systems

---

A complex VLSI chip will consist of interconnected realizations of the logic devices of **Chapter 118** and the state machines of **Chapter 119**. A microprocessor falls within the definition of a sequential circuit or state machine, but the usual state machine notation—that is, state diagrams and state tables—are grossly inadequate for describing *digital hardware systems* with so many possible states. For example, a digital hardware system consisting of only one 32-bit register could assume any one of  $2^{32}$  states. A more powerful medium for describing digital systems—most of

whose memory elements are organized into registers—is depicted in Fig. 120.1. The block labeled "control" represents a state machine that can be described as in Chapter 119. The *data unit* block will consist of registers, *buses*, and combinational logic units. *Definition:* A bus is a conductor or a vector of conductors used for transmitting signals. A bus will typically be driven by multiplexers or tristate logic elements.

**Figure 120.1** Partition into control and data units.



The active state of the control state machine will activate one or more of the control output lines in Fig. 120.1. Each such active control line will trigger one or more activities in the data unit. These activities consist of connecting outputs of logic units or registers to buses; or enabling the loading of the output of a register, logic unit, or bus into a target register. Activities in the data unit will be synchronized by the system master clock, and the control unit may change state once each period of that same clock. One form of notation for describing the structure and function of digital systems conforming to the model of Fig. 120.1 is a *hardware description language*. [One hardware description language that has been approved as a standard is VHDL (Berge 1993).] The following might be a fragment of a description of some digital system in some unspecified hardware description language (HDL):

```

IF control state = q10 THEN
    ABUS = REGISTER4;
    carryin = carry_memory_element
    OBUS = ADD(ABUS; DATABUS; carryin);

```



carry\_memory\_element, REGISTER5 <- OBUS.

By itself this partial description indicates that, whenever the control unit is in state  $q_{10}$ , that REGISTER4 is connected to the bus ABUS. During the same period the sum of the vectors on the ABUS, another bus (perhaps the memory data bus), and the carry memory element propagates through a multiplexer to the OBUS. This sum vector is then loaded into REGISTER5. An adder with bit length compatible with processor data registers will be included in the *arithmetic and logic unit* (ALU) of every microprocessor.

The change of state in the control unit is not described in the fragment just given. To facilitate complete descriptions, the HDL must provide for specification of the appropriate change of state, perhaps as a function of the values of the signals from the data unit to the control unit as shown in [Fig. 120.1](#).

## 120.2 Processors

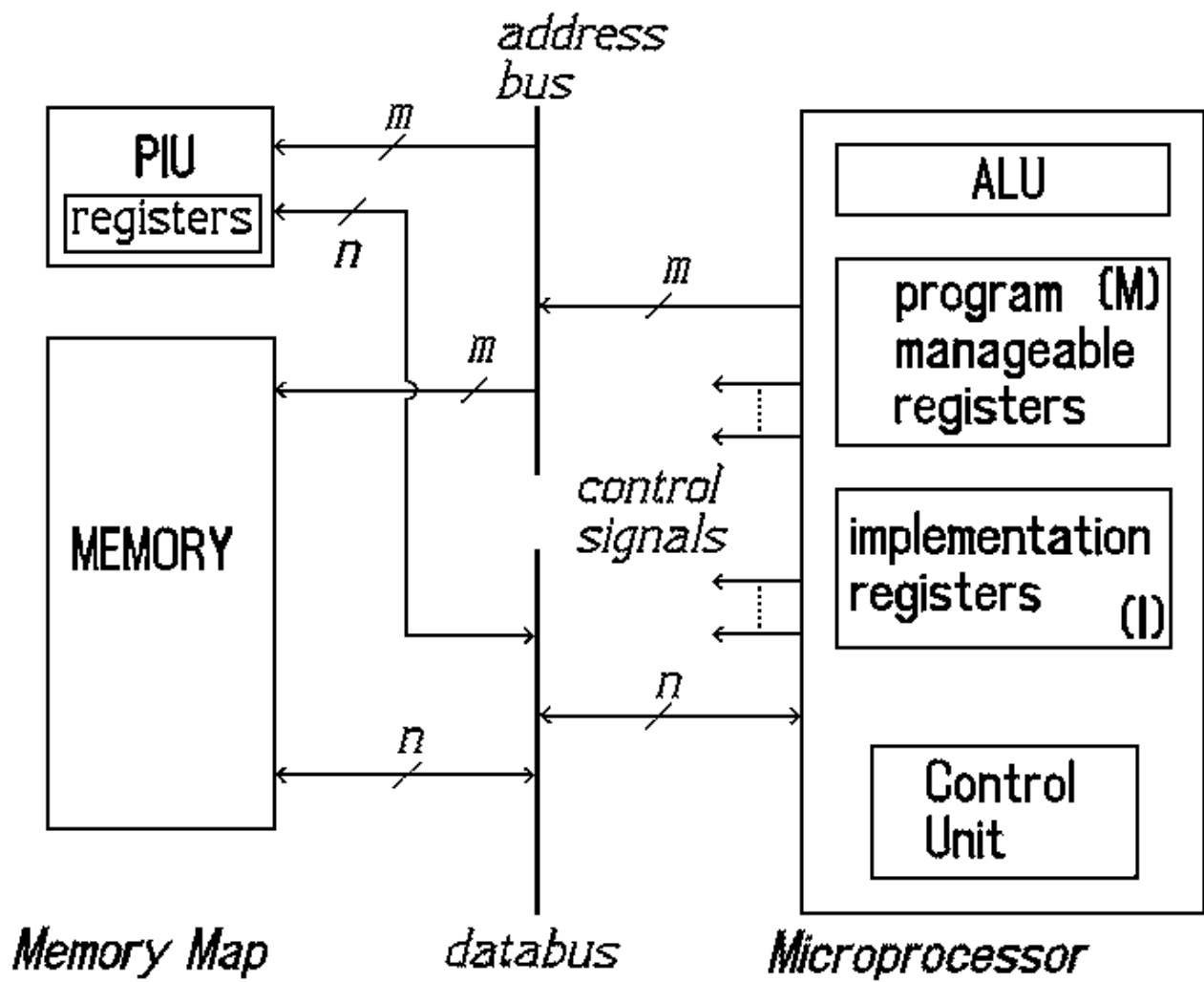
---

Not every digital hardware system satisfying the model of [Fig. 120.1](#) is a processor. A processor must include the capability to access data and process programs from a memory module, as depicted in [Fig. 120.2](#). The memory module will usually consist of several VLSI chips separate from the processor chip. The ALU incorporates the combinational logic necessary to implement the arithmetic, logical, or shifting operations that can be carried out by the processor. As suggested by the HDL fragment in the previous section, the arguments of these operations are selected from among the program-managed registers in block M and possibly a data vector arriving from memory on DATABUS. The result is loaded into a register in block M.

The set of instructions that is implemented in the ALU and the program manageable register configuration are often called the *architecture* of the processor. A physical implementation of that architecture will require additional logic and some registers hidden from the programmer in block I. [1] One register that must always be present in some form is the *program counter*, PC. The PC controls the sequencing through instructions in the program stored in memory. In the classical processor each instruction cycle begins with the *instruction fetch*. To initiate the instruction fetch PC is connected to the address bus. The instruction stored at the addressed location is then connected by the memory to DATABUS. The processor loads the instruction from data bus into some form of hidden *instruction register*, IR, that controls the execution of the instruction through subsequent periods of the master clock. Execution may require one or more similar memory transactions via the address and data bus to fetch operands from memory and/or to store a result in memory. At the end of the instruction cycle the program counter is incremented to sequence to the instruction stored at the next memory location.

The *instruction set* of every processor must include one or more *branch instructions* to permit the sequence of execution of instructions to be altered based on the results of previous instructions.

**Figure 120.2** Microprocessor system.



The execution of a branch is simply logical determination of a new address followed by its loading into the program counter. In most microprocessors the conditional branches are based on the contents of a set of individual memory elements or *flags* that may be set or cleared by each execution of an instruction through the ALU. Typical flags indicate whether the result was negative, zero, included a carry, or resulted in overflow (a result too large in magnitude to be stored in the intended target register). The flags are often collected to form an 8-bit register in block M.

The term  $n$  in Fig. 120.2 indicates the number of binary bits in a word (vector) that can be passed across the data bus and, therefore, the number of bits in each word of memory. Eight, 16, 32, and 64 are all possible values of  $n$  for a microprocessor. The length of a data register in box M of the processor is often equal to  $n$ . One exception was the original Motorola MC68000, which had 32-bit internal data registers but a 16-bit data bus. The width of the data bus became a limiting factor on the rate at which data could be processed, so the MC68000 was classified as a 16-bit processor.

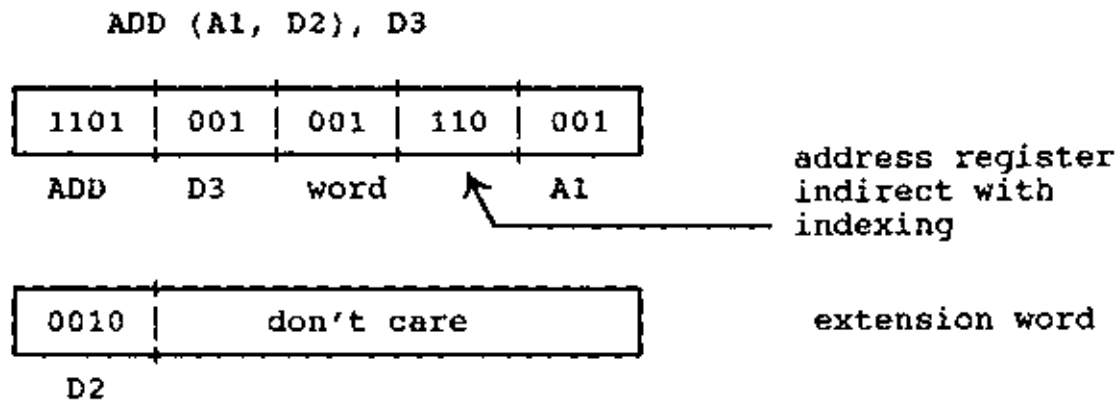
The number of bits in an address is indicated as  $m$  in Fig. 120.2. Typically,  $16 \leq m \leq 32$ . If  $m = 23$ , there are  $2^{23}$  words in the *memory map*. In most processors internal address registers are included among the program-accessible registers in block M. Sometimes the internal address registers will address data in finer units than the words that can be passed across the data bus. The smallest unit of separately addressed data is the 8-bit byte. For example, consider a processor with a 16-bit data bus and 24-bit internal address registers that map the memory as  $2^{24}$  8-bit bytes (16 megabytes). Only the most significant 23 bits of the address register are connected to the address bus ( $m = 23$ ). Once a 16-bit word is accepted from the data bus the least significant address register bit can be used to select between the two bytes of the data word.

Not every location in the memory map is physically implemented as a word in a *random access memory*. (See **Chapter 121**.) As illustrated in Fig. 120.2, control and data registers within *peripheral interface units* (PIUs) occupy locations in the memory map. Once data are written to the data registers of a PIU, the output process is completed under the control of that interface (a digital hardware system). The process takes place in reverse for input. The processor can provide direction to an I/O process by first writing to the control registers of a PIU. This approach to input/output is called *memory-mapped I/O*. Other blocks of consecutive words in the memory map are occupied by self-describing *read-only memory* (ROM), whereas still other blocks in the memory map remain unused.

## 120.3 Assembly Language

---

Instructions must be stored in memory as vectors of 1s and 0s. A typical instruction word will be partitioned into several fields. The string of 1s and 0s in each field codes some separate feature of what the instruction is to accomplish. Coding instructions in this format and interpreting the coding is a tedious task for the programmer. All information found in a machine-coded instruction for a particular processor may be represented in alphanumeric form in the *assembly language* unique to that processor. The following is one MC68000 assembly language instruction with the corresponding two consecutive machine-coded words.



The instruction will add the operand specified by (A1, D2) to the least significant 16-bit word of the data register D3 and will leave the result in D3. The field specified as 110 tells the processor that (Ax, X) address register, **indirect** with indexing, is the **addressing** mode for the first operand. The extension word is necessary in this case only to indicate that the index register is the data register D2. The address of the first operand is obtained by adding address register A1 to index register D2. This address is then placed on the address bus and the required operand is found on the data bus.

A program that will translate an assembly language program for a particular processor to the corresponding machine-coded program is called an *assembler*. A programmer may write a program directly in assembly language or use a *compiler* to translate to assembly language from a high-level language such as C.

## 120.4 Some Real Microprocessors and Microcontrollers

Some features of a selected subset of microprocessors and microcontrollers are given in [Table 120.1](#). Some of these features are defined at the end of the chapter. The overall performance of the microprocessor chips is directly related to the clock rate, the tabulated parameters, and the other listed features. Sets of standard benchmark programs have been devised to provide empirical measurement of performance. These benchmarks are usually applied to complete microcomputer systems (workstations), including memory, secondary storage, and input/output mechanisms, rather than bare processors.

**Table 120.1** Features of Select Microprocessors and Microcontrollers

Processor	IntroDate	Data Bus	ClockRate	Address Bus	Features <sup>*</sup>
Intel 8080	1975	8 bits		16 bits	
MC6800	1976	8 bits	2 MHz	16 bits	
Intel 8086	1978	16 bits	8 MHz	20 bits	FPCP, DNR
MC6805	1979	8 bits	2 MHz	16 bits	CNTRL, simplified 6800
MC68000	1981	16 bits	8 MHz	23 bits	FPCP, DNR

MC68HC11	1984	8 bits	4 MHz	16 bits	CNTRLR
MC68040	1989	32 bits	25 MHz	30 bits	C, FPPL, IPL, DNR
Intel 486	1991	32 bits	66 MHz	30 bits	C, IPL, FP, DNR
MC68332	1991	16 bits	17 MHz	24 bits	CNTRLR, 68000 based
DEC alpha	1992	64 bits	200 MHz	34 bits	C, FPPL, IPL
SuperSparc	1992	32 bits	100 MHz	32 bits	C, FPPL, IPL
MPC601	1993	64 bits	80 MHz	32 bits	C, FPPL, IPL
Intel Pentium	1993	32 bits	66 MHz	30 bits	C, FPPL, IPL, DNR

\* Features: CNTRLR, microcontroller; C, on-chip **cache**; FPCP, floating point **coprocessor** chip available; FP, internal floating point; FPPL, floating point **pipeline**; IPL, integer **pipeline**; DNR, definitely not **RISC**.

Notice the 20-bit address bus of the Intel 8086, which could address  $2^{20} = 2^{10} \cdot 2^{10} = 1 \text{ kilo- byte} \cdot 2^{10} = 1 \text{ megabyte}$  of memory. In 1978 this appeared to be a very large memory space. The decision to tie the operating system MSDOS firmly to that amount of memory was a very costly one to those responsible for applications software development over the life of that operating system.

Only Motorola microcontrollers are listed in Table 120.1, but most IC vendors offer a line of microcontroller chips. The user pressure for conformity to a single architecture and a single operating system—which allowed Intel to be the volume-dominant producer of microprocessor chips—is not operative for microcontrollers. Compatibility is not important where programs are hidden from the user in noncomputer products. The advanced features that increase the aggregate rate at which information is processed are not of much interest in microcontroller applications. In place of the features listed in Table 120.1 small blocks of random access memory, user-programmable ROM (PROM), and PIUs are included on the same VLSI chip as the processor.

## Defining Terms

**Cache:** A high-speed buffer storage containing continuously updated copies of the most recently used words from main memory.

**Coprocessor:** A digital hardware system that does not interpret or execute programs but performs operations on data under the direction of a processor.

**Indirect addressing:** An addressing mode in which the first address determined in the execution of an instruction is not the address of the operand but is the *address of the address* of the operand.

**Pipeline:** Beginning execution of a new instruction prior to completion of another instruction of the same class. Several instructions may be in various stages of completion in a pipeline at one time.

**RISC:** Reduced instruction set computer. A processor described as RISC will usually have few instruction formats and a single instruction length. It will have few addressing modes and no indirect addressing. Memory access and operations on data are not specified in the same

instruction.

## Reference

Berge, J.-M. 1993. *VHDL '92: The New Features of the VHDL Hardware Description Language*. Kluwer Academic, Boston, MA.

## Further Information

*Reference manuals* on currently available microprocessors and microcontrollers are readily available from the respective manufacturers.

Geppert, L. 1993. Not your father's CPU. *IEEE Spectrum*. 30(12).

Wray, W. C. and Greenfield, J. D. 1994. *Using Microprocessors and Microcomputers*, 3rd ed. Prentice Hall, Englewood Cliffs, NJ.

Hill, F. J. and Peterson, G. R. 1987. *Digital Systems: Hardware Organization and Design*, 3rd ed. John Wiley & Sons, New York.

Sandige, R. S. "Memory Systems"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

[121.1 CPU, Memory, and I/O Interface Connections](#)

[121.2 CPU Memory Systems Overview](#)

[121.3 Common and Separate I/O Data Buses](#)

[121.4 Single-Port RAM Devices](#)

[121.5 Additional Types of Memory Devices](#)

[121.6 Design Examples](#)

**Richard S. Sandige**

*University of Wyoming*

This chapter deals primarily with relatively fast electronic memory systems. Slower magnetic media, other storage media, and input output interfaces are only briefly discussed. Read/write memory, read-only memory, static memory, dynamic memory, cache memory, volatility of memory, relative speed of memory, and applications of memory are the major topics that are discussed. Trade-offs, advantages, and disadvantages are considered for various memory systems in the following sections. This approach provides insight as to the applications for which each type of memory system is used and why.

---

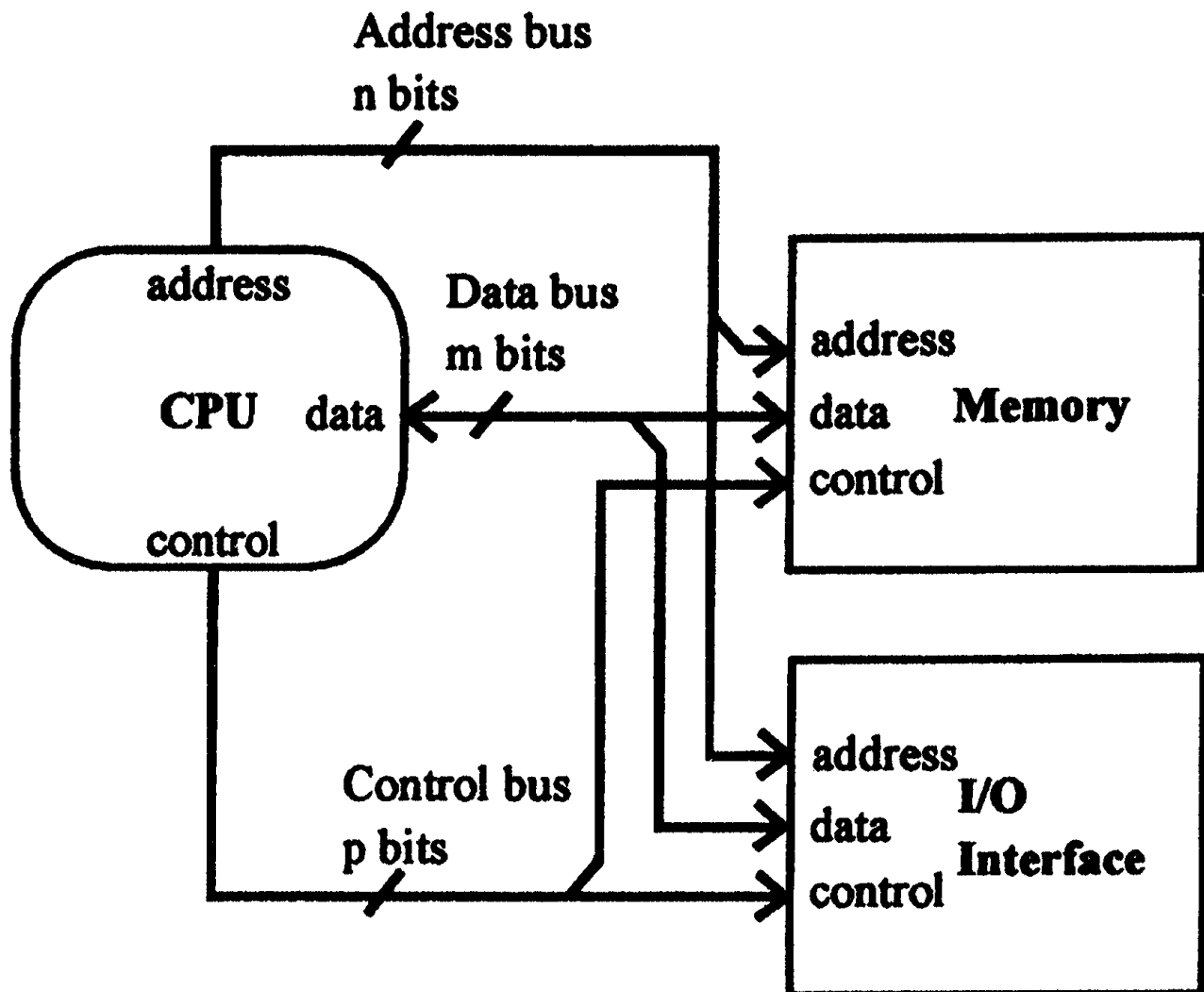
## 121.1 CPU, Memory, and I/O Interface Connections

---

Most applications that employ memory systems contain a processor or a central processing unit (CPU). The CPU carries out basic instructions that require moving data between internal registers and memory or between internal registers and an input/output (I/O) interface. Accomplishing this task usually requires a minimum of three types of buses, as illustrated in [Fig. 121.1](#). The CPU provides the locations or addresses where the data must come from during a memory read cycle or go to during a memory write cycle. The bus that conveys these locations to the memory and I/O interface is called the *address bus*. For an address bus of  $n$  bits there are  $2^n$  storage locations available. When the memory and I/O interface are both accessed via the same address lines and the same control lines, as illustrated in [Fig. 121.1](#), the interface is called *memory-mapped I/O*.



**Figure 121.1** CPU, memory, and I/O interface for a memory-mapped I/O architecture.



When the I/O interface is accessed as an address space separate from the memory address space, the address bus lines are interpreted as port addresses of the CPU, and separate control bus lines exist between the CPU and the I/O interface to direct data into or out of requested ports on the CPU via special I/O instructions. In the case of processors that have separate I/O and memory control signals, memory-mapped I/O can always be used by ignoring both the separate I/O control signals and the special I/O instructions and using only those processor instructions that relate to memory. Using only memory instructions to access both memory and I/O in the memory space of the CPU generally increases the flexibility of the overall system and also makes the system easier to program. A necessary requirement for the programmer of a memory-mapped I/O memory

system is a **memory map** that shows where all device locations exist.

Data, the contents of an address in a memory-mapped I/O system as illustrated in [Fig. 121.1](#), are read from the memory space on a CPU memory read cycle and written to the memory space on a CPU memory write cycle. An example of a memory read cycle and a memory write cycle timing diagram for a static RAM device will be presented later. Data travel on the **bidirectional data bus** in one direction at a time. The number of bits (data bus lines) associated with the data bus is dependent on the design of the CPU and ranges from a minimum of 4 bits for early microprocessors to as many as 32 bits for current microprocessors. Since heavy competition exists between major companies that manufacture microprocessors such as Motorola, Intel, IBM, Hewlett-Packard, and others, the push for 64 data bits and beyond for newer designs is underway. A larger number of data bits usually allows programs to be written more efficiently and also provides speed improvements in program execution since more bytes of data can be moved with fewer memory read or write cycles.

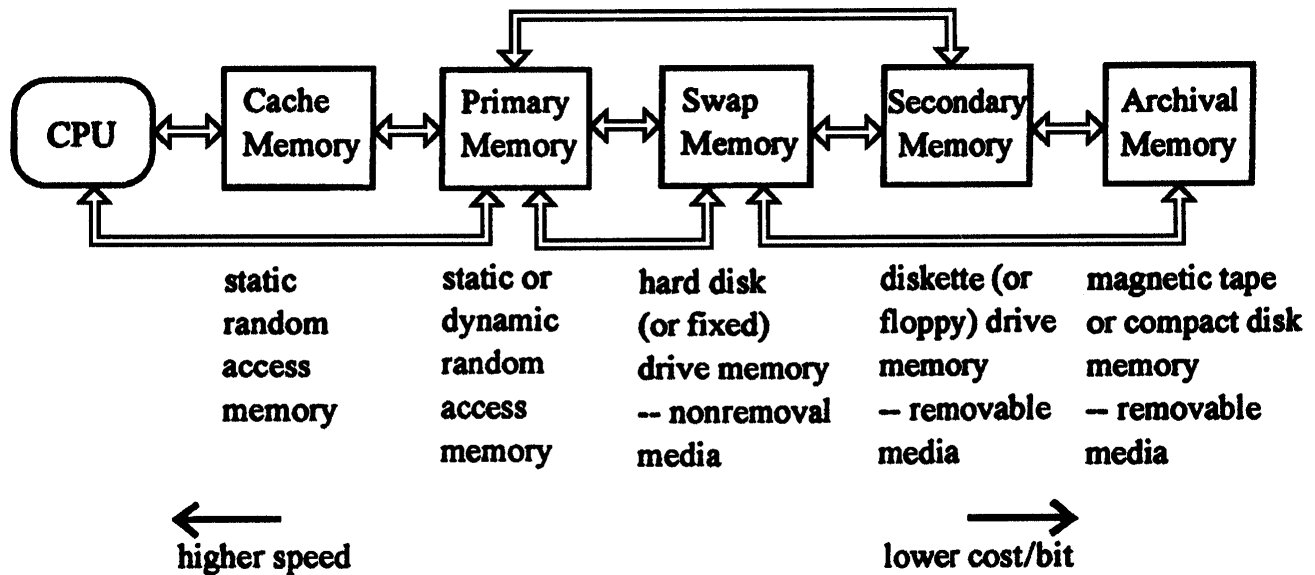
The control bus lines in a memory-mapped I/O system are routed to the memory and the I/O interface as illustrated in [Fig. 121.1](#). The I/O interface provides a data path for reading and writing data to the outside world that is beyond the system's existing memory system. The number of bits associated with the control bus is dependent on the architectural design of the complete system.

## 121.2 CPU Memory Systems Overview

---

A few of the various types of memory devices that can be connected to a CPU are illustrated in the memory systems connected to the CPU in [Fig. 121.2](#). The interconnections (all possible interconnections are not shown) must each have an address bus, a data bus, and a control bus, as discussed earlier. The **cache memory** block closest to the CPU usually contains static random access memory (RAM). Static RAM can also be used for primary memory; however, primary memory is usually designed with dynamic RAM, which operates at a slower speed and is less expensive than static RAM. Dynamic RAM must be constantly refreshed (about every 2 to 8 milliseconds) or stored data is lost. Only one transistor per storage location is required in a dynamic RAM, compared to several transistors per storage location in a static RAM, where refresh is not required. [Fig. 121.2](#) also provides a general indication of where other types of memory such as hard (or fixed) drive memory, diskette (or floppy) drive memory, and magnetic tape or compact disk memory fit into a memory system relative to speed and cost/bit. Both static and dynamic RAMs are volatile types of memory. When power is removed, the content of the data in all storage locations is lost. The memory devices used for swap memory, secondary memory, and archival memory shown in [Fig. 121.2](#) are all nonvolatile memory. When power is removed from nonvolatile memory, the content of the data in all storage locations is preserved. Swap, secondary, and archival memories are all slower and cost less per bit than static and dynamic memories. This chapter will concentrate only on the faster or higher-speed types of static and dynamic memories that are contained in blocks closer to the CPU.

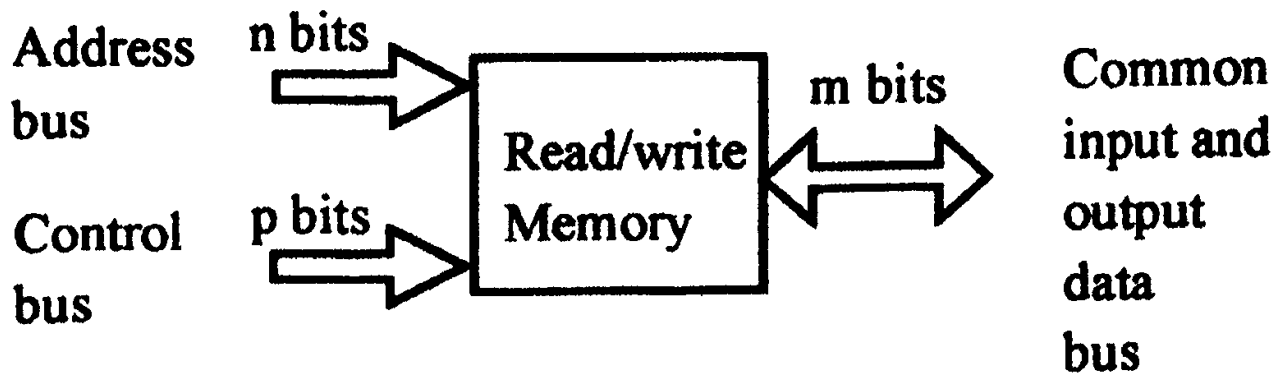
**Figure 121.2** Overview of a CPU memory system.



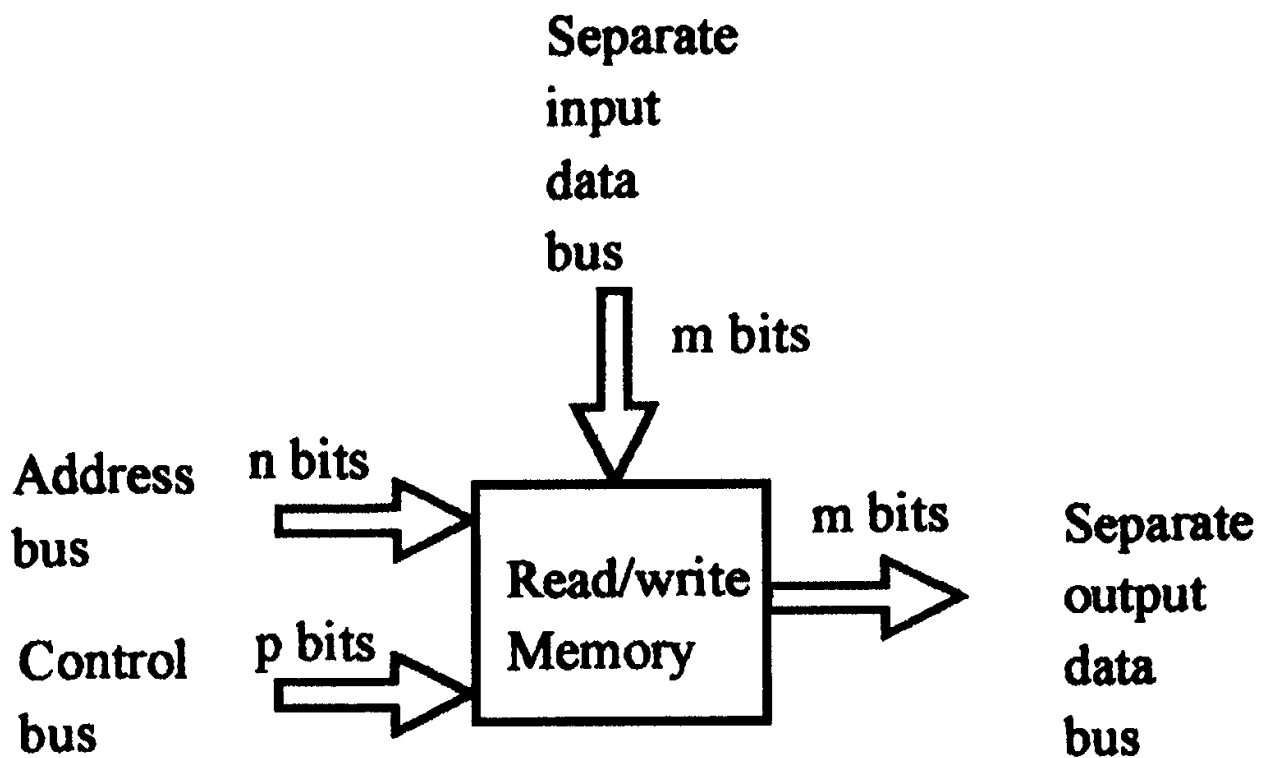
### 121.3 Common and Separate I/O Data Buses

Memory system blocks that consist of RAMs either utilize RAM devices that provide the input and output data lines via a common I/O data bus, as shown in Fig. 121.3(a), or utilize RAM devices that provide the input and output data lines on separate I/O data buses, as shown in Fig. 121.3(b). Memory systems that use memory blocks with a common I/O data bus have the advantage of costing less and taking up less printed circuit (PC) board space than those that use memory blocks with separate I/O data buses. Systems that use memory blocks with separate I/O data buses, however, have the advantage that **three-state buffers** can be added to the separate output data bus to eliminate **bus contention**, a problem in high-speed designs. Memory blocks that use either a common I/O bus or separate I/O buses are referred to as *single-port memory blocks* since data can only be accessed sequentially via a memory read cycle or a memory write cycle.

**Figure 121.3** (a) Memory block with a common I/O data bus; (b) memory block with separate I/O data buses.



(a)



(b)

Some memory system blocks utilize dual-port memory devices, which allow data to be written to an input port while other data are read at a separate output port at the same time. The video

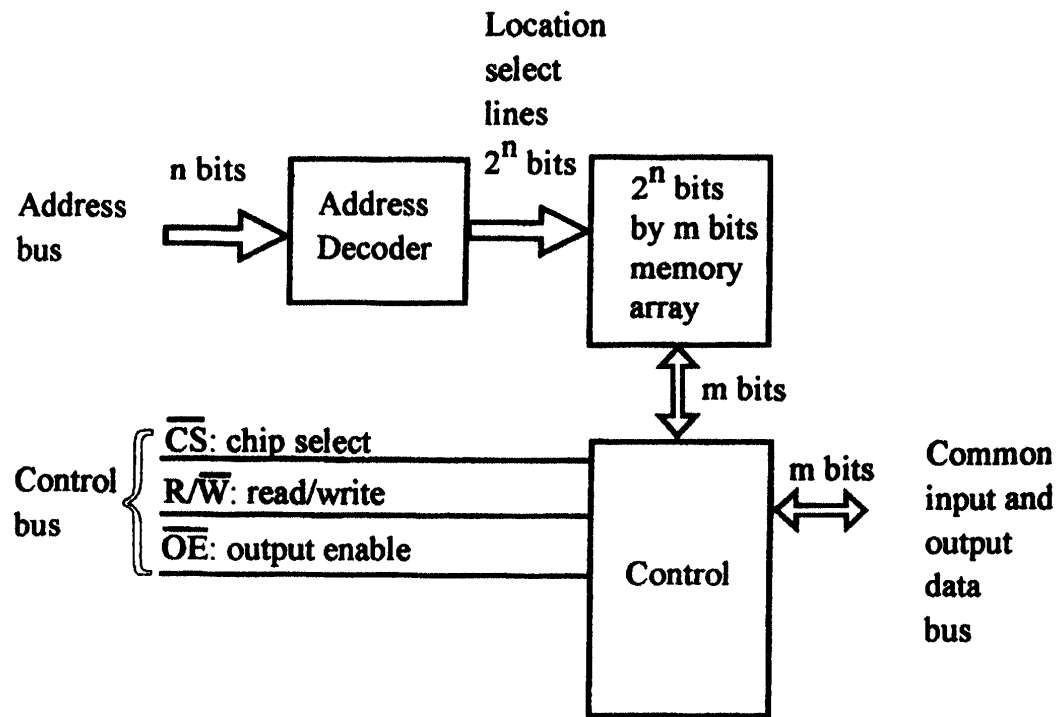
memory of fast computer workstations is often designed using dual-port static memory devices called *VRAMs* (video random access memory). VRAMs have two ways to access data simultaneously. Data can be accessed through a parallel RAM port and through a serial access memory (SAM) port. VRAMs are specialized devices that are specifically designed for use as display memories in **bit-mapped** graphic systems. Because of their multiple-access capability and functional flexibility, VRAM devices are usually the most expensive memory devices. It is also common in high-speed designs to have a dedicated video processor that issues video commands to speed up the video memory system.

## 121.4 Single-Port RAM Devices

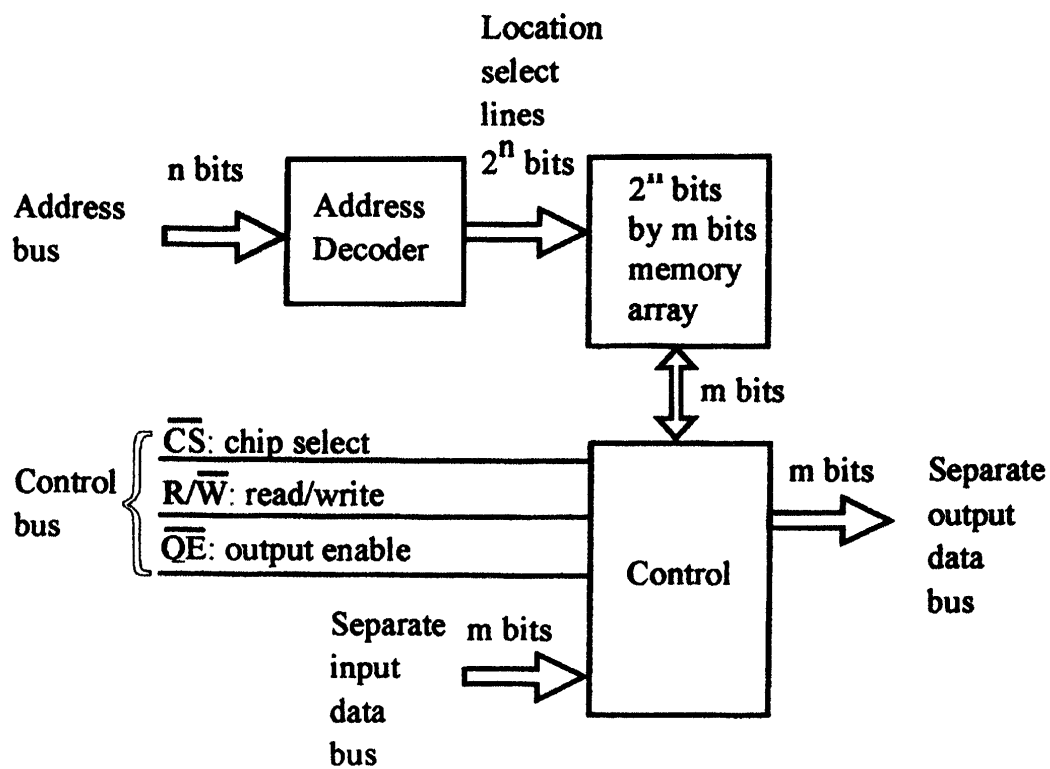
---

[Figure 121.4\(a\)](#) shows a block diagram of a static RAM device with a common I/O data bus. A block diagram of a static RAM device with separate I/O data buses is shown in [Figure 121.4\(b\)](#). In both types of RAM devices an address **decoder** is required to decode the  $n$  bits of address to determine one among the  $2^n$  location in the memory array where data is stored (written) or retrieved (read). Both types of RAM devices also can contain a chip select line, a read/write line, and an output enable line. Output data is read from a static RAM device via the data bus—that is, data out valid—when the chip select line ( $\overline{\text{CS}}$ ) is activated, the read/write control line ( $\text{R}/\overline{\text{W}}$ ) is activated for read, and output enable line ( $\overline{\text{OE}}$ ) is activated. These details are illustrated in the simplified timing diagram for the memory read cycle of a static RAM device as shown in [Fig. 121.5\(a\)](#). The  $\overline{\text{CS}}$  and  $\overline{\text{OE}}$  lines must both be pulled to a low-voltage logic state to be activated, since they both contain an overbar. The  $\text{R}/\overline{\text{W}}$  line must be pulled to a high-voltage logic state to be activated for read (no overbar over R). Abbreviations in the memory read cycle timing diagram are provided in [Table 121.1](#).

**Figure 121.4** (a) Static RAM device with a common I/O data bus; (b) static RAM device with separate I/O data buses.

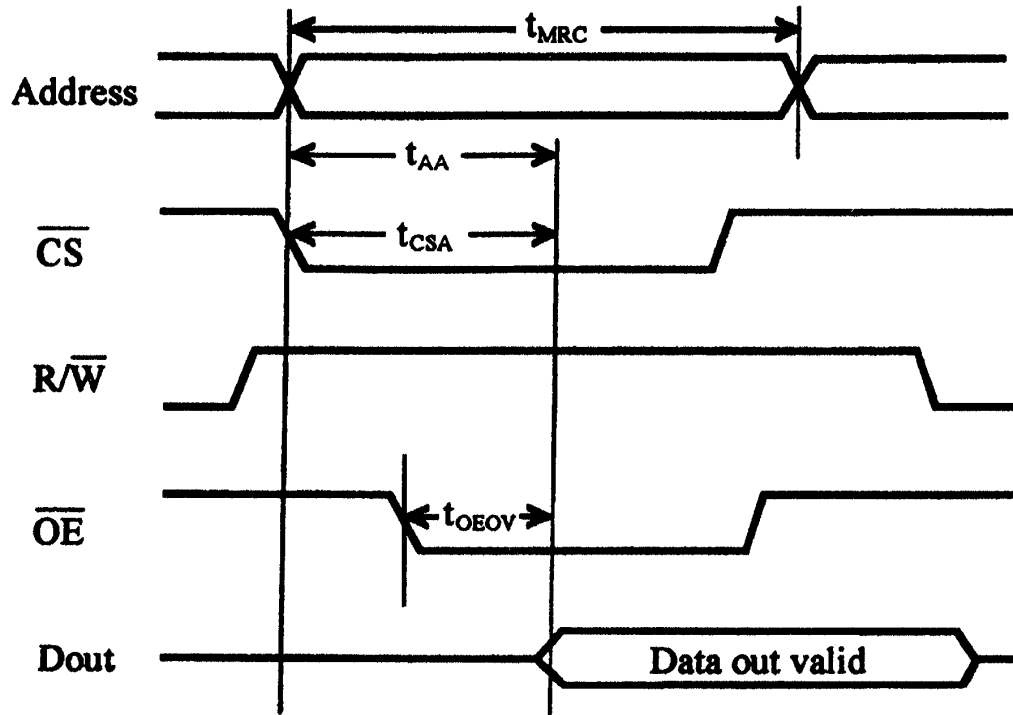


(a)

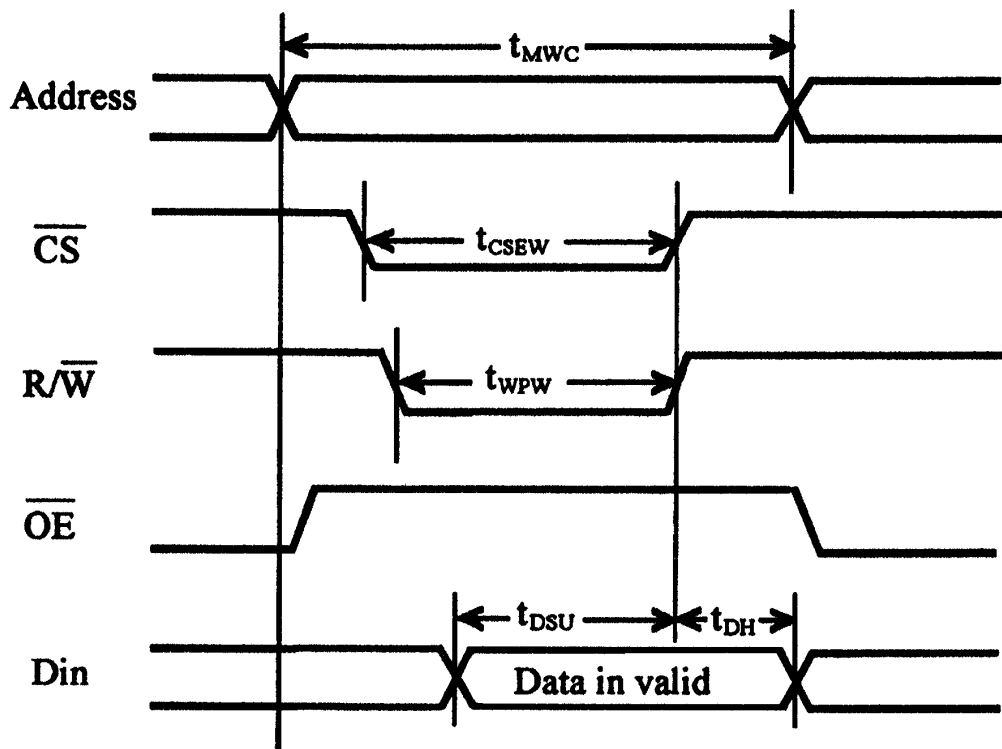


(b)

**Figure 121.5** (a) Simplified timing diagram for a memory read cycle; (b) simplified timing diagram for a memory write cycle.



(a)



(b)

**Table 121.1** Abbreviations for the Memory Read Cycle Timing Diagram

Abbreviation	Meaning
$t_{\text{MRC}}$	Memory read cycle time
$t_{\text{AA}}$	Address access time
$t_{\text{CSA}}$	Chip select access time
$t_{\text{OEOV}}$	Output enable to output valid time

Input data are written to a static RAM device via the data bus—that is, data in valid—when the chip select line ( $\overline{\text{CS}}$ ) is activated, the read/write line ( $\text{R}/\overline{\text{W}}$ ) is activated for write (pulled to a low-voltage logic state), and the output enable line ( $\overline{\text{OE}}$ ) is not activated (pulled to a high-voltage logic state). These details are illustrated in the simplified timing diagram for the memory write cycle of a static RAM device as shown in Fig. 121.5(b). Abbreviations in the memory write cycle timing diagram are provided in Table 121.2.

**Table 121.2** Abbreviations for the Memory Write Cycle Timing Diagram

Abbreviation	Meaning
$t_{\text{MWC}}$	Memory write cycle time
$t_{\text{CSEW}}$	Chip select to end of write time
$t_{\text{WPW}}$	Write pulse width time
$t_{\text{DSU}}$	Data setup time
$t_{\text{DH}}$	Data hold time

## 121.5 Additional Types of Memory Devices

Other types of memory devices that can be located close to the CPU are **read-only memories** (ROMs), programmable read-only memories (PROMs), ultraviolet read-only memories (UVPROMs), and electrically erasable read-only memories (EEPROMs). Notice that all of these devices are read-only devices in normal operation. Data can be stored permanently in the case of the ROM and PROM devices or can be restored up to about 100 times in the case of the UVPROM and EEPROM devices. ROMs are programmed at the factory by metallic masks and cannot be reprogrammed by the user. PROMs are generally programmed in the field using a **universal programmer** by blowing or leaving fuses intact; these are called *one-time-programmable* (OTP) devices. UVPROMs are programmed in the field using a universal programmer, and programmed data can be erased using an ultraviolet light source through a quartz window located on top of the chip. The time it takes to erase the device is about 15 to 45 minutes, depending on the intensity of the light source. Once erased, the UVPROM can be reprogrammed. EEPROMs, the latest technology, can be programmed using a universal programmer, and then they can be erased by electrical pulses in less than a minute.

When developing a ROM-based system, engineers often use EEPROMs, since the turnaround



time (the time for erasing and reprogramming) is very small to fix mistakes in programmed bit patterns. Engineers usually prefer to work with EEPROMs, until all the bugs are ironed out of a design, before committing to a ROM or PROM for the final design. ROMs are used in memory systems in computers to store information that needs to be accessed fast at power-up (such as the boot ROM that starts a PC system). EEPROMs are also used to store information that doesn't change very often, such as date, time, and configuration information.

ROMs and PROMs are permanently nonvolatile, whereas UVPROMs and EEPROMs are nonvolatile for approximately ten years after being programmed. ROMs are very inexpensive per part after an up-front high cost is paid to the manufacturer for making the mask. To program PROMs, UVPROMs, and EEPROMs requires the user to purchase a universal programmer. A high-end universal programmer can run a few thousand dollars, whereas a programmer that programs only certain types of PROMs can cost as little as four or five hundred dollars. PROMs are a little more expensive per part compared to ROMs and require that the user purchase a programmer. The next most expensive devices are UVPROMs and EEPROMs. To erase UVPROMs requires that the user purchase an ultraviolet light source in addition to a universal programmer. EEPROMs can be programmed and erased with a universal programmer.

Speeds of the various devices are generally given with the ROM as the fastest, the PROM the next fastest, and the UVPROM and EEPROM much slower. [Table 121.3](#) provides a quick reference concerning cost/part, speed, and so forth for ROM and PROM-type devices.

**Table 121.3** Summary of ROM and PROM-Type Devices

Device	Cost/Part	Speed	Nonvolatile	Programmer
ROM	Lowest <sup>1</sup>	Fastest	Permanently	N/A
PROM	Low	Fast	Permanently	Required
UVPROM	Higher	Slow	Semipermanently <sup>2</sup>	Required <sup>3</sup>
EEPROM	Higher	Slow	Semipermanently <sup>2</sup>	Required

<sup>1</sup>Up-front high cost must be paid to manufacturer for making the mask.

<sup>2</sup>Stores data for approximately ten years.

<sup>3</sup>An ultraviolet light source is also required to erase these devices.

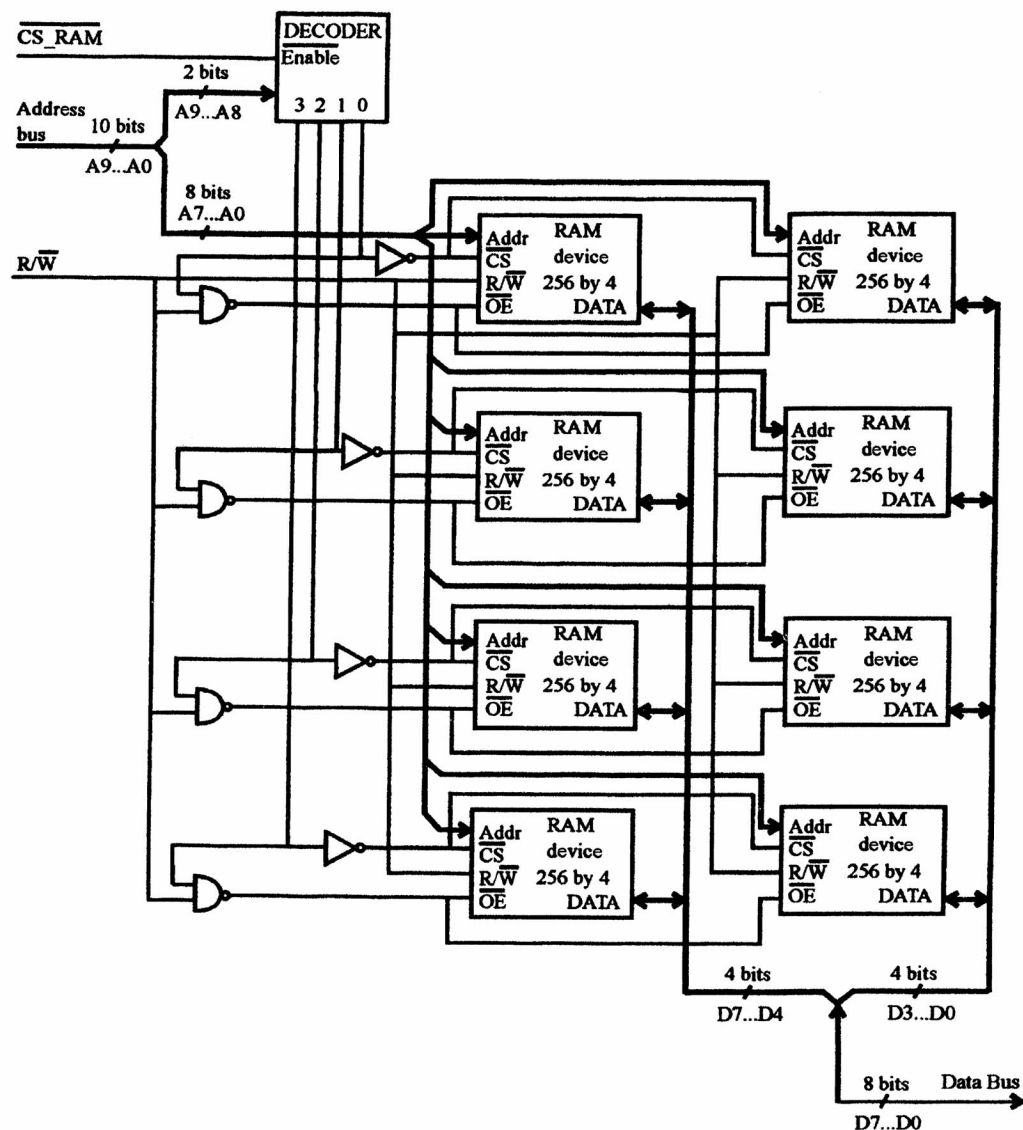
The timing diagram of interest for ROM users and the various PROM devices is the memory read cycle. The universal programmer provides the required memory write cycle for data storage for PROM devices. The memory read cycle is basically the same as the memory read cycle shown in [Fig. 121.5\(a\)](#). The  $R/\overline{W}$  signal for PROM devices is replaced by a program signal  $\overline{PGM}$  used to perform a memory write cycle. After the device is programmed the program signal  $\overline{PGM}$  is disabled or not activated (pulled to a high-voltage logic state) so that the device can only be read.

## 121.6 Design Examples

In general, a cache memory system and a primary memory system often require a larger number of address bits and data bits than single-chip RAM devices can supply. This condition requires that several RAM devices be connected together to form the design for a cache or primary memory system.

Figures 121.6 and 121.7 illustrate the designs for two separate memory system blocks using static RAM devices. The first memory system design shown in Fig. 121.6 addresses 1024 locations each 8 bits wide. The RAM devices used in this design have only 256 addresses with a common I/O 4-bit-wide data bus. This situation requires using a decoder to decode four ranges of addresses and two RAM devices simultaneously. The decoder requires 2 bits from the address bus to select the four different address ranges. Output enable for the two RAMs that are selected for a memory read cycle is obtained by ANDing the read/write signal with the decoder output that selects the required address range. Each RAM device requires 8 bits from the address bus to select among the 256 locations in its memory. Each RAM device only provides 4 bits of content at each memory location from 0 through 1023, thus requiring two RAM devices for each address.

**Figure 121.6** A  $1024 \times 8$ -bit static RAM memory system using  $256 \times 4$ -bit RAM devices with a common I/O data bus.



**Figure 121.7** A  $1024 \times 8$ -bit static RAM memory system using  $1024 \times 1$ -bit RAM devices with a common I/O data bus.

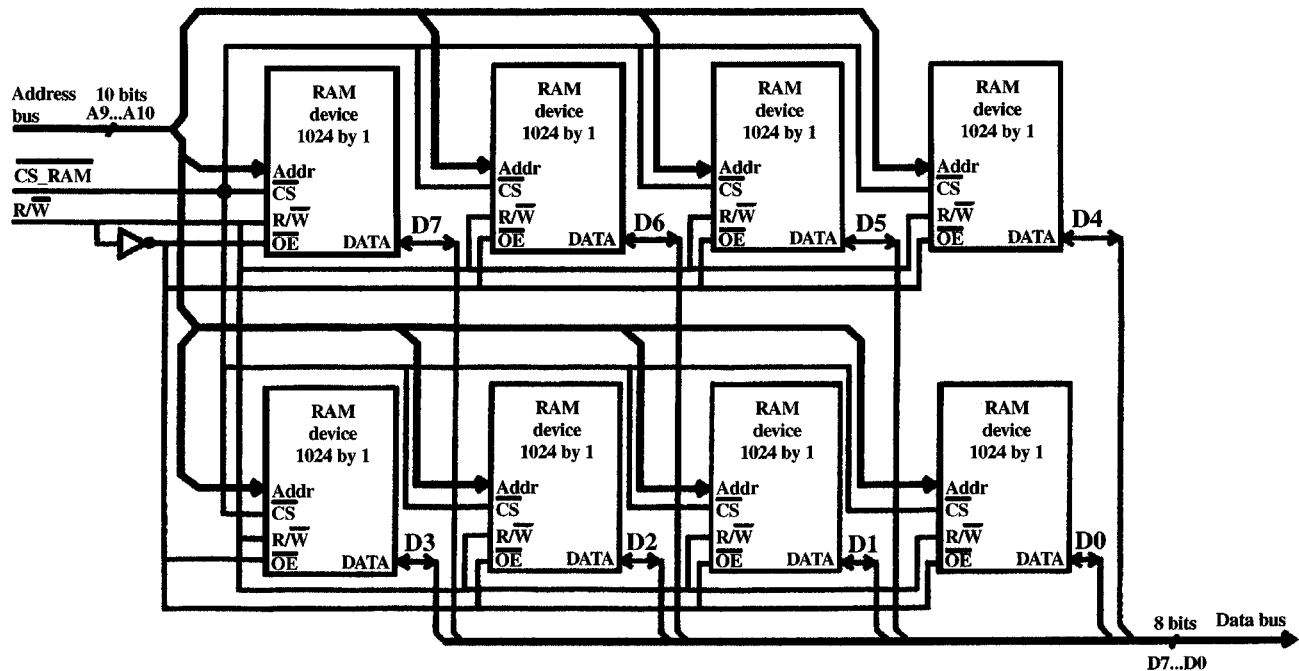


Fig. 121.7 illustrates a simpler memory system design that also addresses 1024 locations each 8 bits wide. In this design 1024-bit RAM devices organized as 1024 words of 1 bit are used. Eight RAM devices must be used to obtain 8 bits of data. The address bus supplies all the RAM devices simultaneously without the need for an external address decoder, thus simplifying the design. Expanding the memory to 8192 addresses using instances of this design requires an external address decoder for chip select logic.

The simplest design for a memory system that addresses 1024 locations each 8 bits wide uses a single 8192-bit RAM device organized as 1024 words of 8 bits. In this design the address bus and the data bus are connected directly to the RAM device. Cost, speed, printed circuit board space, and so forth are factors that need to be considered in choosing the best memory system design for a particular application.

## Defining Terms

**Bus contention:** A condition that occurs on a bus when two or more devices try to output opposite logic levels on the same bus line.

**Bidirectional data bus:** A bus that allows data to flow in either direction.

**Bit mapped:** A graphic display that has 1 bit (or more) of display memory for each possible pixel or dot on the display.

**Cache memory:** A high-speed memory that can contain small segments of a program that can be accessed faster than primary or bulk memory.

**Decoder:** A logic device that converts a binary code applied to  $i$ -input lines to  $2^i$  different output lines.

**Memory map:** A table that specifies the location or address of every memory or I/O device that a

CPU can access.

**Read-only memories:** Memories that are preprogrammed with data that can only be read under normal operation.

**Three-state buffers:** Buffers that have a third output state—that is, a state that is off, or high impedance, to effectively disconnect the outputs—in addition to the normal high- and low-voltage output states.

**Universal programmer:** A programming unit that can program various varieties of PROMs, erase EEPROMs, and program programmable logic array (PLA) and programmable array logic (PAL) devices.

## References

- Gaonkar, R. S. 1984. *Microprocessor Architecture, Programming and Applications with the 8085/8080A*. Merrill, Columbus, OH.
- Gibson, G. A. 1991. *Computer Systems Concepts and Design*. Prentice Hall, Englewood Cliffs, NJ.
- Greenfield, J. D. and William, C. W. 1988. *Using Microprocessors and Microcomputers, the Motorola Family*, 2nd ed. John Wiley & Sons, New York.
- Hayes, J. P. 1988. *Computer Architecture and Organization*, McGraw-Hill, New York.
- Hitachi. *Hitachi IC Memories DataBook*. #M11. Hitachi America, Ltd., San Jose, California.
- Motorola, 1990. *Motorola Memories*. DL113 REV 6. Motorola, Phoenix, Arizona.
- Pollard, H. L. 1990. *Computer Design and Architecture*. Prentice Hall, Englewood Cliffs, NJ.
- Slater, M. 1989. *Microprocessor-Based Design, A Comprehensive Guide to Effective Hardware Design*. Prentice Hall, Englewood Cliffs, NJ.
- Wear, L. L., Pinkert, J. R., Wear, C. W., and Land, W. G. 1991. *Computers, An Introduction to Hardware and Software Design*. McGraw-Hill, New York.

## Further Information

- Dorf, R. C. (Ed.) 1993. *The Electrical Engineering Handbook*. CRC Press, Boca Raton, FL.
- IEEE Transactions on Education*, published quarterly by the Institute of Electrical and Electronic Engineers.
- Rigby, W. H. and Dalby, T. 1995. *Computer Interfacing, A Practical Approach to Data Acquisition and Control*. Prentice Hall, Englewood Cliffs, NJ.

Ciletti, M. D. "Computer-Aided Design and Simulation"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Computer-Aided Design and Simulation

---

## 122.1 Design Flow

## 122.2 Schematic Entry

## 122.3 Hardware Description Languages

Simulation of HDL-Modeled Circuits

## 122.4 Trade-offs between HDLs and Schematic Entry

## 122.5 HDLs and Synthesis

## 122.6 Transistor-Level Design and Simulation

Conclusions

### Michael D. Ciletti

*University of Colorado, Colorado Springs*

The size and complexity of very-large-scale integrated (VLSI) circuits preclude manual design. Designers of VLSI circuits typically use specialized software tools in a workstation-based interactive environment. This chapter reviews important aspects of computer-aided circuit design and simulation and presents an introduction of the use of hardware description languages for design description and design simulation/verification of digital circuits. Digital simulation of analog circuits is also introduced.

## 122.1 Design Flow

---

The design flow of a methodology for designing VLSI circuits consists of a structured sequence of steps, beginning with design entry and culminating in the generation of a database containing geometric detail of the masks that will be used to fabricate the design. These steps can be summarized as follows:

- Create a description of the design (design entry).
- Establish testability of the design.
- Verify the functionality of the design (simulation).
- Develop a gate-level realization of the design.
- Verify the timing specifications.
- Place and route the design.
- Evaluate the timing performance of the routed design.
- Produce database for mask generation.

Multiple passes may be necessary through all or part of this flow. Other design flows are

possible, and this design flow can be modified depending on the particular technology that is being used. For example, formal verification tools may be used in place of simulation, and a synthesis tool may be used to produce the gate-level realization.

Design specifications summarize the functional behavior and timing requirements of the design. They may include power and area constraints and additional information that is relevant to the task. Design entry is the step of encapsulating a representation of the design. This representation may be in a variety of forms, such as a schematic. The testability of the design is generally addressed early in the design process to allow detection of untestable circuitry. The addition of hardware might be required, such as the insertion of a scan path into a sequential circuit. If the design is already in a form that is bound to a particular hardware realization, the timing specifications of the design can be verified by a static analysis of the paths in the circuit or by simulation. The place and route step may involve a full-custom layout of high-performance circuitry or semicustom layout of standard cells or gate arrays (field- or mask-programmable). In either case the physical layout must be re-verified to confirm that the implementation not only realizes the desired functional behavior but also meets the externally imposed timing constraints of the design and the timing constraints imposed by storage elements (flip-flops) used in the design itself.

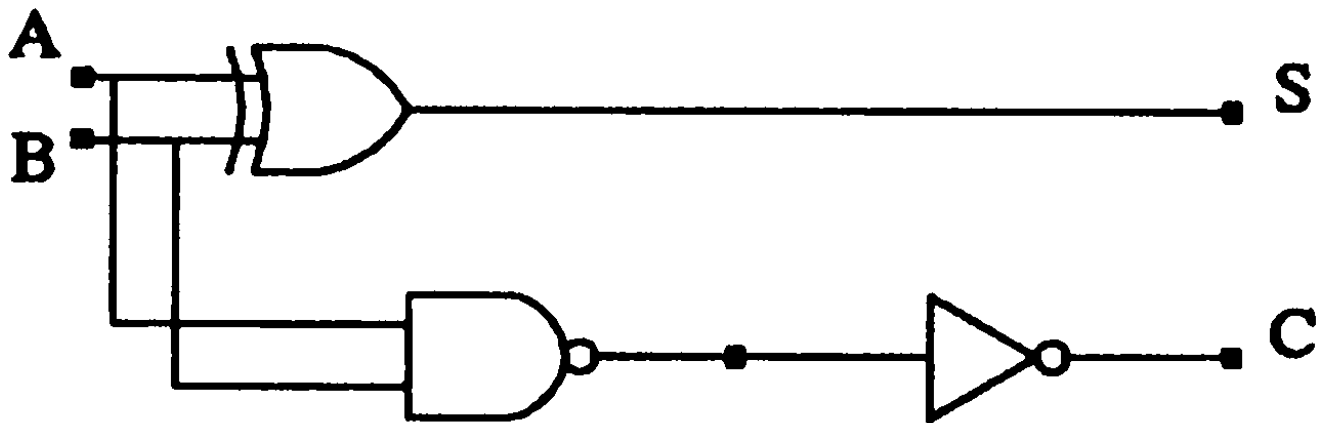
This chapter will focus on two steps in the overall design flow: design entry and simulation. *Design entry* is the process by which a description of the design is encapsulated in a database that serves all subsequent steps in the design flow. A designer may perform this step by drawing paper-and-pencil schematics, by using a schematic entry tool, by using a hardware description language, or by selecting and interconnecting high-level macros representing hardware functional components. The two modes of entry that are of interest here are schematics and hardware description languages.

## 122.2 Schematic Entry

---

A schematic entry tool allows the designer to select and interconnect schematic symbols (icons) representing hardware components. Connections are made by graphical wires and buses representing physical signals in a circuit. The schematic entry software creates and manages a database containing the topological and incidental information created at the schematic, and this software also creates interfaces that will allow other design flow steps to access the database. For example, the information in the topological description can be used by a simulation tool to establish a database for simulating the design represented by the schematic. [Figure 122.1](#) shows a simple schematic of a half-adder logic circuit generated by a schematic entry tool. Schematic entry tools have become popular during the past ten years as the computational power of desktop interactive computer graphics has become capable of supporting complex design tasks at reasonable speeds.

**Figure 122.1** Schematic of a half-adder.



## 122.3 Hardware Description Languages

Schematic entry focuses attention on structural detail of the design and is an appealing mode of entry because engineers are familiar with this traditional visual format. The objects in a schematic representation are high-level blocks and/or gate-level circuits. Structural modeling consists of interconnecting these objects to create a structure that has a desired behavior. A **hardware description language** (HDL) is a computer-based programming language having special constructs and semantics to model, represent, and simulate the functional behavior and timing of digital hardware. An HDL supports structural, behavioral, and mixed descriptions of a design. Unlike schematic-based design, behavioral modeling uses HDL-based constructs and/or procedural code to describe the desired behavior without explicit binding to hardware.

Recently developed HDLs, such as Verilog [Thomas, 1991; Sternheim *et al.*, 1993] and VHDL [Navabi, 1993], provide an alternative mode of design entry by allowing the designer to create a text description of the circuit without relying on a schematic. The text itself can be generated on an ordinary terminal and is very portable. Some examples of Verilog descriptions are given as follows:

```
module Flip_flop(q,data_in,clk,rst);
input data_in,clk,rst;
output q;
always @ (posedge clk)
begin
    if(rst ==1) q = 0;
    else q = data_in;
end
endmodule
```

The basic element of design encapsulation in Verilog is a "module." The code in the flip-flop module declared here updates the value of the output  $q$  whenever the clock has a positive (rising)



edge, provided that the reset line is not asserted.

Verilog also supports a *register transfer logic* (RTL) description with several built-in language operators. The use of RTL is illustrated as follows:

```
module bitwise_or (y,A,B); // RTL model
  input [7:0] A,B;
  output [7:0] y;
  assign y = A | B;
endmodule
```

This bitwise-or module uses the built-in operator "|" to implement the "bitwise or" of the data words. The "assign" keyword effects an event-scheduling rule that updates the value of y whenever A or B changes. An event is said to occur whenever a signal changes value. Note that the Verilog language operators used in RTL design (such as the "|" operator) have implied logic. Their use simplifies the task of writing Verilog descriptions of behavior.

The text that follows declares a model of a half-adder circuit (see [Fig. 122.1](#)) as a structural connection of CMOS standard cells xorf201, nanf201, and invf101.

```
module Add_half_structural (S,C,A,B);
  output S,C;
  input A, B;
  wire C_bar;
  xorf201 G1 (S,A,B);
  nanf201 G2 (C_bar,A,B);
  invf101 G3 (C,C_bar);
endmodule
```

The description declares the name of the module, the input and output ports of the module, a wire that is used internally for a connection, and a list of instantiations of the library cells/ modules. The arguments of the instantiated modules correspond directly to the physical wires that would interconnect their hardware counterparts.

HDLs allow a design to be represented abstractly, or behaviorally, without any binding to particular hardware elements. The fragments of text that follow show two alternative descriptions of a half-adder. The first fragment uses built-in language operators in an RTL style to implement a 4-bit-slice adder. (Here { } denotes a concatenation operator, which in this example creates a 5-bit wide data path from operations on the 4-bit buses and the c\_in bit.) Verilog features built-in data types and operators that make this style of modeling very easy to implement.

```
module adder_4_RTL (sum, c_out, a,b,c_in);
  output [3:0] sum;
  output c_out;
  input [3:0] a,b;
  assign {c_out,sum} = a + b + c_in
```

**endmodule**

The next description is a fragment of procedural code that implements a behavioral model of a carry look-ahead adder [[Sternheim \*et al.\*, 1993](#)].

```
module add_4_CLA (sum, c_out, a,b,c_in);
  output [3:0] sum;
  output c_out;
  input [3:0] a,b;
  reg [3:0] carrychain;
  wire [3:0] gen = a & b; // bitwise and (carry generate);
  wire [3:0] prop = a ^ b; // bitwise xor (carry propagate);
  always @ a or b or c_in // event "or"
    begin: carry_generation_block
      integer i;
      carrychain[0] = gen[0] + (prop[0] & c_in);
      for(i = 1; i <= 3; i = i + 1)
        begin
          #0 carrychain[i] = gen[i] + (prop[i] & carrychain[i-1]);
        end
      end
    end
  wire [4:0] shiftedcarry = {carrychain, c_in};
  wire [3:0] sum = prop ^ shiftedcarry;
  wire c_out = shiftedcarry[4];
endmodule
```

A module that is modeled by procedural code has no predetermined binding to hardware. Admittedly, built-in language operators, such as  $+$ , have an implicit binding to hardware, but this binding can be deferred to later in the design flow. This allows the designer to focus on functionality rather than implementation. Both Verilog and VHDL support algorithmic description of behavior through the mechanism of procedural code, which executes serially.

CAD/CAE software vendors provide development and debug environments supporting the designer of HDL-based descriptions of a design. These may include text editors and interactive debuggers.

VHDL, the VHSIC (very-high-speed integrated circuit) HDL, was created under the auspices of the Department of Defense and is now an IEEE standard (IEEE 1076-1987). Verilog was created as a proprietary language, became very popular as a widely used industry tool, and then was placed in the public domain in 1990. It is presently in the final stages of becoming an IEEE standard [IEEE 1364]. Both languages support high-level abstract descriptions of digital systems; Verilog also has built-in gate-level and switch-level functional primitives.

## Simulation of HDL-Modeled Circuits

Simulators are available for simulating the behavior of digital circuits described by either Verilog or VHDL. These simulators provide a fast, efficient, visual representation of the behavior of a digital circuit. Logic simulation is usually done on event-driven simulators, which exploit the topological latency that is characteristic of digital circuits.

Simulation of a circuit that has been modeled by an HDL requires that a *design unit test bench* (DUTB) be developed to apply stimulus to the unit under test (UUT). A test bench for a nand latch (cross-coupled nand primitives) module is given in the following:

```
module DUTB_Nand_latch;
  reg preset, clear;
  wire q, qbar;
  Nand_latch(q, qbar, preset, clear); // Instantiate the UUT
  initial
  begin // Create stimulus to UUT
    $monitor ($time, "preset = %b   clear = %b   q = %b   qbar = %b", preset, clear, q, qbar);
    #10  preset = 0;      clear = 1;
    #10  preset = 1;
    #10  clear = 0;
    #10  clear = 1;
    #10  preset = 0;
    #10  $finish;
  end
endmodule
```

This DUTB applies stimulus to the inputs of the nand latch at intervals of 10 simulation units. The \$monitor system task effects a listing of the output shown in the following table. (Note: In this example the nand latch has a unit delay between an event on one of its inputs and the resulting event on its output. The value "x" denotes an unknown logic value.)

0	preset = x	clear = x	q = x	qbar = x
10	preset = 0	clear = 1	q = x	qbar = x
11	preset = 0	clear = 1	q = 1	qbar = x
12	preset = 0	clear = 1	q = 1	qbar = 0
20	preset = 1	clear = 1	q = 1	qbar = 0
30	preset = 1	clear = 0	q = 1	qbar = 0
31	preset = 1	clear = 0	q = 1	qbar = 1
32	preset = 1	clear = 0	q = 0	qbar = 1
40	preset = 1	clear = 1	q = 0	qbar = 1

In addition to providing the standard output listing, a variety of tools offered by vendors of CAD/CAE software provide graphical output of simulation results.

## 122.4 Trade-offs between HDLs and Schematic Entry

---

There are some trade-offs that can be noted between schematic entry and HDL-based design entry. From the standpoint of support, a schematic-driven paradigm requires a color-graphic workstation (or suitably enhanced PC); language-based entry is easily done at a terminal, and an engineer can work at a remote site without requiring local support.

Editing a design that is described by an HDL can be shorter and simpler than the task of editing a design described by schematics. The task of removing, relocating, and rebinding schematic objects can be time consuming if the schematic itself is very dense and/or complex.

HDLs support a higher level of abstraction than can be described by schematics. A schematic can certainly have a symbol of any functional unit, but this must eventually be represented in terms of lower-level detail that is ultimately expressed as a structural description. An HDL can embody a behavior with no reference whatsoever to structural detail.

Schematic entry focuses the designer's attention on structural detail; a designer using an HDL can focus on structural detail, functional behavior, or a mixture of the two. HDLs support a top-down design methodology (TDM), in which a design is hierarchically decomposed in simpler, hierarchically organized functional units. In Verilog the nested instantiation of modules is the mechanism by which hierarchical decomposition of a design is accomplished. The actual partitioning may be done according to functionality, but no restriction is implied by the language itself. In a TDM the design is created in a top-down fashion; it is verified in a bottom-up sequence, beginning with verification of the lowest levels of the hierarchy and proceeding to verification of the integrated design. The hierarchical decomposition of a 4-bit adder is given below. The top-level module contains four instantiations of full adders, which themselves contain instantiations of half adders and glue logic; the half adders are defined in terms of modules declared in the cell library.

```
module Add_rca_4 (Sum,C_out,A,B,C_in);
  output Sum,C_out;
  input A,B,C_in;
  wire [3:0] Sum,A,B;
  wire C_out,C_in4,C_in3,C_in2,C_in;
  Add_full G1 (Sum[3],C_out,A[3],B[3],C_in4);
  Add_full G2 (Sum[2],C_in4,A[2],B[2],C_in3);
  Add_full G3 (Sum[1],C_in3,A[1],B[1],C_in2);
  Add_full G4 (Sum[0],C_in2,A[0],B[0],C_in);
endmodule

module Add_full(S,C_out,A,B,C_in);
  output S,C_out;
  input A,B,C_in;
  wire S1,C1,C2,C_out_bar;
  Add_half G1 (S1,C1,A,B);
  Add_half G2 (S,C2,S1,C_in);
  norf201 G3 (C_out_bar,C1,C2);
```

```

    invf101 G4 (C_out,C_out_bar);
endmodule
module Add_half(S,C,A,B);
    output S,C;
    input A,B;
    wire C_bar;
    xorf201 G1 (S,A,B);
    nanf201 G2 (C_bar,A,B);
    invf101 G3 (C,C_bar);
endmodule

```

HDLs support rapid prototyping of a design by allowing a designer to focus attention on the functionality of a design rather than its physical/structural implementation. This factor dramatically shortens the time required to create and verify a design. Shortening the design cycle allows more changes to be made in less time, thereby increasing the likelihood that a design error will be found before its effects became widespread. The ease of considering design alternatives in an HDL context can stimulate and encourage consideration of design alternatives.

Tools now exist that automatically create a schematic from the HDL description, resulting in another attractive feature of HDLs. Thus a schematic is actually a by-product of the HDL design flow. This approach to design greatly shortens the design cycle. The designer must, however, conform to a style that ensures synthesizable results.

## 122.5 HDLs and Synthesis

---

The Verilog HDL is a key element in modern design flows that incorporate logic synthesis tools. These tools begin with a behavioral description of the functionality of the design and then create an optimal logic-level description. This description can then be mapped onto a particular technology to meet timing and area constraints.

## 122.6 Transistor-Level Design and Simulation

---

Digital simulation of a circuit's analog waveforms has been popular for decades, especially since the creation of SPICE [Tuinenga, 1988]. This tool uses mathematical descriptions of active and passive circuit elements such as transistors and other components. These models provide a closer approximation of the actual analog waveforms that will be generated in the physical circuit. This additional resolution/detail comes at the price of greatly increased simulation times and memory requirements. The underlying models of the devices themselves are typically more complex than those used to support digital simulation.

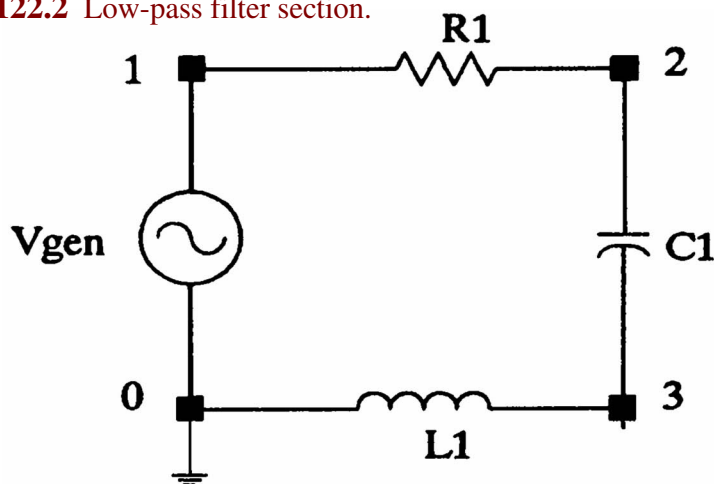
Transistor-level simulators rely on numerical integration techniques, such as the Newton-Raphson method, and exploit the sparsity of the underlying matrices that result from applying Kirchhoff's laws to the possibly nonlinear circuit. In practice these simulators are used effectively to verify the timing performance of critical subcircuits, such as a detailed transient analysis of a memory design, rather than that of an entire system. The text that follows shows a

fragment of SPICE code that describes the simple double-pole low-pass filter section shown in [Fig. 122.2](#).

```
Vgen 1 0 AC 1
R1 1 2 500
L1 2 3 10 mH
C1 3 0 1 uF
.AC DEC 100 100 HZ 10 KHz
.PROBE
.END
```

The first line of the code specifies an independent AC voltage source connected between node 1 and the reference node 0. This source has a level of 1. Each of the next three lines begins with a label specifying the type and identity index of a component, the indices of a pair of nodes to which the component is connected, and the value of the component in default or specified physical units. Then the AC source is specified to have a logarithmic sweep in decades from 100 Hz to 10 kHz in steps of 100 Hz. The .PROBE command is a PSpice [Tuinenga, 1988] statement that creates a database from simulation of the circuit. This database can then be examined to view various features of the response of the circuit. In this example a Bode plot of the filter response will be created.

**Figure 122.2** Low-pass filter section.



The following code describes a transient analysis of the same circuit that was considered earlier:

```
Vgen 1 0 pw1(0,0.1m,1 5.1m, 0)
R1 1 2 500
L1 2 3 10 mH
C1 3 0 1 uF
.TRAN 1m 10m
.PROBE
.END
```

Here, the source is programmed as a piecewise linear waveform describing a pulse that has a 0.1 ms transition from 0 to 1 and lasts for 5 ms. The .TRAN statement specifies that the transient waveform should be plotted in steps of 1 ms beginning at time = 0 (default) and ending at time 10 ms.

SPICE-like software typically supports the following analysis tasks:

- Transient analysis
- Steady state analysis
- Temperature analysis
- Frequency response
- Small-signal transfer function
- Sensitivity analysis
- Thevenin equivalent circuit
- Monte Carlo analysis
- Group delay analysis

Many variations of SPICE are available from CAD/CAE software vendors.

## Conclusions

Powerful software tools exist to support circuits designers. These tools offer substantial gains in productivity in the overall engineering effort.

## Defining Terms

**Hardware description language:** A computer-based programming language having special constructs and semantics to model, represent, and simulate the functional behavior and timing of digital hardware.

**SPICE:** A software language used to create digital simulations of analog circuits. The acronym stands for *simulation program with integrated circuit emphasis*. SPICE was developed at the University of California, Berkeley; it is a public domain tool.

**Verilog:** An HDL that is widely used in industry to describe and simulate digital systems. It is in the final stages of becoming an IEEE standard. It supports hierarchical decomposition, switch- and gate-level structural modeling, RTL/data flow modeling, and procedural modeling, including concurrent activity flows. It has built-in data types.

**VHDL:** An HDL that was developed under the support of the Department of Defense and is an IEEE standard. It supports hierarchical decomposition, data flow, and procedural modeling. It has user-defined data types.

## References

- Navabi, Z. 1993. *VHDL Analysis and Modeling of Digital Systems*. McGraw-Hill, New York.
- Sternheim, E. 1993. *Digital Design and Synthesis with Verilog HDL*. Automata, San Jose, CA.
- Thomas, D. E. and Moorby, P. 1991. *The Verilog Hardware Description Language*. Kluwer Academic, Boston, MA.
- Tuinenga, P. W. 1988. *SPICE: A Guide to Circuit Simulation & Analysis Using PSpice*. Prentice Hall, Englewood Cliffs, NJ.

## Further Information

For additional information and examples of design using the Verilog HDL, see *Digital Design and*

*Synthesis with Verilog HDL*, by Sternheim. For comprehensive information about VHDL, see the *IEEE Standard VHDL Language Reference Manual*, published by IEEE, 345 East 47th St., New York, 1988. For additional information about recent research in computer-aided design, including simulation tools, see the *IEEE Transactions on Computer-Aided Design of Circuits and Systems*. For information on the mathematical models in SPICE-like software, see *Computer-Aided Analysis of Electronic Circuits: Algorithms and Computational Techniques*, published by Prentice Hall, Englewood Cliffs, NJ, 1975.



Mourad, S., Haydt, M. S. "Logic Analyzers"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

- 123.1 Nature of Digital Signals
- 123.2 Signal Sampling
- 123.3 Timing Analysis
- 123.4 State Analysis
- 123.5 Components of a Logic Analyzer
- 123.6 Advanced Features of Logic Analyzers
- 123.7 Applications of Logic Analyzers

A Paradox

**Samiha Mourad**

*Santa Clara University*

**Mary Sue Haydt**

*Santa Clara University*

There are several types of equipment that are used to verify and test digital circuits. The type depends on the design itself and the stage at which it is tested. Logic analyzers are "testers" that combine general purpose characteristics with ease of use. They allow engineers to measure digital signals in a fashion similar to oscilloscopes; the  $x$  axis represents the time and the  $y$  axis, the voltage sampled. Logic analyzers evolved from oscilloscopes in the early 1970s; in a later section, we will distinguish between the use of analyzers and the use of oscilloscopes. Logic analyzers come in various sizes and with a variety of features. They may be used to measure one signal or several signals simultaneously. To explain how they work, it is important to present some basic concepts. The following sections will first describe the nature of digital signals, then explain what an analyzer is and how it is used in testing digital circuits.

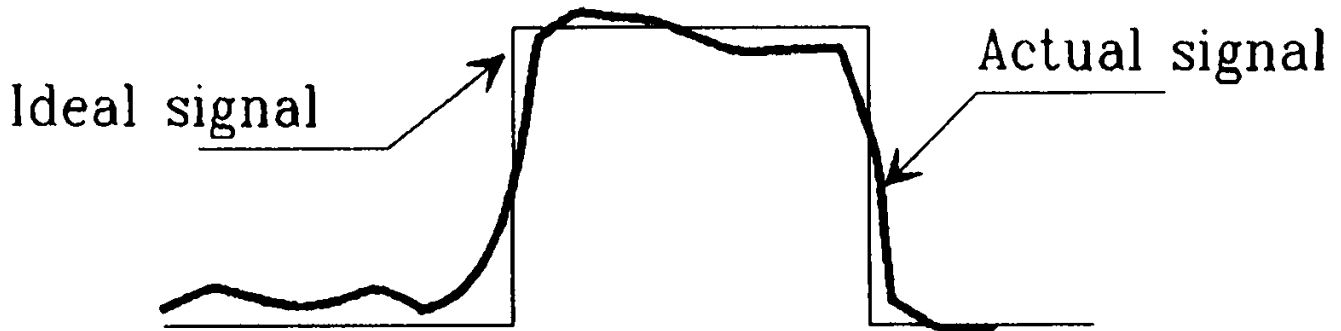
## **123.1 Nature of Digital Signals**

---

Digital signals are waveforms that assume either a low-limit  $V_L$  (logic state 0) or a high-limit  $V_H$  (logic state 1). The exact values of these two limits are technology-dependent. An example of a digital signal is shown in [Fig. 123.1](#). In its ideal form it is a square wave but not necessarily periodic. Due to delays within the devices and interconnect, the rise and fall of the signal are not instantaneous. For a typical pulse, the rate of rise (or fall) is usually called the *slew rate*. The higher the slew rate is, the faster the operation of the circuit. Proper operation of digital circuits is

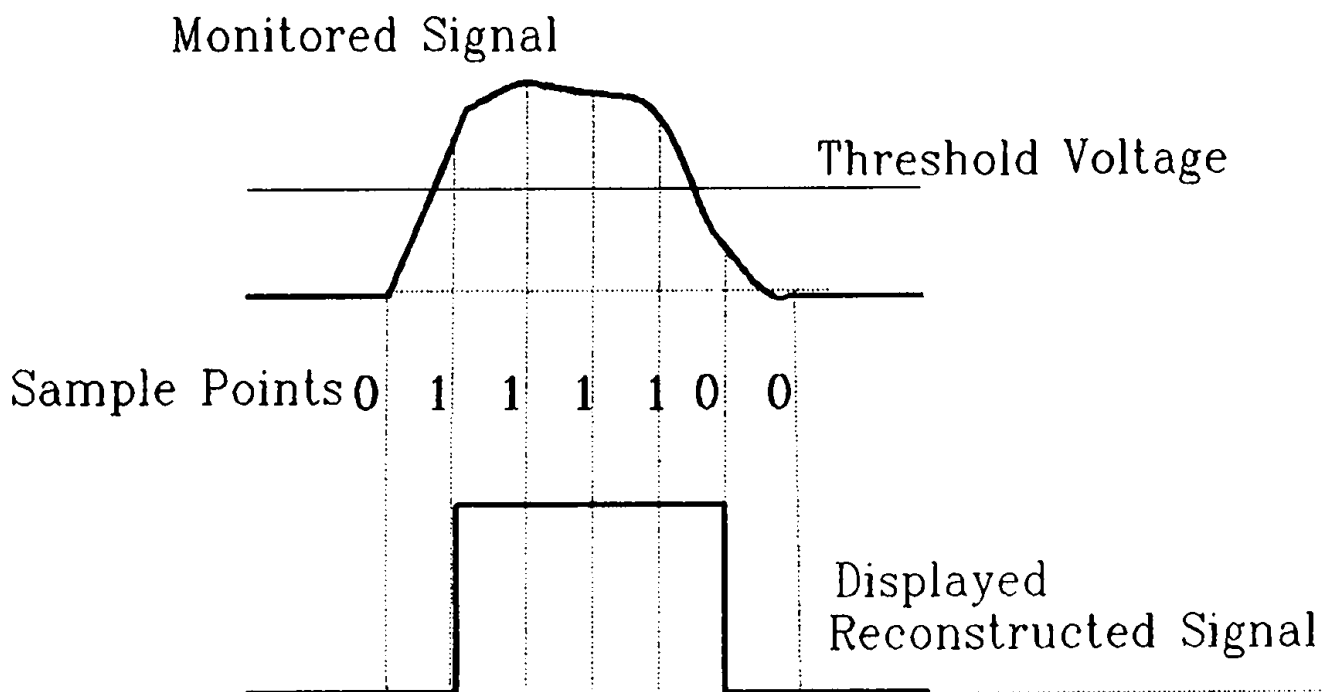
dependent on the arrival of the waveform at various nodes of the circuit at the appropriate time.

**Figure 123.1** Digital signals: ideal versus actual.



Verifying the correctness of a digital signal involves state and timing measurement. The state is the value of the signal at any time, either state 0 or state 1. If the signal is above a certain reference  $V$ , it is logic 1; if it is below this reference, it is logic 0. This is illustrated in Fig. 123.2. For the logic analyzer to reconstruct the waveform examined, measurements need to be taken at various time instants. The collection of such measurements forms a sample that is interpreted by the analyzer, displayed, and possibly stored for further analysis. This process is called **sampling** and is done at various sampling points. In a sense, sampling is like taking snapshots of the waveform. Sampling also serves in time analysis. In the remainder of this chapter we will learn how the analyzer works in both state and timing capacities, but first let us understand sampling and its importance in logic analyzer operations.

**Figure 123.2** Sampling a waveform and reconstructing it.



## 123.2 Signal Sampling

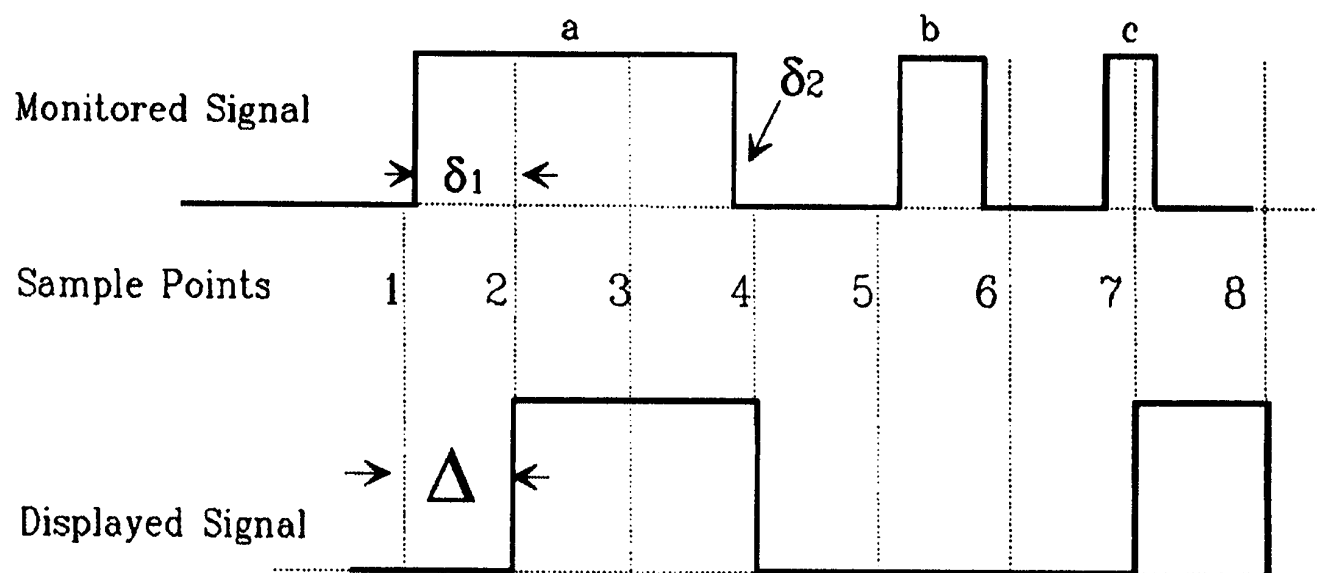
Figure 123.2 shows a digital waveform, the sampling points, and the possible displays of the analyzer. The collected sample is stored in a buffer and used to reconstruct the waveform as illustrated in the figure. Sampling is a very important process in the operation of logic analyzers. If the signal is changing at 1MHz, then it alternates between high and low every  $T = 0.5 \mu\text{s}$ . To accurately reconstruct this signal, we must sample it at least every  $0.5 \mu\text{s}$ . That is, the sampling rate has to be at least double that of the rate at which the signal is changing. The size of the sample is limited to the size of the buffer. A normal buffer size contains 1K (1024) sample points. For a sampling rate of 2 megasamples per second, the buffer will hold data for  $512 \mu\text{s}$ .

The logic analyzer is always sampling. When the user triggers it for display, it will show the waveform before and after the **trigger**. The triggering is done in one of two modes: level and edge. Both modes will be described later in conjunction with timing and state analysis.

## 123.3 Timing Analysis

Time analysis is very crucial to the performance of digital circuits. For proper functioning of a circuit the signal must arrive at various nodes at precise times within a tolerance. Whenever the state changes from 0 to 1 or vice versa for two consecutive sampling points, the analyzer recognizes that the signal went through a transition during the period between these two points. This transition is then interpreted as having occurred at the second sampling instant, as illustrated in Fig. 123.3. However, there is uncertainty about where the transition has really occurred. The first transition, 0 to 1, occurred just after sample point 1, but it is only recorded at sample point 2. Similarly, the transition from 1 to 0 occurred right before sample point 4, where it was recorded. Since  $\delta_1 > \delta_2$ , the recording at sample point 4 is more accurate than at point 2.

**Figure 123.3** Sampling accuracy.

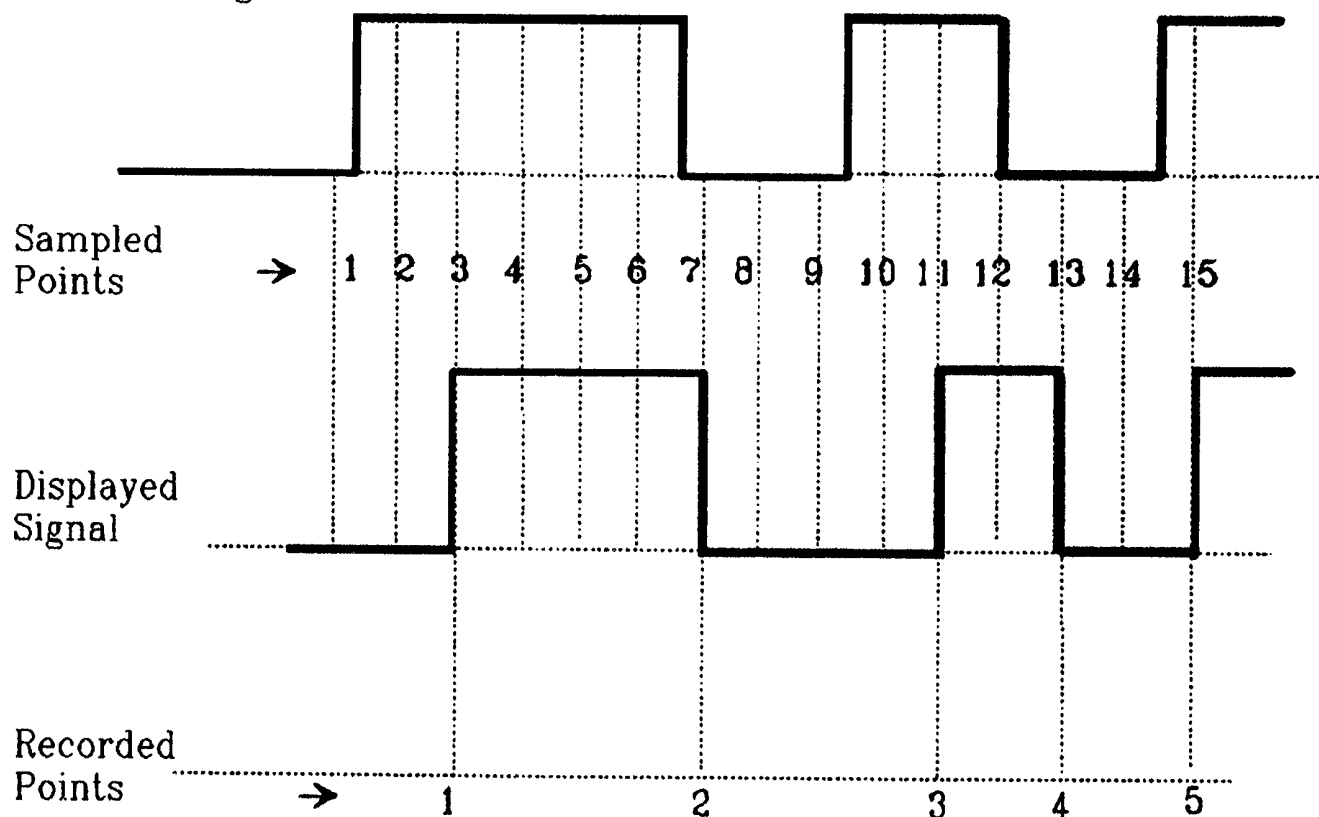


The interval of time between two sample points, the *sampling period*, denoted in the figure by  $\Delta$ , is the maximum uncertainty in timing measurement. Thus the shorter the sampling period is, the higher is the resolution. This period has to be smaller than any pulse width measured by the logic analyzer. Otherwise, the pulse may not be recorded at all, or it might be recorded as having the sampling period as its width. Both cases are shown in Fig. 123.3 for the pulses b and c.

Increasing the sampling rate for a certain **buffer depth** results in a narrower window. Alternately, keeping the same window requires a *deeper buffer*. It is possible to reduce the buffer depth by storing only the sample points following transitions, as illustrated in Fig. 123.4. Of all fifteen sample points, only five are stored. Since the analyzer is sampling digital data, only the transition is relevant. This scheme is called **transitional sampling**.

**Figure 123.4** Transitional sampling.

Monitored Signal



Logic analyzers are not intended as parametric measuring instruments. They are used to determine timing relationship. We will consider such use of logic analyzers later in the chapter. Next, we will examine the second type of analysis—state analysis.

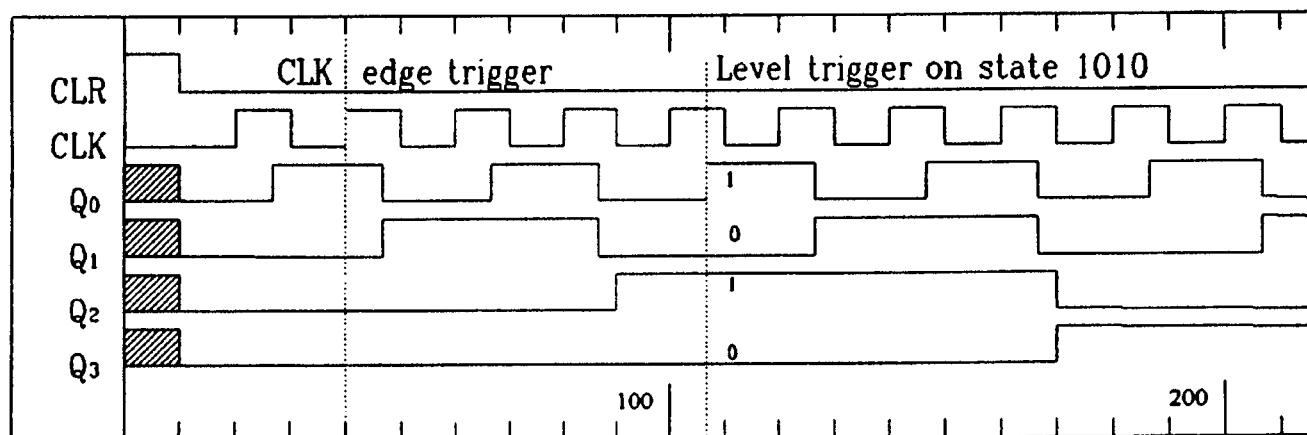
## 123.4 State Analysis

In an earlier section we used the term *state* to refer to the value of the signal on a certain node of the circuit. However, the advantage of the logic analyzer is in providing a means to determine the signal value on several nodes of the circuit. The **state** of a certain circuit is the collection of signals (0 or 1) on some lines of the circuit. The set of nodes may be the lines of a bus or the output of a counter.

The operation of the majority of digital circuits is synchronized by one signal called the *clock*. As the clock transitions from 0 to 1 (positive edge) or from 1 to 0 (negative edge), the signals propagate through the circuit and stabilize before the next clock edge. The clock is usually periodic, with duty cycle less than 50%. Thus, to observe the state of a circuit, it is convenient to use the system clock as the reference. The logic analyzer will capture the state of the circuit as the edge of the clock occurs. This mode of operation is called *edge triggering*.

Consider the example shown in Fig. 123.5 for a 4-bit binary counter. The count is read on  $Q_1 Q_2 Q_3 Q_4$ . These nodes form the state of the counter. Displayed also are the system clock and an asynchronous clear that returns the counter to 0 whenever asserted.

**Figure 123.5** Edge- and level trigger.



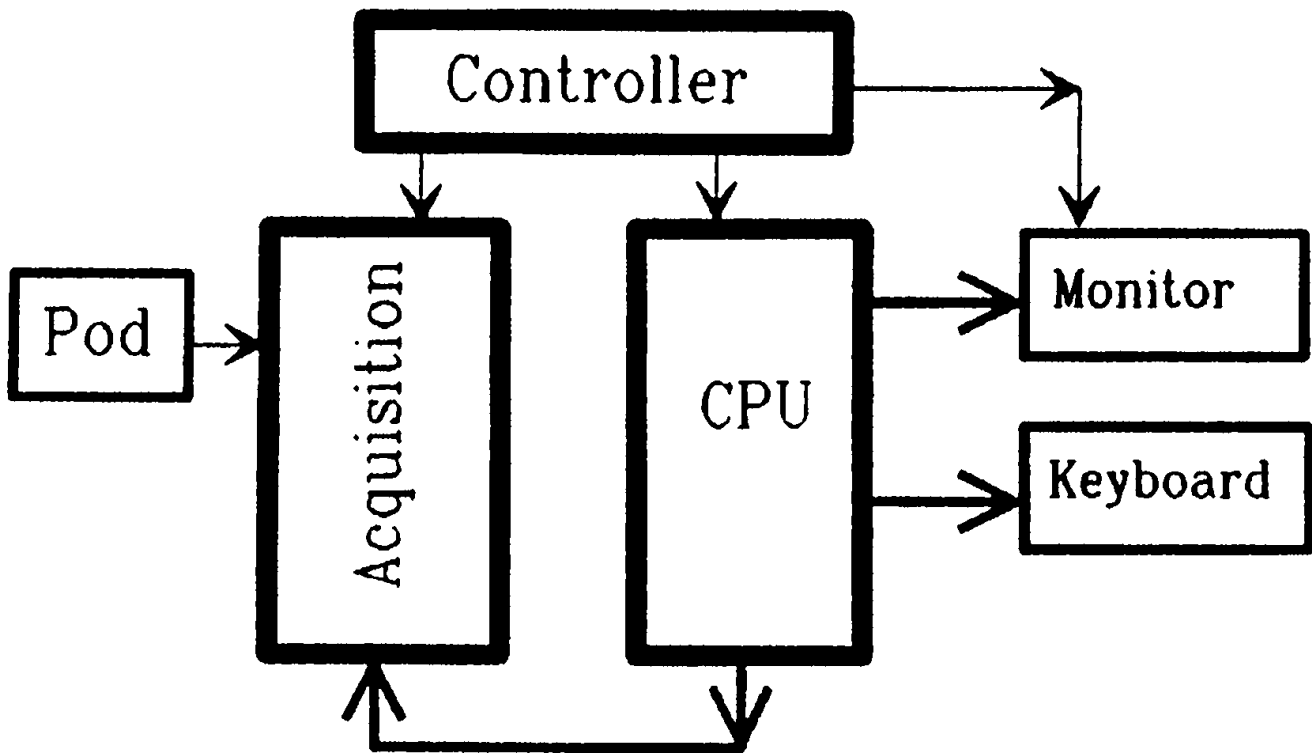
It is also possible to request the data when the circuit is in a certain state, say 1010. In such a case, *level trigger*—specification of a level to start the capture of the data—is used. Every signal of the state—here, four—is measured by a channel. Besides observing the display, users may need to examine the values. For this 4-bit counter, using binary notations is manageable. It would be difficult, though, to use this representation for a 32-bit bus; instead, hexadecimal numbers would be used.

Now that we are familiar with the operation of logic analyzers, we will briefly describe its components.

## 123.5 Components of a Logic Analyzer

A logic analyzer consists of the functional blocks shown in Fig. 123.6: a data acquisition block; a data analyzer block that is part of the CPU; a memory; a display; and several ports that allow connection to a PC, printer, and other instruments. The interface between the data acquisition and the unit under test (UUT) is called the *pod*. The CPU is microprocessor based, but the data acquisition is usually an application-specific integrated circuit (ASIC). As IC technology advances the size and the cost of analyzers decreases. Portable analyzers may be used for on-site testing.

**Figure 123.6** Components of a logic analyzer.



## 123.6 Advanced Features of Logic Analyzers

---

- Modern logic analyzers interface with PCs for postprocessing of the data. This data can be edited by the PC editor and also exported to a spreadsheet for further analysis.
- A logic analyzer can be bundled with a digitizing oscilloscope and thus work as three instruments in one—for timing analysis, state analysis, and parametric measurements. For intermodule analysis the data is viewed simultaneously by the three modules.
- Utilities are also available to translate state analyses files into test pattern generation files to minimize time-consuming data entries.

## 123.7 Applications of Logic Analyzers

---

Although logic analyzers bear resemblance to digitizing oscilloscopes, their applications are not identical. Scopes are usually used when high fidelity is necessary in replicating every variation, no matter how small, in the waveform, and when timing between two or more events needs to be measured with high accuracy. The scope is suitable for parametric measurements.

Logic analyzers, on the other hand, are not as accurate in measuring time and voltages. They do have, however, many other equally important assets that have helped advance the state of the art of digital design. In the digital domain the signal is either high or low; the ripples in the waveform profile are of no consequence to proper operation of the product. It is more important to show that the actual waveforms of the circuit are as the designer expected. Logic analyzers are also useful in determining timing relationships among data lines of a bus or any group of various nodes in the UUT. This fact was illustrated earlier with the example of the 4-bit counter.

Logic analyzers are used for design verification during the development and integration cycle of digital systems. They are also useful in production testing. The product itself may be a microprocessor-based circuit or an ASIC.

The effectiveness of the logic analyzer depends on its characteristics. For example, analyzers with only a few channels are not efficient for testing a large circuit. If the maximum speed of an analyzer is under that of the UUT, it is useless for such a UUT. Also, logic analyzers must have sufficient capacity for data acquisition.

## A Paradox

Logic analyzers are built of digital ICs that need to function at a speed higher than that of the UUT. They are used to develop tomorrow's ICs that are supposed to have higher performance than the products of today. To build a design tool for a customer who uses state-of-the-art devices, one must have the next-generation devices to make the tool effective for the present generation. And that is physically impossible.

## Defining Terms

**Buffer depth:** The size of the acquisition memory.

**Sampling:** The process of recording signals at specific moments, called *sample points*, over a period of time.

**State:** The logic values of a collection of nodes in the circuit that depend on the transition in the clock of the circuit.

**Transitional sampling:** Storing samples indicating change of level.

**Trigger:** Signal entered by the user to flag the acquisition.

## References

- Hewlett-Packard. 1988. *Feeling Comfortable with Logic Analyzers*. Part # 5954-2686. Hewlett-Packard, Palo Alto, CA.
- Miner, G. F. and Comer, D. J. 1992. *Physical Data Acquisition for Digital Processing: Components, Parameters, and Specifications*. Prentice Hall, Englewood Cliffs, NJ.

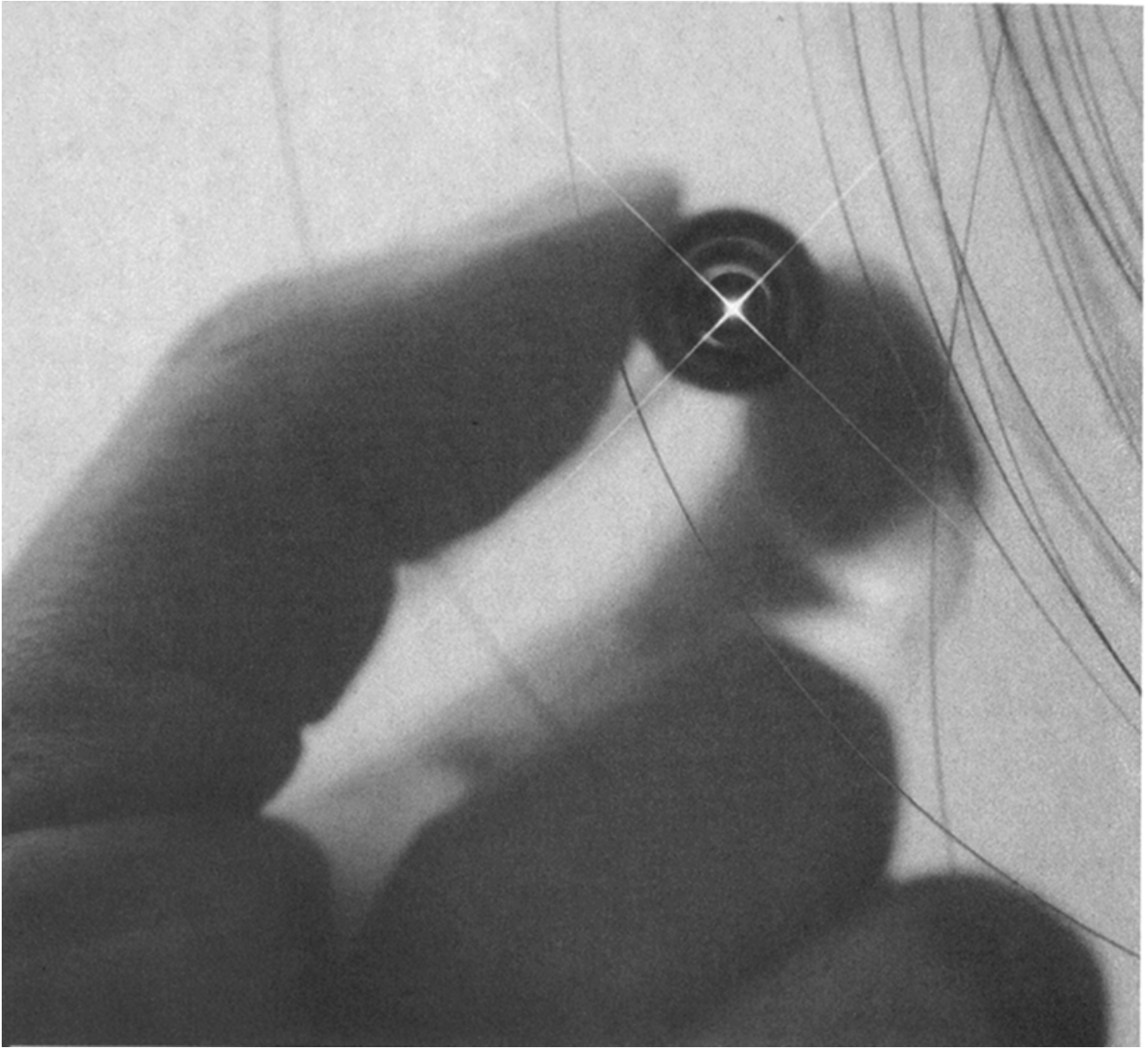
## Further Information

- Bauer, N. 1989. Logic analyzers vs. IC design verification systems, *Test & Meas. World*. p. 21.
- DeSena, A. 1980. Logic analyzers, new capabilities, and challenges. *Electronic Test*. pp. 24-27.



- Editorial Staff. 1991. Designing a logic analyzer to cost and market needs. *Electron. Eng.* pp. 41-44.
- Hewlett-Packard. 1986. *Bandwidth and Sampling Rate in Digitizing Oscilloscopes*. Application Note 344. Hewlett-Packard, Palo Alto, CA.
- Jacob, G. 1990. Versatile trigger and data interpretation. *Eval. Eng.* October.
- Jacob, G. 1991. Faster processors place demands on logic analyzers. *Eval. Eng.* October.
- Jacob, G. 1992. Analyzers meet continuing "wider-deeper-faster" demands. *Eval. Eng.* October.

Poor, H. V. "Communications and Signal Processing"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



Hair-thin fibers of ultrapure glass are now transmitting voice, data, and video communications in many parts of the globe in the form of digital pulses emitted by semiconductor lasers the size of a grain of salt. Such fiber-optic systems are now capable of transmitting a half million simultaneous conversations. (Photo courtesy of AT&T Bell Laboratories.)

# Communications and Signal Processing

---

**H. Vincent Poor**

*Princeton University*

- 124 **Transforms and Fast Algorithms** *A. D. Poularikas*  
Fourier Transforms • Walsh-Hadamard Transform
- 125 **Digital Filters** *B. W. Bomar and L. M. Smith*  
Finite Impulse Response Filter Design • Infinite Impulse Response Filter Design • Digital Filter Implementation
- 126 **Modulation and Detection** *H. V. Poor*  
Analog Modulation and Detection • Digital Modulation and Detection • Further Issues
- 127 **Coding** *S. L. Miller and L. W. Couch II*  
Block Codes • Convolutional Codes • Trellis-Coded Modulation
- 128 **Computer Communication Networks** *J. N. Daigle*  
General Networking Concepts • Computer Communication Network Architecture • Local-Area Networks and Internets • Some Additional Recent Developments
- 129 **Satellites and Aerospace** *S. W. Fordyce and W. W. Wu*  
Communications Satellite Services and Frequency Allocation • Information Transfer and Link Margins—Ground to Space (Up-Link) • Communication Satellite Orbits • Launch Vehicles • Spacecraft Design • Propagation • Earth Stations
- 130 **Mobile and Cellular Radio Communications** *T. S. Rappaport, R. Muhamed, M. Buehrer, and A. Doradla*  
Paging Systems • Cordless Telephone Systems • Cellular Telephone Systems • Personal Communications System (PCS) • The Cellular Concept and System Fundamentals • System Capacity and Performance of Cellular Systems • Mobile Radio Systems Around the World
- 131 **Optical Communications** *J. C. Palais*  
Optical Communications Systems • Topologies • Fibers • Other Components • Signal Quality

TELECOMMUNICATIONS IS ONE OF THE WORLD'S most influential and most rapidly evolving technologies. New and emerging telecommunications services hold the promise of dramatically changing the speed, extent, and connectivity of communications among humans and of information sharing among computers. Such services include wireless personal communications connected globally through satellite networks, and optical fiber communication networks that permit the transmission of data at enormous rates.

Telecommunications technology is based largely on the two closely allied engineering fields of communications and signal processing. Fundamentally, communications deals with the transmission of information through physical channels by electronic means, while signal processing deals with methods of transforming or combining electronic signals into more useful

forms. Thus, the principles of communications provide functional descriptions of ways in which information can be incorporated in and extracted from physical signals to maximize communications capacity, and the principles of signal processing provide efficient transforms, algorithms, and filters to perform these functions.

This section provides information on the underlying principles and some major applications of the two fields of communications and signal processing. The contributions to this section can be grouped into three categories: signal processing principles, communications principles, and communications applications. The first two contributions (**Chapters 124 and 125**) cover the three fundamental areas of signal processing noted above, namely, transforms, algorithms, and filter design. The techniques described in these chapters provide the means by which signals can be treated to perform many of the functions required by communication systems. The next three contributions (**Chapters 126–128**) describe the basic principles underlying three major functional areas of electronic communications: modulation and detection, coding, and networking. Modulation refers to the impression of messages onto signals that are suitable for transmission through a physical communication channel, and detection is the reverse process; coding deals with the treatment of the messages themselves to counteract errors incurred in the communications channel; and networking deals with the issues arising when multiple communicating partners are connected through a shared medium. The final three contributions (**Chapters 129–131**) cover the basics of the three most rapidly developing areas of commercial communications: satellite communications, mobile radio communications, and optical communications.

Poularikas, A. D. "Transforms and Fast Algorithms"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Transforms and Fast Algorithms

---

## 124.1 Fourier Transforms

Properties of the DFT • Relation between DFT and Fourier Transform • Power, Amplitude, and Phase Spectra • Data Windowing • Fast Fourier Transform • Computation of the Inverse DFT

## 124.2 Walsh-Hadamard Transform

Walsh Functions • Walsh-Ordered Walsh-Hadamard Transform (WHT<sub>w</sub>) • Fast Walsh-Ordered Walsh-Hadamard Transform (FWHT<sub>w</sub>)

**Alexander D. Poularikas**

*University of Alabama, Huntsville*

## 124.1 Fourier Transforms

---

One method used extensively calls for replacing the continuous Fourier transform by an equivalent *discrete Fourier transform* (DFT) and then evaluating the DFT using the discrete data. But evaluating a DFT with 512 samples (a small number in most cases) requires more than  $1.5 \cdot 10^6$  mathematical operations. It was the development of the **fast Fourier transform (FFT)**, a computational technique that reduces the number of mathematical operations in the evaluation of the DFT to  $N \log_2(N)$  (approximately  $2.5 \cdot 10^4$  operations for the 512-point case mentioned above), that made DFT an extremely useful tool in almost all fields of science and engineering.

A data sequence is available only within a finite time window from  $n = 0$  to  $n = N - 1$ . The transform is discretized for  $N$  values by taking samples at the frequencies  $2\pi/NT$ , where  $T$  is the time interval between sample points. Hence, we define the DFT of a sequence of  $N$  samples for  $0 \leq k \leq N - 1$  by the relation

$$\begin{aligned}
 F(k-\Omega) &\doteq \mathcal{F}_d\{f(nT)\} = T \sum_{n=0}^{N-1} f(nT) e^{-j2\pi nkT/NT} \\
 &= T \sum_{n=0}^{N-1} f(nT) e^{-j\Omega Tnk}, \\
 n &= 0, 1, \dots, N-1 \quad (124.1)
 \end{aligned}$$

where

$$\begin{aligned}
 N &= \text{number of sample values} \\
 T &= \text{sampling time interval} \\
 (N-1)T &= \text{signal length} \\
 f(nT) &= \text{sampled form of } f(t) \text{ at points } nT \\
 \Omega &= \frac{2\pi}{T} \frac{1}{N} = \frac{\omega_s}{N} = \text{the frequency sampling interval} \\
 e^{-j\Omega T} &= N\text{th principal root of unity} \\
 j &= \sqrt{-1}
 \end{aligned}$$

The inverse IDFT is given by

$$\begin{aligned}
 f(nT) &\doteq \mathcal{F}_d^{-1}\{F(k-\Omega)\} = \frac{1}{NT} \sum_{k=0}^{N-1} F(k-\Omega) e^{j2\pi nkT/NT} \\
 &= \frac{1}{NT} \sum_{k=0}^{N-1} F(k-\Omega) e^{j\Omega Tnk} \quad (124.2)
 \end{aligned}$$

The sequence  $f(nT)$  can be viewed as representing  $N$  consecutive samples  $f(n)$  of the continuous signal, whereas the sequence  $F(k-\Omega)$  can be considered as representing  $N$  consecutive samples  $F(k)$  in the frequency domain. Therefore, Eqs. (124.1) and (124.2) take the compact form

$$\begin{aligned}
 F(k) &\doteq \mathcal{F}_d\{f(n)\} \\
 &= \sum_{n=0}^{N-1} f(n) e^{-j2\pi nk/N} = \sum_{n=0}^{N-1} f(n) W_N^{nk}, \\
 k &= 0, \dots, N-1 \quad (124.3)
 \end{aligned}$$

$$\begin{aligned}
 f(n) &\doteq \mathcal{F}_d^{-1}\{F(k)\} \\
 &= \frac{1}{N} \sum_{k=0}^{N-1} F(k) e^{j2\pi nk/N} = \sum_{k=0}^{N-1} F(k) W_N^{-nk}, \\
 n &= 0, \dots, N-1 \quad (124.4)
 \end{aligned}$$

where  $W_N = e^{-j2\pi/N}$ , and  $j = \sqrt{-1}$ . An important property of the DFT is that  $f(n)$  and  $F(k)$  are uniquely related by the transform pair in Eqs. (124.3) and (124.4).

We observe that the functions  $W^{kn}$  are  $N$ -periodic; that is,



$$W_N^{kn} = W_N^{(k+N)n} = W_N^{k(n+N)}, \quad k, n = 0, \pm 1, \pm 2, \dots \quad (124.5)$$

As a consequence, the sequence  $f(n)$  and  $F(k)$  as defined by Eqs. (124.3) and (124.4) are also  $N$ -periodic.

It is generally convenient to adopt the convention

$$\{f(n)\} \longleftrightarrow \{F(k)\} \quad (124.6)$$

to represent the transform pair in Eqs. (124.3) and (124.4).

## Properties of the DFT

A detailed discussion of the properties of DFT can be found in the references cited at the end of this chapter. The following list gives a few of these properties that are of value for the development of the fast Fourier transform.

### 1. Linearity

$$\{af(n) + by(n)\} \longleftrightarrow \{aF(k) + bY(k)\} \quad (124.7a)$$

### 2. Complex conjugate. If $N/2$ is an integer and $\{f(n)\} \longleftrightarrow \{F(k)\}$ , then

$$F\left(\frac{N}{2} + \ell\right) = F^*\left(\frac{N}{2} - \ell\right), \quad \ell = 0, 1, \dots, \frac{N}{2} \quad (124.7b)$$

where  $F^*(k)$  denotes the complex conjugate of  $F(k)$ . This identity shows the folding property of the DFT.

### 3. Reversal

$$\{f(-n)\} \longleftrightarrow \{F(-k)\} \quad (124.8)$$

### 4. Time shifting

$$\{f(n + \ell)\} \longleftrightarrow \{W^{-\ell k} F(k)\} \quad (124.9)$$

### 5. Convolution of real sequences. If

$$y(n) = \frac{1}{N} \sum_{\ell=0}^{N-1} f(\ell)h(n - \ell), \quad n = 0, 1, \dots, N - 1 \quad (124.10)$$

then

$$\{y(n)\} \longleftrightarrow \{F(k)H(k)\} \quad (124.11)$$

6. *Correlation of real sequences.* If

$$y(n) = \frac{1}{N} \sum_{\ell=0}^{N-1} f(\ell)h(n + \ell), \quad n = 0, 1, \dots, N - 1 \quad (124.12)$$

then

$$\{y(n)\} \longleftrightarrow \{F(k)H^*(k)\} \quad (124.13)$$

7. *Symmetry*

$$\left\{ \frac{1}{N} F(n) \right\} \longleftrightarrow \{f(-k)\} \quad (124.14)$$

8. *Parseval's theorem*

$$\sum_{n=0}^{N-1} f^2(n) = \frac{1}{N} \sum_{k=0}^{N-1} |F(k)|^2 \quad (124.15)$$

where  $|F(k)| = F(k)F^*(k)$  .

**Example 124.1.** Verify Parseval's theorem for the sequence  $\{f(n)\} = \{1, 2, -1, 3\}$  .

**Solution.** With the help of Eq. (124.3) we obtain

$$\begin{aligned} F(k)|_{k=0} &= F(0) = \sum_{n=0}^3 f(n)e^{-j\frac{2\pi}{4}kn} \Big|_{k=0} \\ &= (1e^{-j\frac{\pi}{2}\cdot 0\cdot 0} + 2e^{-j\frac{\pi}{2}\cdot 0\cdot 1} - e^{-j\frac{\pi}{2}\cdot 0\cdot 2} + 3e^{-j\frac{\pi}{2}\cdot 0\cdot 3}) \\ &= 5 \end{aligned}$$

Similarly, we find

$$F(1) = 2 + j; \quad F(2) = -5; \quad F(3) = 2 - j$$

Introducing these values in Eq. (124.15) we obtain

$$\begin{aligned}
1^2 + 2^2 + (-1)^2 + 3^2 &= \\
\frac{1}{4}[5^2 + (2+j)(2-j) + 5^2 + (2-j)(2+j)] & \\
\text{or} & \\
15 &= \frac{60}{4}
\end{aligned}$$

which is an identity as it should have been.

## Relation between DFT and Fourier Transform

The sampled form of a continuous function  $f(t)$  can be represented by the  $N$  equally spaced sampled values  $f(n)$  such that

$$f(n) = f(nT), \quad n = 0, 1, \dots, N-1 \quad (124.16)$$

where  $T$  is the sampling interval. The length of the continuous function is  $L = NT$ , where  $f(N) = f(0)$ .

We denote the sampled version of  $f(t)$  by  $f_s(t)$ , which can be represented by the expression

$$f_s(t) = \sum_{n=0}^{N-1} [Tf(n)]\delta(t - nT) \quad (124.17)$$

where  $\delta(t)$  is the Dirac or impulse function.

Taking the Fourier transform of  $f_s(t)$  in Eq. (124.17) we obtain

$$\begin{aligned}
F_s(\omega) &= T \int_{-\infty}^{\infty} \sum_{n=0}^{N-1} f(n)\delta(t - nT)e^{-j\omega t} dt \\
&= T \sum_{n=0}^{N-1} f(n) \int_{-\infty}^{\infty} \delta(t - nT)e^{-j\omega t} dt \\
&= T \sum_{n=0}^{N-1} f(n)e^{-j\omega nT} \quad (124.18)
\end{aligned}$$

Equation (124.18) yields  $F_s(\omega)$  for all values of  $\omega$ . However, if we are interested only in values of  $F_s(\omega)$  at a set of discrete equidistant points, then Eq. (124.18) is expressed in the form [see also Eq. (124.1)]

$$F_s(k) = T \sum_{n=0}^{N-1} f(n) e^{-jkn} \Omega T, \quad k = 0, \pm 1, \pm 2, \dots, \pm \frac{N}{2} \quad (124.19)$$

where  $\Omega = 2\pi/L = 2\pi/NT$ . Therefore, comparing Eqs. (124.3) and (124.19), we observe that we can find  $F(\omega)$  from  $F_s(\omega)$  using the relation

$$F(k) = F_s(\omega)|_{\omega=k\Omega} \quad (124.20)$$

## Power, Amplitude, and Phase Spectra

If  $f(t)$  represents voltage or current waveform supplying a load of 1 ohm, the left-hand side of Parseval's theorem [Eq. (124.15)] represents the power dissipated in the 1-ohm resistor. Therefore, the right-hand side represents the power contributed by each harmonic of the spectrum. Thus the **DFT power spectrum** is defined as

$$P(k) = F(k)F^*(k) = |F(k)|^2, \quad k = 0, 1, \dots, N-1 \quad (124.21)$$

For real  $f(n)$  there are only  $(N/2 + 1)$  independent DFT spectral points as the complex conjugate property shows [Eq. (124.7)]. Hence, we write

$$P(k) = |F(k)|^2, \quad k = 0, 1, \dots, \frac{N}{2} \quad (124.22)$$

The *amplitude spectrum* is readily found from that of a power spectrum, and it is defined as

$$A(k) = |F(k)|, \quad k = 0, 1, \dots, N-1 \quad (124.23)$$

The power and amplitude spectra are invariant with respect to shifts of the data sequence  $\{f(n)\}$ .

The **phase spectrum** of a sequence  $\{f(n)\}$  is defined as

$$\varphi_f(k) = \tan^{-1} \left[ \frac{\text{Im}\{F(k)\}}{\text{Re}\{F(k)\}} \right], \quad k = 0, 1, \dots, N-1 \quad (124.24)$$

As in the case of the power spectrum, only  $(N/2 + 1)$  of the DFT phase spectral points are independent for real  $\{f(n)\}$ . For a real sequence  $\{f(n)\}$  the power spectrum is an *even function* about the point  $k = N/2$  and the phase spectrum is an *odd function* about the point  $k = N/2$ .

## Observations

1. The frequency spacing  $\Delta\omega$  between coefficients is

$$\Delta\omega = -\Omega = \frac{2\pi}{NT} = \frac{\omega_s}{N} \quad \text{or} \quad \Delta f = \frac{1}{NT} = \frac{f_s}{N} = \frac{1}{T_o} \quad (124.25)$$

2. The reciprocal of the record length defines the frequency resolution.
3. If the number of samples  $N$  are fixed and the sampling time is increased, the record length and the precision of frequency resolution is increased. When the sampling time is decreased, the opposite is true.
4. If the record length is fixed and the sampling time is decreased ( $N$  increases), the resolution stays the same and the computed accuracy of  $F(n-\Omega)$  increases.
5. If the record length is fixed and the sampling time is increased ( $N$  decreases), the resolution stays the same and the computed accuracy of  $F(n-\Omega)$  decreases.

## Data Windowing

To produce more accurate frequency spectra it is recommended that the data are weighted by a **window** function  $\{w(n)\}$ . Hence, the new data set will be of the form  $\{f(n)w(n)\}$ . The following are the most commonly used windows:

1. *Triangle (Fejer, Bartlet) window*

$$w(n) = \begin{cases} \frac{n}{N/2}, & n = 0, 1, \dots, \frac{N}{2} \\ w(N-n), & n = \frac{N}{2}, \dots, N-1 \end{cases} \quad (124.26)$$

2.  $\cos^\alpha(x)$  window

$$\begin{aligned} w(n) &= \sin^2 \left( \frac{n}{N} \pi \right), \quad n = 0, 1, \dots, N-1, \quad \alpha = 2 \\ &= 0.5 \left[ 1 - \cos \left( \frac{2n}{N} \pi \right) \right] \end{aligned} \quad (124.27)$$

This window is also called the *raised cosine* or *Hann window*.

3. *Hamming window*

$$w(n) = 0.54 - 0.46 \cos \left( \frac{2\pi}{N} n \right), \quad n = 0, 1, \dots, N-1 \quad (124.28)$$

4. *Blackman window*

$$w(n) = \sum_{m=0}^K (-1)^m a_m \cos \left( 2\pi m \frac{n}{N} \right),$$

$$n = 0, 1, \dots, N-1, \quad K \leq \frac{N}{2} \quad (124.29)$$

For  $K = 2$ ,  $a_0 = 0.42$ ,  $a_1 = 0.50$ , and  $a_2 = 0.08$ .

5. *Blackman-Harris window*. Harris used a gradient search technique to find three- and four-term expansions of Eq. (124.29) that either minimized the maximum side-lobe level for fixed main-lobe width or traded main-lobe width for minimum side-lobe level (see [Table 124.1](#)).

6. *Centered Gaussian window*

$$w(n) = \exp \left[ -\frac{1}{2} \alpha \left( \frac{n}{N/2} \right)^2 \right],$$

$$0 \leq |n| \leq \frac{N}{2}, \quad \alpha = 2, 3, \dots \quad (124.30)$$

As  $\alpha$  increases, the main lobe of the frequency spectrum becomes broader and the side-lobe peaks become lower.

7. *Centered Kaiser-Bessel window*

$$w(n) = I_0[\pi\alpha \sqrt{1.0 - \left( \frac{n}{N/2} \right)^2}] / I_0[\pi\alpha],$$

$$0 \leq |n| \leq \frac{N}{2} \quad (124.31)$$

where

$I_0(x)$  = zero-order modified Bessel function

$$= \sum_{k=0}^{\infty} \left( \frac{(x/2)^k}{k!} \right)^2$$

$$k! = 1 \times 2 \times 3 \times \dots \times k$$

$$\alpha = 2, 2.5, 3 \quad (\text{typical values}) \quad (124.32)$$

**Table 124.1** Blackman-Harris window parameters

	Parameter Values			
Number of terms in Eq. (124.29)	3	3	4	4
Minimum side lobe (dB)	-70.83	-62.05	-92	-74.39
Parameter				
$a_0$	0.42323	0.44959	0.35875	0.40217
$a_1$	0.49755	0.49364	0.48829	0.49703
$a_2$	0.07922	0.05677	0.14128	0.09892
$a_3$	—	—	0.01168	0.00188

## Fast Fourier Transform

One of the approaches to speed the computation of the DFT of a sequence is the *decimation-in-time method*. This approach involves breaking the  $N$ -point transform into two  $(N/2)$ -point transforms, then breaking each  $(N/2)$ -point transform into two  $(N/4)$ -point transforms, and continuing the above process until the two-point transform is obtained. We start with the DFT expression and factor it into two DFTs of length  $N/2$ :

$$\begin{aligned}
 F(k) &= \sum_{n=0}^{N-2} f(n)W_N^{kn} & n \text{ even} \\
 &+ \sum_{n=1}^{N-1} f(n)W_N^{kn} & n \text{ odd}
 \end{aligned} \quad (124.33)$$

Letting  $n = 2m$  in the first sum and  $n = 2m + 1$  in the second, Eq. (124.33) becomes

$$F(k) = \sum_{m=0}^{(N/2)-1} f(2m)W_N^{2mk} + \sum_{m=0}^{(N/2)-1} f(2m+1)W_N^{(2m+1)k} \quad (124.34)$$

However, because of the following identities,

$$W_N^{2mk} = (W_N^2)^{mk} = e^{-j\frac{2\pi}{N}2mk} = e^{-j\frac{2\pi}{N/2}mk} = W_{N/2}^{mk} \quad (124.35)$$

and the substitution  $f(2m) = f_1(m)$  and  $f(2m+1) = f_2(m)$ ,  $m = 0, 1, \dots, N/2 - 1$ , Eq. (124.33) takes the form

$$\begin{aligned}
F(k) &= \sum_{m=0}^{(N/2)-1} f_1(m) W_{N/2}^{mk} \\
(N/2) &= \text{point DFT of even indexed sequence} \\
&+ W_N^k \sum_{m=0}^{(N/2)-1} f_2(m) W_{N/2}^{mk} \\
(N/2) &= \text{point DFT of odd indexed sequence} \\
k &= 0, \dots, N/2 - 1
\end{aligned} \tag{124.36}$$

We can also write Eq. (124.36) in the form

$$\begin{aligned}
F(k) &= F_1(k) + W_N^k F_2(k), \\
k &= 0, 1, \dots, N/2 - 1 \\
F\left(k + \frac{N}{2}\right) &= F_1(k) + W_N^{k+N/2} F_2(k) \\
&= F_1(k) - W_N^k F_2(k), \\
k &= 0, 1, \dots, N/2 - 1
\end{aligned} \tag{124.37}$$

where  $W_N^{k+N/2} = -W_N^k$  and  $W_{N/2}^{m(k+N/2)} = W_{N/2}^{mk}$ . Since the DFT is periodic,  $F_1(k) = F_1(k + N/2)$  and  $F_2(k) = F_2(k + N/2)$ .

Next we apply the same procedure to each  $N/2$  sample, where  $f_{11}(m) = f_1(2m)$  and  $f_{21}(m) = f_2(2m + 1)$ ,  $m = 0, 1, \dots, (N/4) - 1$ . Hence,

$$\begin{aligned}
F_1(k) &= \sum_{m=0}^{(N/4)-1} f_{11}(m) W_{N/4}^{mk} + W_N^{2k} \sum_{m=0}^{(N/4)-1} f_{21}(m) W_{N/4}^{mk}, \\
k &= 0, 1, \dots, \frac{N}{4} - 1
\end{aligned} \tag{124.38}$$

or



$$\begin{aligned}
F_1(k) &= F_{11}(k) + W_N^{2k} F_{21}(k) \\
F_1\left(k + \frac{N}{4}\right) &= F_{11}(k) - W_N^{2k} F_{21}(k), \\
k &= 0, 1, \dots, \frac{N}{4} - 1 \quad (124.39)
\end{aligned}$$

Therefore, each of the sequences  $f_1$  and  $f_2$  have been split into two DFTs of length  $N/4$ .

**Example 124.2.** To find the FFT of the sequence  $\{2, 3, 4, 5\}$  we first bit-reverse the elements from position  $\{00, 01, 10, 11\}$  to position  $\{00, 10, 01, 11\}$ . The new sequence is  $\{2, 4, 3, 5\}$  (see also Fig. 124.1). Using Eqs. (124.36) and (124.37) we obtain

$$\begin{aligned}
F_1(0) &= \sum_{m=0}^1 f_1(m)_2^{m \cdot 0} \\
&= f_1(0)W_2^0 + f_1(1)W_2^0 \\
&= f(0) \cdot 1 + f(2) \cdot 1
\end{aligned}$$

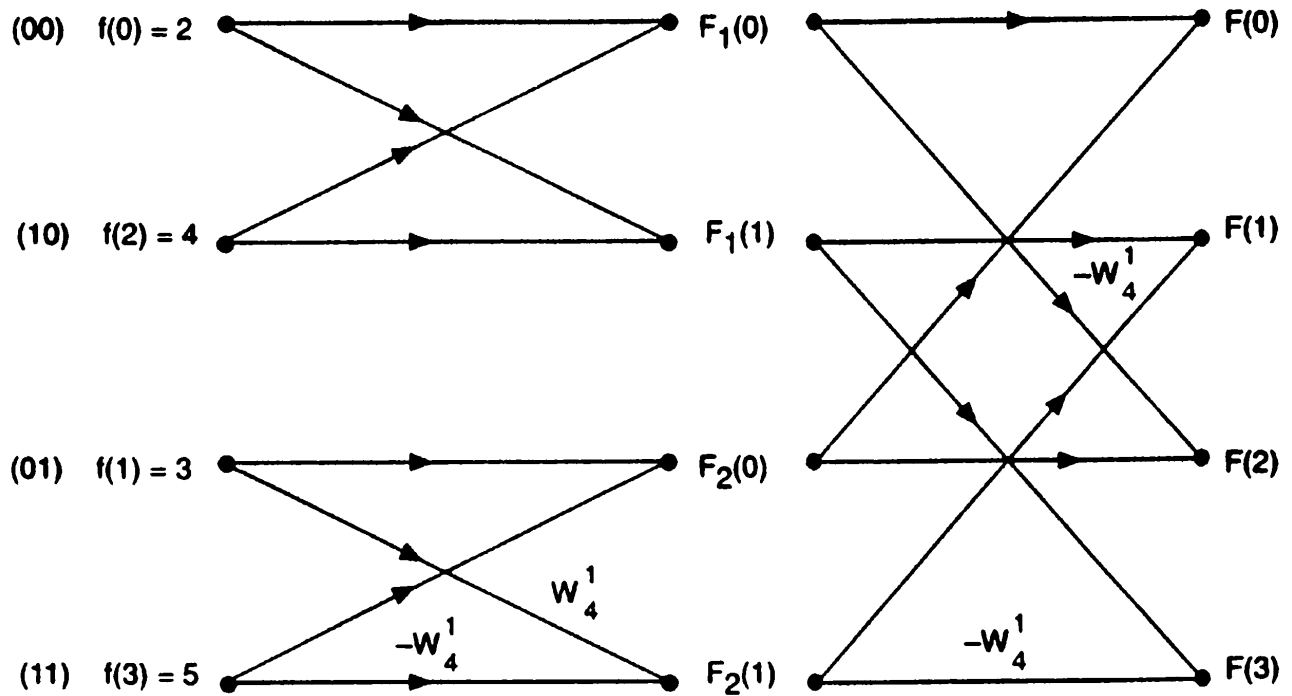
$$\begin{aligned}
F_1(1) &= \sum_{m=0}^1 f_1(m)W_2^{m \cdot 1} \\
&= f_1(0)W_2^{0 \cdot 1} + f_1(1)W_2^1 \\
&= f(0) + f(2)(-j)
\end{aligned}$$

$$\begin{aligned}
F_2(0) &= W_4^0 \sum_{m=0}^1 f_2(m)W_2^{m \cdot 0} \\
&= f_2(0)W_2^0 + f_2(1)W_2^0 \\
&= f(1) + f(3)
\end{aligned}$$

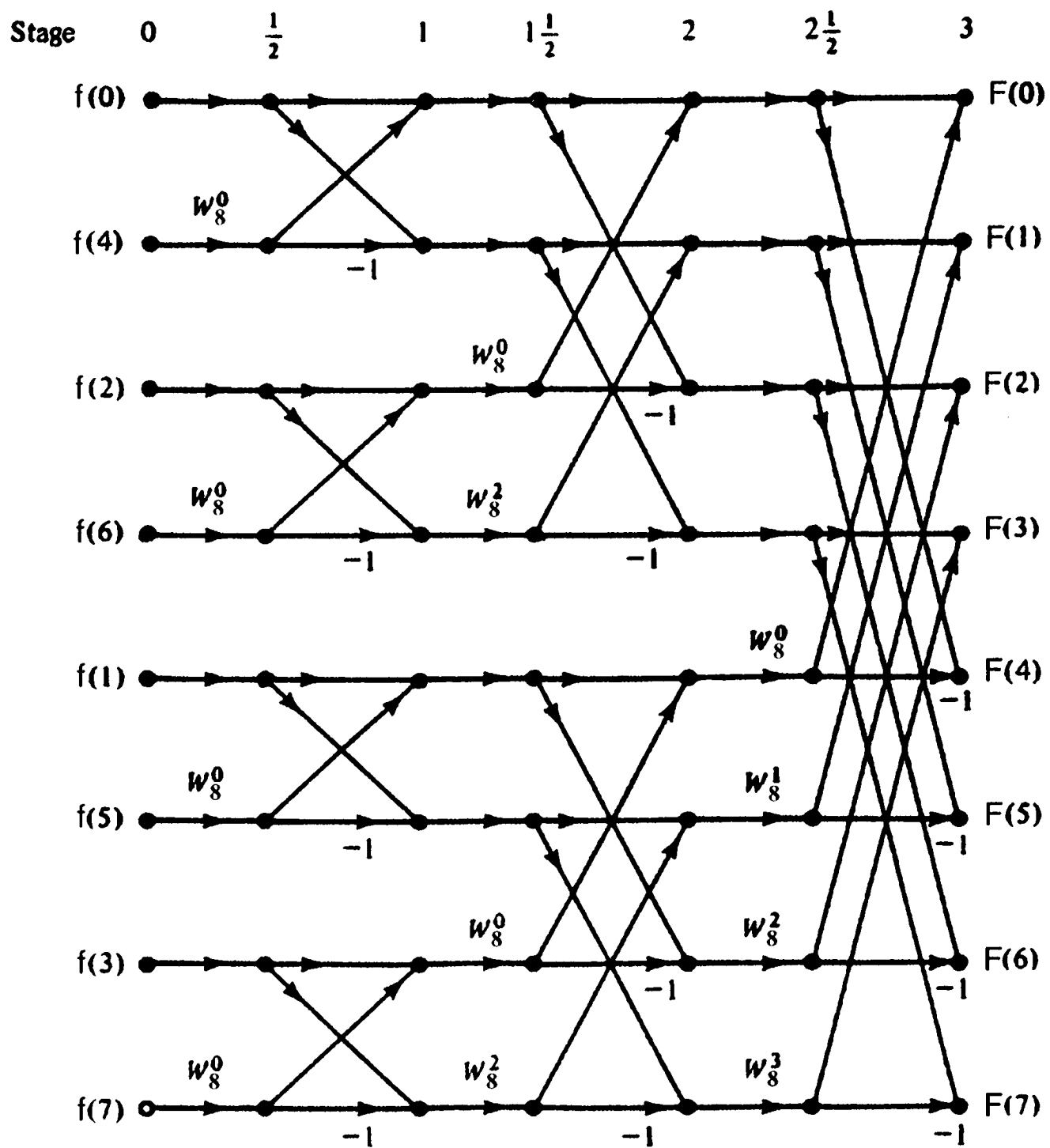
$$\begin{aligned}
F_2(1) &= W_4^1 \sum_{m=0}^1 f_2(m)W_2^{m \cdot 1} \\
&= W_4^1 [f(1)W_2^0 + f(3)W_2^1] \\
&= W_4^1 f(1) - W_4^1 f(3)
\end{aligned}$$

From Eq. (124.37) the output is  $F(0) = F_1(0) + W_4^0 F_2(0)$ ,  $F(1) = F_1(1) + W_4^1 F_2(1)$ ,  $F(2) = F_1(0) - W_4^0 F_2(0)$ , and  $F(3) = F_1(1) - W_4^1 F_2(1)$ . Figure 124.2 shows an eight-point decimation-in-time FFT. (See also Table 124.2)

**Figure 124.1** Illustration of Example 124.1. (Source: Dorf, R. C. (Ed.) 1993. *The Electrical Engineering Handbook*. CRC Press, Boca Raton, FL. With permission.)



**Figure 124.2** An eight-point decimation-in-time FFT.



**Table 124.2** Fast Fourier Transform Subroutine

```

SUBROUTINE FOUR1 (DATA, NN, ISIGN)
  Replaces DATA by its discrete Fourier transform, if ISIGN is input as 1; or replaces
  DATA by NN times its inverse discrete Fourier transform, if ISIGN is input as -1. DATA
  is a complex array of length NN or equivalently, a real array of length 2*NN. NN must
  be an integer power of 2.
  REAL*8 WR,WI,WPR,WPI,WTEMP,THETA Double precision for the trigonometric recurrences.
  DIMENSION DATA(2*NN)
  N=2*NN
  J=1
  DO 11 I=1,N,2 This is the bit-reversal section of the routine.
    IF (J.GT.I) THEN
      TEMPR=DATA(J) Exchange the two complex numbers.
      TEMPI=DATA(J+1)
      DATA(J)=DATA(I)
      DATA(J+1)=DATA(I+1)
      DATA(I)=TEMPR
      DATA(I+1)=TEMPI
    ENDIF
    M=N/2
  1 IF ((M.GE.2).AND.(J.GT.M)) THEN
      J=J-M
      M=M/2
    GO TO 1
  ENDIF
  J=J+M
  11 CONTINUE
  MMAX=2 Here begins the Danielson-Lanczos section of the routine.

  2 IF (N.GT.MMAX) THEN Outer loop executed log2 NN times.
    ISTEP=2*MMAX
    THETA=6.28318530717959DO/(ISIGN*MMAX) Initialize for trigonometric recurrence.
    WPR=-2.DO*DSIN(0.5DO*THETA)**2
    WPI=DSIN(THETA)
    WR=1.DO
    WI=0.DO
    DO 13 M=1,MMAX,2 Here are two nested loops.
      DO 12 I=M,N,ISTEP
        J=I+MMAX This is the Danielson-Lanczos formula:
        TEMPR=SNGL(WR)*DATA(J)-SNGL(WI)*DATA(J+1)
        TEMPI=SNGL(WR)*DATA(J+1)+SNGL(WI)*DATA(J)
        DATA(J)=DATA(I)-TEMPR
        DATA(J+1)=DATA(I+1)-TEMPI
        DATA(I)=DATA(I)+TEMPR
        DATA(I+1)=DATA(I+1)+TEMPI
      12 CONTINUE
      WTEMP=WR Trigonometric recurrence.
      WR=WR*WPR-WI*WPI+WR
      WI=WI*WPR+WTEMP*WPI+WI
    13 CONTINUE
    MMAX=ISTEP
  GO TO 2
ENDIF
RETURN
END

```

Source: Press, W. H., Flannery, B. P., Teukolosky, S. A., and Vetterling, W. T. 1986. *Numerical Recipes*. Cambridge University Press, Cambridge, UK. By permission.

## Computation of the Inverse DFT

To find the inverse of FFT using an FFT algorithm, we use the relation

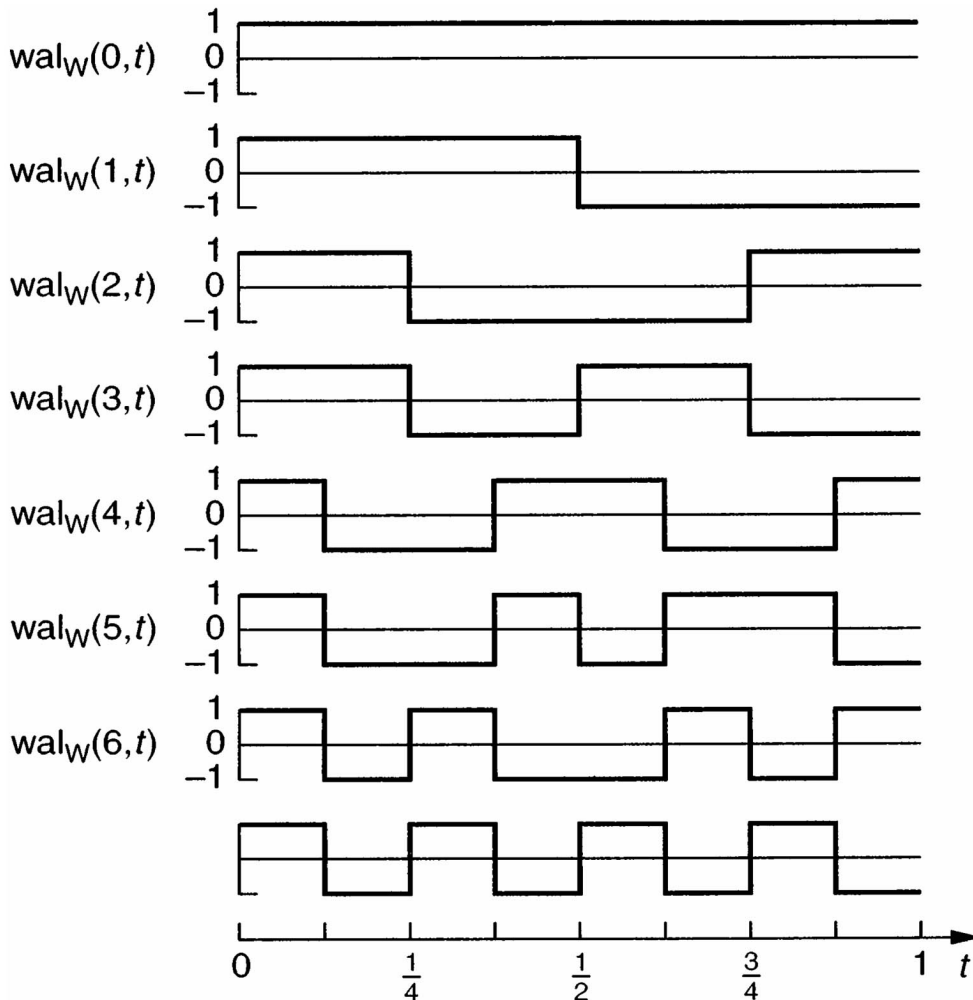
$$f(n) = [\text{FFT}(F^*(k))]^*/N \quad (124.40)$$

## 124.2 Walsh-Hadamard Transform

### Walsh Functions

In 1923 Walsh developed a complete orthogonal set of rectangular functions known as Walsh functions. The first eight functions are shown in Fig. 124.3. These signals have been widely used to perform nonsinusoidal orthogonal transforms in various digital signal processing applications since they can essentially be computed using addition and subtraction only.

**Figure 124.3** Walsh functions.



Every signal  $f(t)$  absolutely integrable in  $0 \leq t \leq 1$  can be equated in a series of the form

$$f(t) = \sum_{k=0}^{\infty} d_k \text{Wal}_w(k, t) \quad (124.41)$$

where

$$d_k = \int_0^1 f(t) \text{Wal}_w(k, t) dt, \quad k = 0, 1, 2, \dots \quad (124.42)$$

The above series converges uniformly to  $f(t)$  if it is continuous in  $0 \leq t \leq 1$  and converges in the mean where  $f(t)$  is discontinuous.

The Walsh-Hadamard transforms are analogous to the discrete Fourier transform. The basis functions are sampled Walsh functions, which can be expressed in terms of the Hadamard matrices  $\underline{H}_w(n)$ . An  $\underline{H}_w(3)$  is shown in Fig. 124.4, where  $n = \log_2 N$ .

**Figure 124.4** Hadamard matrix.

$$\underline{H}_W(3) = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \end{bmatrix}$$

Let  $u_i$  and  $v_i$  be the  $i$ th bits in the binary representation of the integers  $u$  and  $v$ , respectively; then

$$(u)_{\text{decimal}} = (u_{n-1}, u_{n-2}, \dots, u_1 u_0)_{\text{binary}}$$

$$(v)_{\text{decimal}} = (v_{n-1}, v_{n-2}, \dots, v_1 v_0)_{\text{binary}}$$

The Walsh-ordering Hadamard matrix elements are given by

$$h_{uv}^{(w)} = (-1)^{\sum_{i=0}^{n-2} r_i(u) v_i}, \quad u, v = 0, 1, \dots, N-1 \quad (124.43)$$

where

$$\begin{aligned}
 r_0(u) &= u_{n-1} \\
 r_1(u) &= u_{n-1} + u_{n-2} \\
 r_2(u) &= u_{n-2} + u_{n-3} \\
 &\vdots \\
 r_{n-1}(u) &= u_1 + u_0
 \end{aligned}$$

The Hadamard matrices have the following properties:

1.  $\underline{H}_w(k)$  is a symmetric matrix:

$$\underline{H}_w(k) = \underline{H}_w(k)^T \quad (124.44)$$

$T$  stands for transpose.

2.  $\underline{H}_w(k)$  are orthogonal:

$$\underline{H}_w(k) \underline{H}_w(k)^T = 2^k \underline{I}(k) \quad (124.45)$$

where  $\underline{I}(k)$  is a  $(2^k \times 2^k)$  identity matrix.

3. The inverse of  $\underline{H}_w(k)$  is proportional to itself:

$$[\underline{H}_w(k)]^{-1} = (1/2^k) \underline{H}_w(k) \quad (124.46)$$

where  $[\underline{H}_w(k)]^{-1}$  defines the inverse of  $\underline{H}_w(k)$ .

## Walsh-Ordered Walsh-Hadamard Transform (WHT<sub>w</sub>)

The WHT<sub>w</sub> of the data sequence  $\{x(n)\} = \{x(0), x(1), \dots, x(N-1)\}$  is defined by

$$\underline{X}(n) = \frac{1}{N} \underline{H}_w(n) \underline{x}(n) \quad (124.47)$$

where  $\underline{X}(n)$  is the  $k$ th WHT<sub>w</sub> coefficient and

$$\underline{X}(n)^T = [X(0), X(1), \dots, X(N-1)] \quad (124.48)$$

Since  $\underline{H}_w(n)$  is orthogonal and symmetric we find that the inverse Walsh-ordered Hadamard transform IWHT<sub>w</sub> is given by

$$\underline{x}(n) = \underline{H}_w(n) \underline{X}(n) \quad (124.49)$$

**Example 124.3.** Let  $\{x(n)\} = \{1, 2, 2, 1\}$ . To evaluate  $\underline{X}(n)$  for  $k = 0, 1, 2, 3$  we use Eq. (124.47). Hence,

$$\begin{bmatrix} X(0) \\ X(1) \\ X(2) \\ X(3) \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 6/4 \\ 0 \\ -2/4 \\ 0 \end{bmatrix}$$

From the above example we observe that  $N^2$  additions and/or subtractions are required to compute the  $\text{WHT}_w$  coefficients  $\underline{X}(n)$ ,  $n = 0, 1, \dots, N - 1$ .

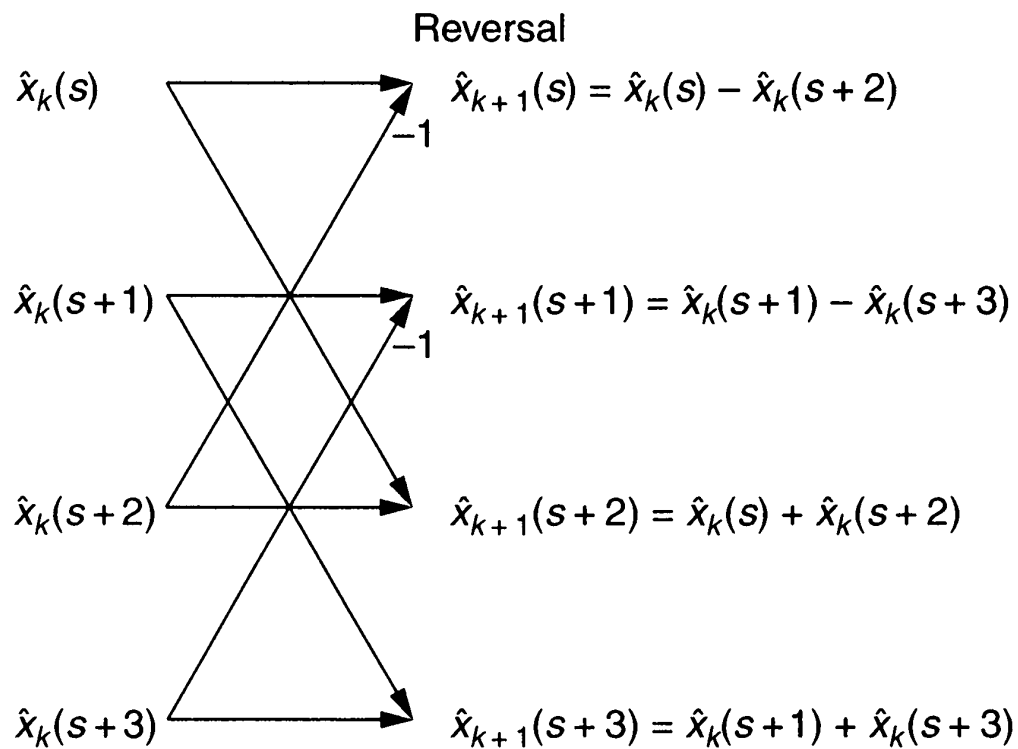
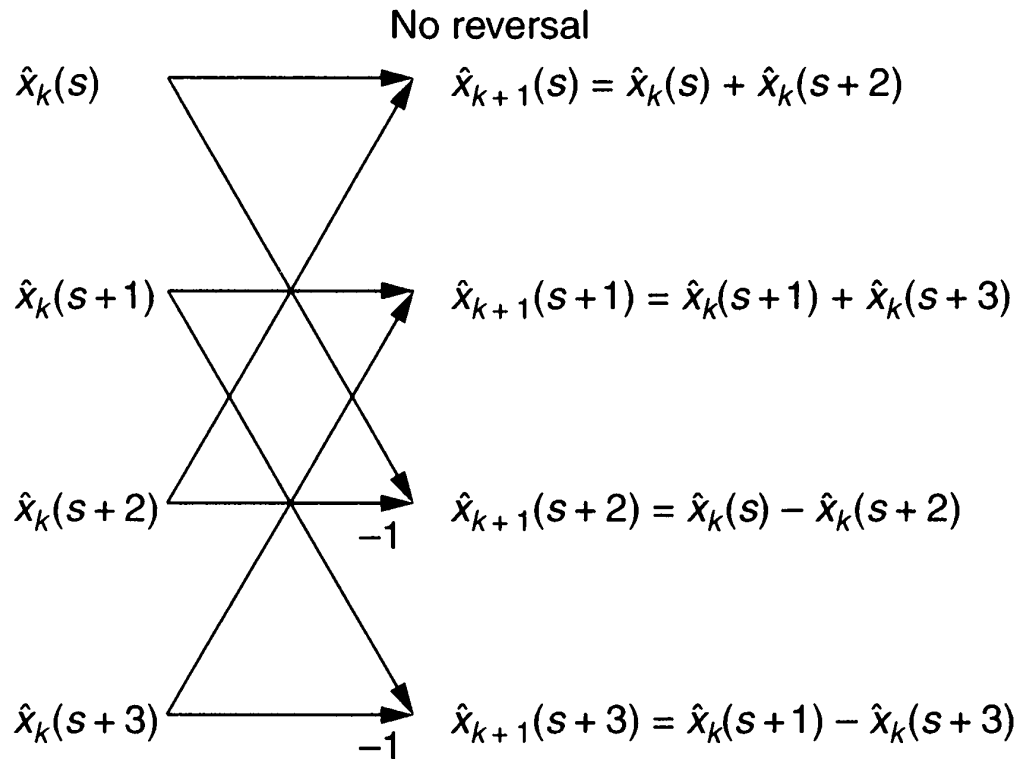
## Fast Walsh-Ordered Walsh-Hadamard Transform ( $\text{FWHT}_w$ )

Manz [1972] introduced an  $\text{FWHT}_w$  that has the following steps:

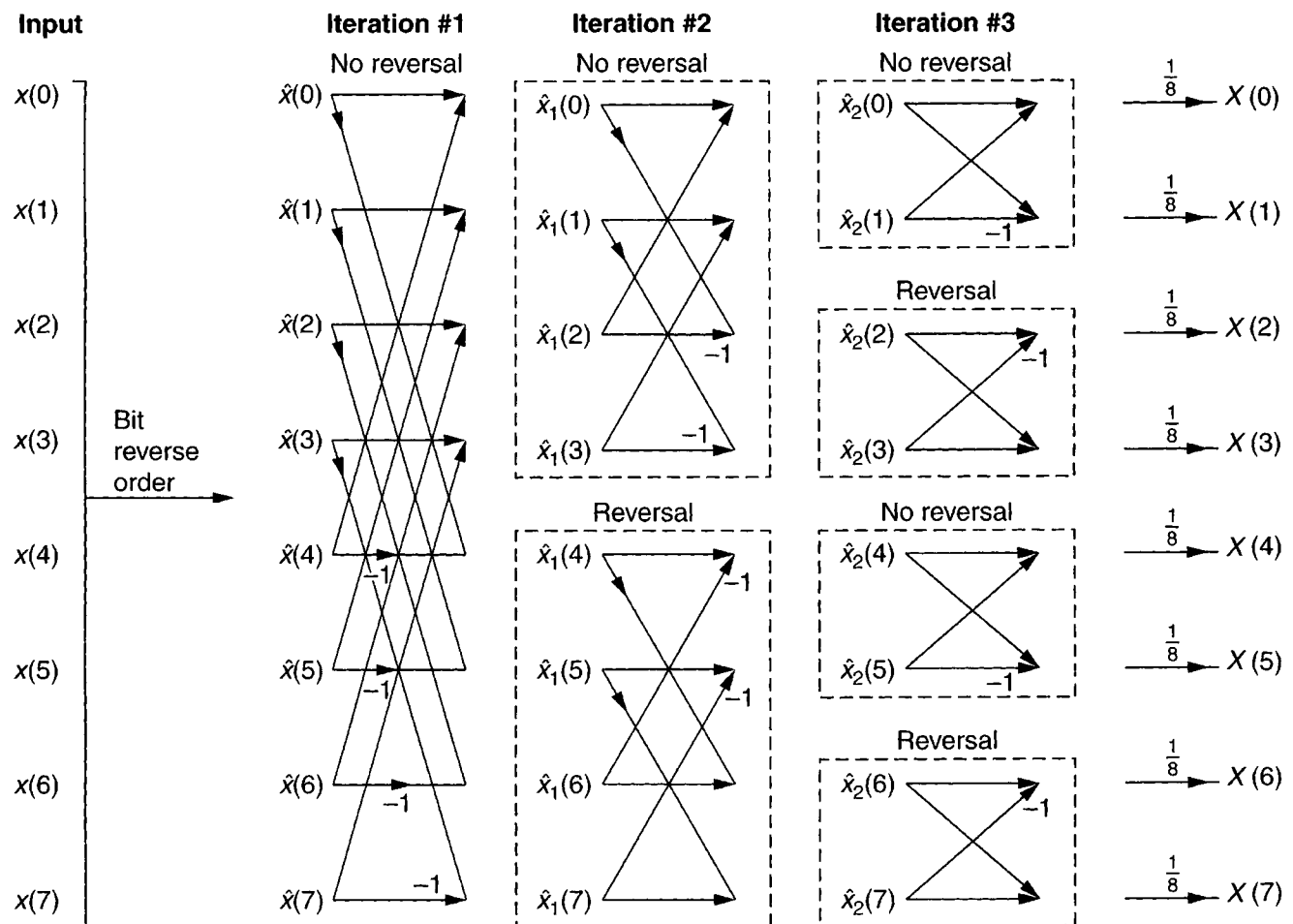
1. Bit-reverse the input sequence and order it in ascending index. For example, if  $\{x(n)\} = \{x(1), x(2), x(3)\} = \{1, 2, 3\}$ , the sequence becomes  $\{\hat{x}(n)\} = \{1, 3, 2\} = \{\hat{x}(0), \hat{x}(1), \hat{x}(2)\}$ .
2. Define a reversal, which is best illustrated by the simple case shown in Fig. 124.5.
3. Define the blocks that are reversed and nonreversed. Figure 124.6 shows the positions of blocks for an  $\text{FWHT}_w$  flow graph for  $N = 8$ .



**Figure 124.5** Reversal and nonreversal steps.



**Figure 124.6** Fast Walsh-Hadamard transform.



**Example 124.4** To find the  $\text{WHT}_w$  of the sequencing  $\{x(n)\} = \{11213213\}$  we proceed as follows:

$x(0) = 1$	$\hat{x}(0)[000] = x(0) = 1$	$x_1(0) = 1 + 1 = 2$	$x_2(0) = 1 + 1 = 2$	$x_3(0) = 1 + 1 = 2$	$\frac{1}{8}X(0) = \frac{14}{8}$
$x(1) = 1$	$\hat{x}(1)[100] = x(4) = 1$	$x_1(1) = 3 + 2 = 5$	$x_2(1) = 5 + 4 = 9$	$x_3(1) = -9 + 5 = -4$	$\frac{1}{8}X(1) = -\frac{4}{8}$
$x(2) = 2$	$\hat{x}(2)[010] = x(2) = 1$	$x_1(2) = 2 + 1 = 3$	$x_2(2) = -3 + 2 = -1$	$x_3(2) = -1 - 1 = -2$	$\frac{1}{8}X(2) = -\frac{2}{8}$
$x(3) = 1$	$\hat{x}(3)[110] = x(6) = 1$	$x_1(3) = 1 + 3 = 4$	$x_2(3) = -4 + 5 = 1$	$x_3(3) = 1 - 1 = 0$	$\frac{1}{8}X(3) = \frac{0}{8}$
$x(4) = 3$	$\hat{x}(4)[001] = x(1) = 1$	$x_1(4) = -1 + 1 = 0$	$x_2(4) = 0 - 1 = -1$	$x_3(4) = -1 + 3 = 2$	$\frac{1}{8}X(4) = \frac{2}{8}$
$x(5) = 2$	$\hat{x}(5)[101] = x(5) = 1$	$x_1(5) = -2 + 3 = 1$	$x_2(5) = 1 + 2 = 3$	$x_3(5) = -3 - 1 = -4$	$\frac{1}{8}X(5) = -\frac{4}{8}$
$x(6) = 1$	$\hat{x}(6)[011] = x(3) = 1$	$x_1(6) = -1 + 2 = 1$	$x_2(6) = 1 + 0 = 1$	$x_3(6) = 1 + 1 = 2$	$\frac{1}{8}X(6) = \frac{2}{8}$
$x(7) = 3$	$\hat{x}(7)[111] = x(7) = 1$	$x_1(7) = -3 + 1 = -2$	$x_2(7) = -2 + 1 = -1$	$x_3(7) = -1 + 1 = 0$	$\frac{1}{8}X(7) = \frac{0}{8}$

The power spectrum of  $\text{WHT}_w$  is given by

$$P_w = X^2(0)$$

$$P_w(s) = X^2(2s - 1) + X^2(2s), \quad s = 1, 2, \dots, \frac{N}{2} - 1$$

$$P_w\left(\frac{N}{2}\right) = X^2(N - 1)$$

## Defining Terms

**Fast Fourier transform (FFT):** A computational technique that reduced the number of mathematical operations in the evaluation of the discrete Fourier transform (DFT) to  $N \log_2(N)$ .

**Phase spectrum:** All phases associated with the spectrum harmonics constitute the phase spectrum.

**Power spectrum:** A power contributed by each harmonic of the spectrum.

**Window:** Any appropriate function that multiplies the data with the intent of minimizing the distortion of the Fourier spectra.

## References

- Ahmed, A. and Rao, K. R. 1975. *Orthogonal Transforms for Digital Signal Processing*. Springer-Verlag, New York.
- Blahut, E. R. 1987. *Fast Algorithms for Digital Signal Processing*. Addison Wesley, Reading, MA.
- Brigham, E. O. 1974. *The Fast Fourier Transform*. Prentice Hall, Englewood Cliffs, NJ.
- Dorf, R. C. (Ed.) 1993. *The Electrical Engineering Handbook*. CRC Press, Boca Raton, FL.
- Elliot, F. D. 1982. *Fast Transforms, Algorithms, Analysis, Application*. Academic Press, New York.

York.

Harmuth, H. F. 1969. *Transmission of Information by Orthogonal Functions*. Springer-Verlag, New York.

Manz, J. W. 1972. A sequence-ordered fast Walsh transform. *IEEE Trans. Audio and Electroacoustics*, AU-20: 204–205.

Nussbaumer, H. J. 1982. *Fast Fourier Transform and Convolution Algorithms*. Springer-Verlag, New York.

Poularikas, A. D. and Seely, S. 1993. *Signals and Systems*, 2nd ed. Krieger, Melbourne, FL.

Press, W. H., Flannery, B. P., Teukolosky, S. A., and Vetterling, W. T. 1986. *Numerical Recipes*. Cambridge University Press, Cambridge, UK.

## Further Information

A historical overview of the fast Fourier transform can be found in the following article: Cooley, J. W., Lewis, P. A. W., and Welch, P. D. 1967. Historical notes on the fast Fourier transform. *IEEE Trans. Audio and Electroacoustics*. AU-15: 76–79.

A source of fast algorithms appears frequently in the monthly magazine *Signal Processing*, which is published by the Institute of Electrical and Electronics Engineering, Inc. For subscriptions or ordering, contact: IEEE Service Center, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1931.

Bomar, B. W., Smith, L. M. "Digital Filters"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

### 125.1 Finite Impulse Response Filter Design

FIR Filter Design by Windowing • Design of Optimal FIR Filters

### 125.2 Infinite Impulse Response Filter Design

Notch Filters

### 125.3 Digital Filter Implementation

#### **Bruce W. Bomar**

*University of Tennessee Space Institute*

#### **L. Montgomery Smith**

*University of Tennessee Space Institute*

Digital filtering is concerned with the manipulation of **discrete data sequences** to remove noise or extract certain desired information. Discrete or digital signals are somewhat intuitive to human beings despite their rarity in nature. That is, people tend to make and tabulate measurements of physical entities in sets of data values with limited precision, while the entities being measured vary continuously both in their functional dependence and their value.

Although an infinite number of numerical manipulations can be applied to discrete data (e.g., finding the mean value, forming a histogram), the primary objective of digital filtering is to form a discrete output sequence  $y(n)$  from a discrete input sequence  $x(n)$ . In some manner or another, each output data sample is computed from the input data sequence—not just from any one sample, but from many, in fact, possibly from all the input samples.

The reader may already be familiar with a few processing schemes. For example, computing a moving average is sometimes used as a means of reducing measurement uncertainty or noise in data. In a three-point symmetric scheme the output sequence is found by

$$y(n) = \frac{x(n-1) + x(n) + x(n+1)}{3}$$

Note that the output sequence in this case is a sum of products of the input sample values with certain coefficients. (In this case the coefficients all have the same value of  $1/3$ .)

To make some headway in analyzing digital filters, it is necessary to restrict in two ways the types of filters being considered. First, a given filter is restricted to being linear. This means that the response of the filter to a sum of inputs is the sum of the outputs corresponding to the inputs individually and that, if the input to the filter is scaled by a multiplicative factor, the output is

scaled by the same factor.

The second restriction placed on the types of filters being analyzed is that they be shift-invariant. This property states that for any given input  $x(n)$  producing output  $y(n)$ , the response of the filter to the shifted input  $x(n - n_o)$  produces the same output sequence of values shifted by the same amount, namely,  $y(n - n_o)$ . Note that the moving average processing scheme satisfies these restrictions. However, a much larger class of processing algorithms also falls into the category of linear shift-invariant filters.

The output of a digital filter to any input can be determined from a formula known as the *convolution sum*. To derive this relation, consider first the unit impulse sequence defined as

$$\delta(n) = \begin{cases} 1 & \text{for } n = 0 \\ 0 & \text{otherwise} \end{cases}$$

By use of this special sequence any discrete sequence can be written

$$x(n) = \sum_{m=-\infty}^{\infty} x(m)\delta(n - m)$$

Suppose this sequence were input to a linear shift-invariant digital filter. Since the filter is linear, the output will be a linear combination of the outputs resulting from shifted unit impulse inputs. Since the filter is shift-invariant, the outputs will also be shifted by the same amounts. Thus, if the output of the filter in response to a single unit impulse at  $n = 0$  is the sequence  $h(n)$ , the output of the filter to the input  $x(n)$  will be the linear combination of shifted  $h(n)$  sequences weighted with the coefficients of the  $x(n)$  sequence:

$$y(n) = \sum_{m=-\infty}^{\infty} x(m)h(n - m)$$

By a change of dummy index of summation, it is easily shown that this can also be written

$$y(n) = \sum_{m=-\infty}^{\infty} h(m)x(n - m)$$

These equations are known as the **convolution summation** and describe the output of the filter to any arbitrary input. Thus, in theory, the response of a given filter to any input can be computed from knowledge of the sequence  $h(n)$ , which is referred to as the *impulse response* of the filter. It is a sequence that is of extreme importance in digital filtering, since it completely describes the filter.

A digital filter is often conveniently described in terms of its frequency characteristics, which are given by the Fourier transform of its impulse response. The impulse response and frequency response make up the Fourier transform pair,

$$H(e^{j\omega}) = \sum_{n=-\infty}^{\infty} h(n)e^{-j\omega n}, \quad -\pi \leq \omega \leq \pi$$

$$h(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H(e^{j\omega})e^{j\omega n} d\omega, \quad -\infty \leq n \leq \infty$$

where  $\omega$  is the digital radian frequency. The discrete-time Fourier transform is expressed as a function of  $e^{j\omega}$  rather than simply as a function of  $\omega$  to distinguish the discrete-time transform from its continuous-time counterpart. The radian frequency ranging from  $\omega = 0$  to  $\omega = \pi$  corresponds to the "real-world" frequency range from 0 to  $F_s/2$ , where  $F_s$  is the sample rate of the discrete-time data. For example, if the sample rate of data into a digital filter is 2 kHz, then the frequency 100 Hz corresponds to  $\omega = 0.1\pi$ .

Just as in the analysis of continuous-time systems, the operation of convolution is equivalent to the product of Fourier transforms in the frequency domain,  $Y(e^{j\omega}) = H(e^{j\omega})X(e^{j\omega})$ . Therefore,  $H(e^{j\omega})$  may be treated as the *transfer function* of the digital filter since it relates the input Fourier transform to the Fourier transform of the output, specifying how each frequency component in the filter input is altered by the filter.

Closely related to the Fourier transform of  $h(n)$  is the  $z$  transform defined by

$$H(z) = \sum_{n=-\infty}^{\infty} h(n)z^{-n}$$

$H(z)$  is referred to as the  $z$ -domain transfer function of the filter. The Fourier transform is then the  $z$  transform evaluated on the unit circle in the  $z$  plane ( $z = e^{j\omega}$ ). An important property of the  $z$  transform is that  $z^{-1}H(z)$  corresponds to  $h(n-1)$ , so  $z^{-1}$  represents a one-sample delay, termed a *unit delay*.

In this chapter the focus will be restricted to **filter design** and **filter implementation** of frequency-selective filters. These filters are intended to pass frequency components of the input sequence in a given band of the spectrum while blocking the rest. Typical frequency-selective filter types are low-pass, high-pass, band-pass, and band-reject filters. Other special-purpose filters exist; however, their design is an advanced topic that will not be addressed here. In addition, special attention is given to causal filters for which the impulse response is identically zero for negative  $n$  and which can thus be implemented in real time.

## 125.1 Finite Impulse Response Filter Design

---

The objective of **finite impulse response (FIR) digital filter** design is to determine  $N + 1$  coefficients

$$h(0), h(1), \dots, h(N)$$



so that the transfer function  $H(e^{j\omega})$  approximates a desired frequency characteristic  $H_d(e^{j\omega})$ . All other impulse response coefficients are zero. An important property of FIR filters for practical applications is that they can be designed to be *linear phase*; that is, the transfer function has the form

$$H(e^{j\omega}) = A(e^{j\omega})e^{-j\omega N/2}$$

where the amplitude  $A(e^{j\omega})$  is a real function of frequency. The desired transfer function can be similarly written

$$H_d(e^{j\omega}) = A_d(e^{j\omega})e^{-j\omega N/2}$$

where  $A_d(e^{j\omega})$  describes the amplitude of the desired frequency-selective characteristics. For example, the amplitude frequency characteristics of an ideal low-pass filter are given by

$$A_d(e^{j\omega}) = \begin{cases} 1 & \text{for } |\omega| \leq \omega_c \\ 0 & \text{otherwise} \end{cases}$$

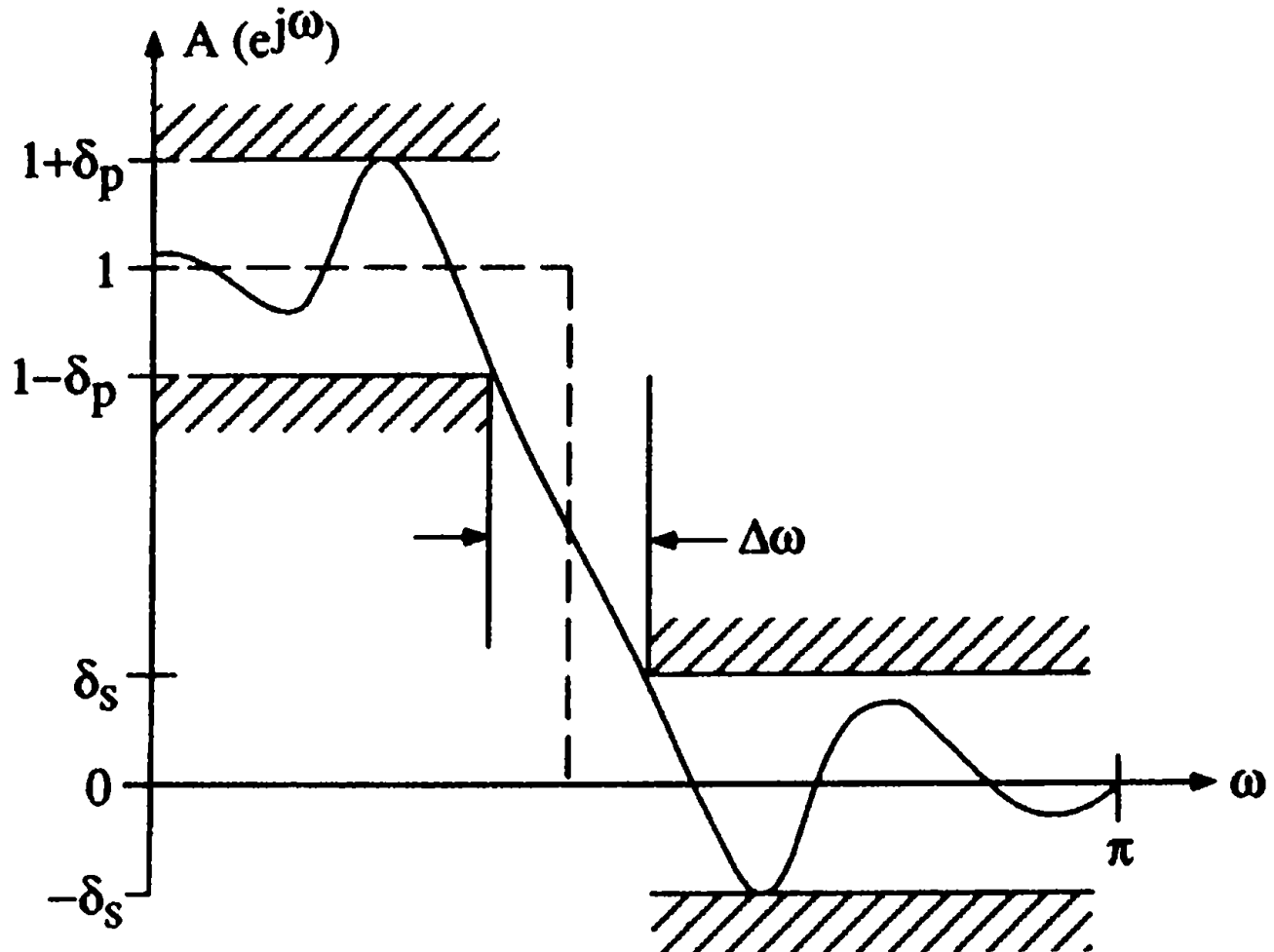
where  $\omega_c$  is the *cutoff frequency* of the filter.

A linear phase characteristic ensures that a filter has a constant delay independent of frequency. Thus, all frequency components in the signal are delayed by the same amount, and the only signal distortion introduced is that imposed by the filter's frequency-selective characteristics.

Since an FIR filter can only approximate a desired frequency-selective characteristic, some measures of the accuracy of approximation are needed to describe the quality of the design. These are the *passband ripple*  $\delta_p$ , the *stopband attenuation*  $\delta_s$ , and the *transition bandwidth*  $\Delta\omega$ . These quantities are illustrated in Fig. 125.1 for a prototype low-pass filter. The passband ripple gives the maximum deviation from the desired amplitude (typically unity) in the region where the input signal spectral components are desired to be passed unattenuated. The stopband attenuation gives the maximum deviation from zero in the region where the input signal spectral components are desired to be blocked. The transition bandwidth gives the width of the spectral region in which the frequency characteristics of the transfer function change from the passband to the stopband values. Passband ripple and stopband attenuation are often specified in decibels, in which case their values are related to the quantities  $\delta_p$  and  $\delta_s$  by the following:

$$\begin{aligned} \text{Passband ripple in dB} &= P = -20 \log_{10} (1 - \delta_p) \\ \text{Stopband attenuation in dB} &= S = -20 \log_{10} \delta_s \end{aligned}$$

**Figure 125.1** Amplitude frequency characteristics of an FIR low-pass filter showing definitions of passband ripple  $\delta_p$ , stopband attenuation  $\delta_s$ , and transition bandwidth  $\Delta\omega$ . (Source: Dorf, R. C. (Ed.) 1993. *The Electrical Engineering Handbook*, p. 240. CRC Press, Boca Raton, FL. With permission.)



## FIR Filter Design by Windowing

The windowing design method is a computationally efficient technique for producing nonoptimal filters. Filters designed in this manner have equal passband ripple and stopband attenuation:

$$\delta_p = \delta_s = \delta$$

The method begins by finding the impulse response of the desired filter from

$$h_d(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} A_d(e^{j\omega}) e^{j\omega(n-N/2)} d\omega$$

For ideal low-pass, high-pass, band-pass, and band-reject frequency-selective filters, the integral can be solved in closed form. The impulse response of the filter is then found by multiplying this ideal impulse response with a window  $w(n)$  that is identically zero for  $n < 0$  and for  $n > N$ :

$$h(n) = h_d(n)w(n), \quad n = 0, 1, \dots, N$$

Some commonly used windows are defined as follows:

Rectangular (truncation)

$$w(n) = \begin{cases} 1 & \text{for } 0 \leq n \leq N \\ 0 & \text{otherwise} \end{cases}$$

Hamming

$$w(n) = \begin{cases} 0.54 - 0.46 \cos(2\pi n/N) & \text{for } 0 \leq n \leq N \\ 0 & \text{otherwise} \end{cases}$$

Kaiser

$$w(n) = \begin{cases} I_0 \left( \beta \sqrt{1 - [(2n - N)/N]^2} \right) / I_0(\beta) & \text{for } 0 \leq n \leq N \\ 0 & \text{otherwise} \end{cases}$$

In general, windows that slowly taper the impulse response to zero result in lower passband ripple and a wider transition bandwidth. Other windows (e.g., Hanning, Blackman) are also sometimes used, but not as often as those shown above.

Of particular note is the Kaiser window where  $I_0(\cdot)$  is the 0th-order modified Bessel function of the first kind, and  $\beta$  is a shape parameter. The proper choice of  $N$  and  $\beta$  allows the designer to meet given passband ripple, stopband attenuation, and transition bandwidth specifications. Specifically, using  $S$ , the stopband attenuation in dB, the filter order must satisfy

$$N = \frac{S - 8}{2.285 \Delta\omega}$$

Then the required value of the shape parameter is given by

$$\beta = \begin{cases} 0 & \text{for } S < 21 \\ 0.5842(S - 21)^{0.4} + 0.07886(S - 21) & \text{for } 21 \leq S \leq 50 \\ 0.1102(S - 8.7) & \text{for } S > 50 \end{cases}$$

Consider as an example of this design technique a low-pass filter with a cutoff frequency of  $\omega_c = 0.4\pi$ . The ideal impulse response for this filter is given by

$$h_d(n) = \frac{\sin[0.4\pi(n - N/2)]}{\pi(n - N/2)}$$

Choosing  $N = 8$  and a Kaiser window with shape parameter of  $\beta = 0.5$  yields the following impulse response coefficients:

$$\begin{aligned}
h(0) &= h(8) = -0.075\,682\,67 \\
h(1) &= h(7) = -0.062\,365\,96 \\
h(2) &= h(6) = 0.093\,548\,92 \\
h(3) &= h(5) = 0.302\,730\,70 \\
h(4) &= 0.400\,000\,00
\end{aligned}$$

## Design of Optimal FIR Filters

The accepted standard criterion for the design of optimal FIR filters is to minimize the maximum value of the error function

$$E(e^{j\omega}) = W_d(e^{j\omega})|A_d(e^{j\omega}) - A(e^{j\omega})|$$

over the full range of  $-\pi \leq \omega \leq \pi$ .  $W_d(e^{j\omega})$  is a desired weighting function used to emphasize specifications in a given frequency band. The ratio of the deviation in any two bands is inversely proportional to the ratio of their respective weighting.

A consequence of this optimization criterion is that the frequency characteristics of optimal filters are *equiripple*: Although the maximum deviation from the desired characteristic is minimized, it is reached several times in each band. Thus, the passband and stopband deviations oscillate about the desired values with equal amplitude in each band. Such approximations are frequently referred to as *mini-max* or *Chebyshev* approximations. In contrast, the maximum deviations occur near the band edges for filters designed by windowing.

Equiripple FIR filters are usually designed using the *Parks-McClellan* computer program [Antoniou, 1979], which uses the *Remez exchange algorithm* to determine iteratively the *extremal frequencies* at which the maximum deviations in the error function occur. A listing of this program along with a detailed description of its use is available in several references, including Parks and Burrus [1987] and DSP Committee [1979]. The program is executed by specifying as inputs the desired band edges, gain for each band (usually 0 or 1), band weighting, and filter impulse response length. If the resulting filter has too much ripple in some bands, those bands can be weighted more heavily and the filter redesigned. Details on this design procedure are discussed in Rabiner [1973] along with approximate design relationships that aid in selecting the filter length needed to meet a given set of specifications.

Although we have focused on the design of frequency-selective filters, other types of FIR filters exist. For example, the Parks-McClellan program will also design linear-phase FIR filters for differentiating broadband signals and for approximating the Hilbert transform of such signals.

For practice in the principles of equiripple filter design, consider an eighth-order low-pass filter with a passband  $0 \leq \omega \leq 0.3\pi$ , a stopband  $0.5\pi \leq \omega \leq \pi$ , and equal weighting for each band. The impulse response coefficients generated by the Parks-McClellan program are as follows:

$$\begin{aligned}
h(0) &= h(8) = -0.063\,678\,59 \\
h(1) &= h(7) = -0.069\,122\,76 \\
h(2) &= h(6) = 0.101\,043\,60 \\
h(3) &= h(5) = 0.285\,749\,90 \\
h(4) &= 0.410\,730\,00
\end{aligned}$$

These values can be compared to those for the similarly specified filter designed in the previous section using the windowing method.

## 125.2 Infinite Impulse Response Filter Design

---

An **infinite impulse response (IIR) digital filter** requires less computation to implement than an FIR digital filter with a corresponding frequency response. However, IIR filters cannot generally achieve a perfect linear phase response.

Techniques for the design of infinite impulse response analog filters are well established. For this reason the most important class of IIR digital filter design techniques is based on forcing a digital filter to behave like a reference analog filter. For frequency-selective filters this is generally done by attempting to match frequency responses. This task is complicated by the fact that the analog filter response is defined for an infinite range of frequencies ( $\Omega = 0$  to  $\infty$ ), whereas the digital filter response is defined for a finite range of frequencies ( $\omega = 0$  to  $\pi$ ). Therefore, a method for mapping the infinite range of analog frequencies,  $\Omega$ , into the finite range from  $\omega = 0$  to  $\pi$  that is termed the *bilinear transform* is employed.

Let  $H_a(s)$  be the Laplace transform transfer function of an analog filter with frequency response  $H_a(j\Omega)$  ( $\Omega$  in radians/s). The bilinear transform method obtains the digital filter transfer function  $H(z)$  from  $H_a(s)$  using the substitution

$$s = \frac{2}{T} \frac{1 - z^{-1}}{1 + z^{-1}}$$

That is,

$$H(z) = H_a(s) \Big|_{s=(2/T)(1-z^{-1})/(1+z^{-1})}$$

This transformation maps analog frequency  $\Omega$  to digital frequency  $\omega$  according to

$$\omega = 2 \tan^{-1} \left( \frac{\Omega T}{2} \right)$$

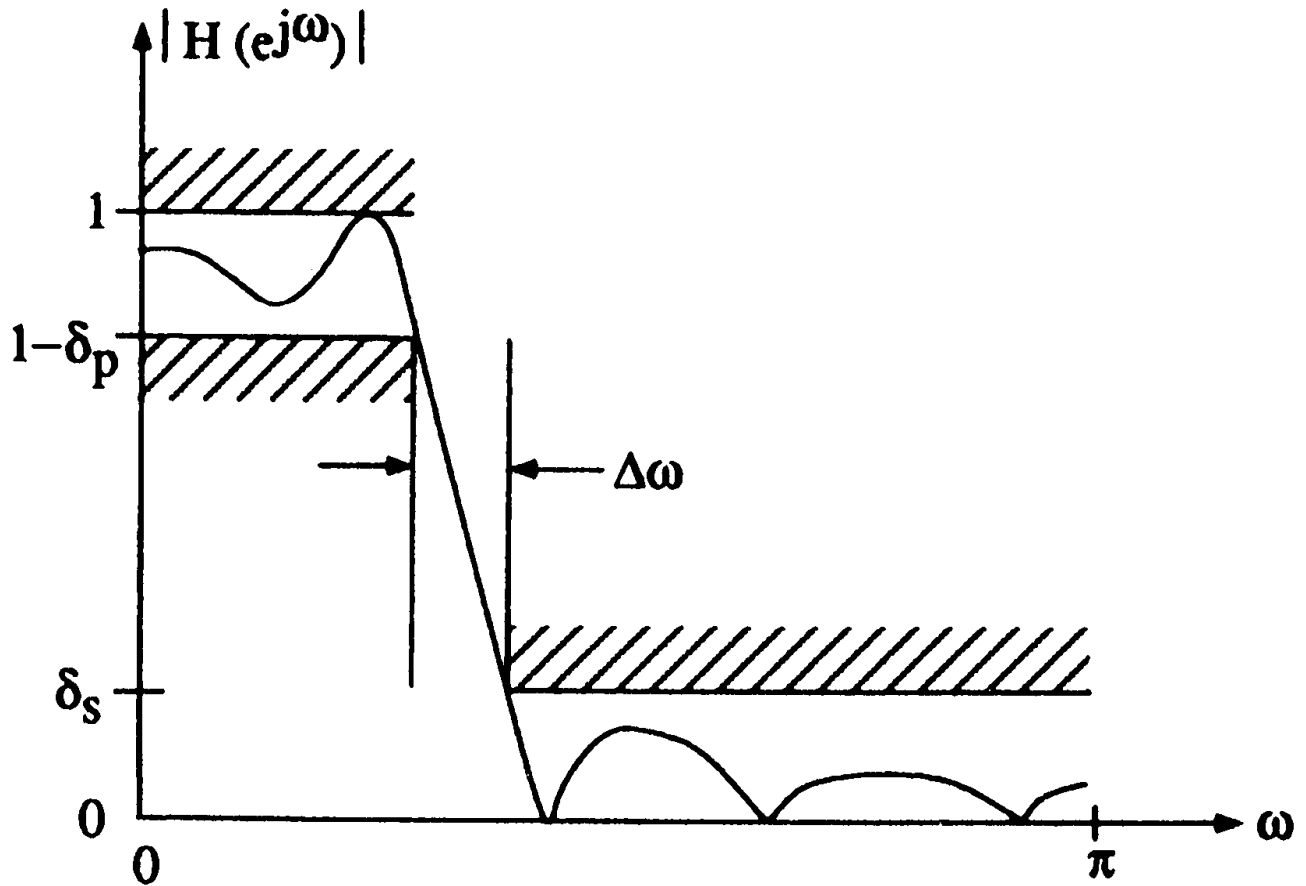
thereby warping the frequency response  $H_a(j\Omega)$  and forcing it to lie between 0 and  $\pi$  for  $H(e^{j\omega})$ . Therefore, to obtain a digital filter with a cutoff frequency of  $\omega_c$ , it is necessary to design an analog filter with cutoff frequency

$$\Omega_c = \frac{2}{T} \tan\left(\frac{\omega_c}{2}\right)$$

This process is referred to as *prewarping* the analog filter frequency response to compensate for the warping of the bilinear transform. Applying the bilinear transform substitution to this analog filter will then give a digital filter that has the desired cutoff frequency.

Analog filters and hence IIR digital filters are typically specified in a slightly different fashion than are FIR filters. Figure 125.2 illustrates how analog and IIR digital filters are usually specified. Notice by comparing to Fig. 125.1 that the passband ripple in this case never goes above unity, where in the FIR case the passband ripple is specified about unity.

**Figure 125.2** Frequency characteristics of an IIR digital low-pass filter showing definitions of passband ripple  $\delta_p$ , stopband attenuation  $\delta_s$ , and transition bandwidth  $\Delta\omega$ . (Source: Dorf, R. C. (Ed.) 1993. *The Electrical Engineering Handbook*, p. 244. CRC Press, Boca Raton, FL. With permission.)



Four basic types of analog filters are generally used to design digital filters: (1) Butterworth filters, which are maximally flat in the passband and decrease monotonically outside the passband, (2) Chebyshev filters, which are equiripple in the passband and decrease monotonically outside the passband, (3) inverse Chebyshev filters, which are flat in the passband and equiripple in the

stopband, and (4) elliptic filters, which are equiripple in both the passband and stopband. Techniques for designing these analog filters are covered elsewhere [see, for example, [Van Valkenberg \(1982\)](#)] and will not be considered here.

To illustrate the design of an IIR digital filter using the bilinear transform, consider the design of a second-order Chebyshev low-pass filter with 0.5 dB of passband ripple and a cutoff frequency of  $\omega_c = 0.4\pi$ . The sample rate of the digital filter is to be 5 Hz, giving  $T = 0.2$  seconds. To design this filter we first design an analog Chebyshev low-pass filter with a cutoff frequency of

$$\Omega_c = \frac{2}{0.2} \tan(0.2\pi) = 7.2654 \text{ rad/s}$$

This filter has a transfer function

$$H(s) = \frac{0.9441}{1 + 0.1249s + 0.01249s^2}$$

Substituting

$$s = \frac{2}{0.2} \frac{z - 1}{z + 1}$$

gives

$$H(z) = \frac{0.2665(z + 1)^2}{z^2 - 0.1406z + 0.2695}$$

Computer programs are available that accept specifications on a digital filter and carry out all steps required to design the filter, including prewarping the frequencies, designing the analog filter, and performing the bilinear transform. Two such programs are given in Parks and Burrus [1987] and Antoniou [1979].

## Notch Filters

An important special type of IIR filter is the notch filter. Such filters can remove a very narrow band of frequencies and are useful for applications like removing 60 Hz noise from data. The following second-order  $z$ -domain transfer function realizes a notch filter at frequency  $\omega_0$ :

$$H(z) = \frac{1 - 2 \cos \omega_0 z^{-1} + z^{-2}}{1 - 2r \cos \omega_0 z^{-1} + r^2 z^{-2}}$$

The parameter  $r$  is restricted to the range  $0 < r < 1$ . Substituting  $z = e^{j\omega}$ , multiplying the numerator and denominator by  $e^{j\omega}$ , and applying some trigonometric identities yields the following frequency response for the filter:

$$H(e^{j\omega}) = \frac{2 \cos \omega - 2 \cos \omega_o}{(1 + r^2) \cos \omega - 2r \cos \omega_o + j(1 - r^2) \sin \omega}$$

From this expression it can easily be seen that the response of the filter at  $\omega = \omega_o$  is exactly zero. Another property is that, for  $r \cong 1$  and  $\omega \neq \omega_o$ , the numerator and denominator are very nearly equal, and so the response is approximately unity. Overall, this is a very good approximation of an ideal notch filter as  $r \rightarrow 1$ .

## 125.3 Digital Filter Implementation

---

For FIR filters the convolution sum represents a computable process, and so filters can be implemented by directly programming the arithmetic operations

$$y(n) = h(0)x(n) + h(1)x(n-1) + \cdots + h(N)x(n-N)$$

However, for an IIR filter the convolution sum does not represent a computable process. Therefore, it is necessary to examine the general transfer function, which is given by

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\gamma_0 + \gamma_1 z^{-1} + \gamma_2 z^{-2} + \cdots + \gamma_M z^{-M}}{1 + \beta_1 z^{-1} + \beta_2 z^{-2} + \cdots + \beta_N z^{-N}}$$

where  $Y(z)$  is the  $z$  transform of the filter output,  $y(n)$ , and where  $X(z)$  is the  $z$  transform of the filter input,  $x(n)$ . The unit-delay characteristic of  $z^{-1}$  then gives the following **difference equation** for implementing the filter

$$y(n) = \gamma_0 x(n) + \gamma_1 x(n-1) + \cdots + \gamma_M x(n-M) - \beta_1 y(n-1) - \cdots - \beta_N y(n-N)$$

In calculations of  $y(0)$ , the values of  $y(-1), y(-2), \dots, y(-N)$  represent initial conditions on the filter. If the filter is started in an initially relaxed state, then these initial conditions are zero.

### Defining Terms

**Convolution summation:** A possibly infinite summation expressing the output of a digital filter in terms of the impulse response coefficients of the filter and present and past values of the input to the filter. For FIR filters this summation represents a way of implementing the filter.

**Difference equation:** An equation expressing the output of a digital filter in terms of present and past values of the filter input and past values of the filter output. For IIR filters a difference equation must be used to implement the filter since the convolution summation is infinite.

**Discrete data sequence:** A set of values constituting a signal whose values are known only at distinct sampled points.

**Filter design:** The process of determining the impulse response coefficients or coefficients of a difference equation to meet a given frequency or time response characteristic.

**Filter implementation:** The numerical method or algorithm by which the output sequence of a



digital filter is computed from the input sequence.

**Finite impulse response (FIR) digital filter:** A filter whose output in response to a unit impulse function is identically zero after a given bounded number of samples.

**Infinite impulse response (IIR) digital filter:** A filter whose output in response to a unit impulse function remains nonzero for indefinitely many samples.

## References

- Antoniou, A. 1979. *Digital Filters: Analysis and Design*. McGraw-Hill, New York.
- DSP Committee, IEEE ASSP. (Eds.) 1979. *Programs for Digital Signal Processing*. IEEE Press, New York.
- Parks, T. W. and Burrus, C. S. 1987. *Digital Filter Design*. John Wiley & Sons, New York.
- Rabiner, L. R. 1973. Approximate design relationships for low-pass FIR digital filters. *IEEE Trans. Audio Electroacoust.* AU-21:456–460.
- Van Valkenberg, M. E. 1982. *Analog Filter Design*. Holt, Rinehart & Winston, New York.

## Further Information

The monthly journals *IEEE Transactions on Circuits and Systems II* and *IEEE Transactions on Signal Processing* routinely publish articles on the design and implementation of digital filters. The bimonthly journal *IEEE Transactions on Instrumentation and Measurement* also contains related information. The use of digital filters for integration and differentiation is discussed in the December 1990 issue (pp. 923–927).

H. Vincent Poor. "Modulation and Detection"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Modulation and Detection

---

## 126.1 Analog Modulation and Detection

Linear Filtering and Fourier Transforms • Linear Modulation • Angle Modulation • Signal-to-Noise Ratio Analysis

## 126.2 Digital Modulation and Detection

On-Off Keying (OOK) • Phase-Shift Keying (PSK) • Frequency-Shift Keying (FSK) • Bit-Error Rates and Bandwidth Efficiency

## 126.3 Further Issues

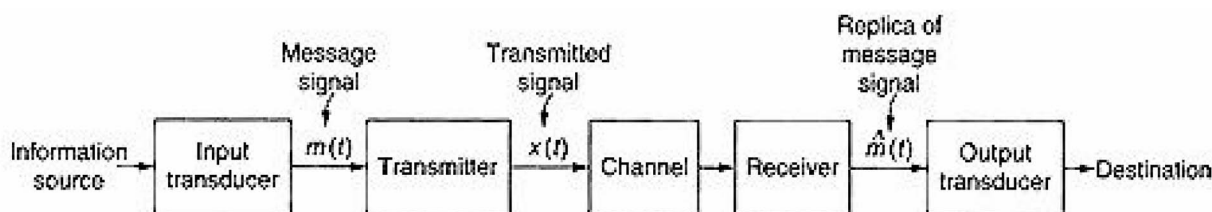
Pulse Modulation • Multiplexing and Multiple Access • Spread-Spectrum Signaling • Sequences of Symbols • Implementation Issues

### H. Vincent Poor

*Princeton University*

A general model of an electrical communication system is shown in [Fig. 126.1](#). It consists of a *transmitting* side and a *receiving* side. The transmitting side includes an *information source* that produces an *input message*, a transducer that converts this input message to an electrical **message signal**, and a **transmitter** that couples the message signal to the communication **channel**. The receiving side consists of a **receiver** that converts the signal received from the channel back into a replica of the electrical message signal, which is in turn converted to an *output message* via an output transducer. The information source can be either continuous (i.e., analog) or discrete in nature. Typical continuous information sources include speech and video, whereas discrete information sources include computer data and text. In the model of [Fig. 126.1](#) the channel comprises all of the physical media affecting the transmitted signal. Depending on the type of communication system being modeled, these media could include coaxial cables, microwave links, telephone lines, optical fibers, the electronics in the receiver, and so forth. A communication channel has essentially two characteristics that affect the ability to transmit information through it: limited **bandwidth** and noise. The bandwidth of a channel describes the maximum rate of change in the input to which the channel can respond. Obviously, this quantity limits the signaling rate that the channel can accommodate. *Noise* refers to random distortion that the channel introduces into the transmitted signal. Such noise consists of atmospheric noises arising in the physical channel connecting the transmitter and receiver, internally generated thermal noises due to random electron motion in the receiver electronics, and other effects such as amplitude fading. In addition to bandwidth limitations and noise, some channels introduce other changes, such as nonlinear distortion, into the transmitted signal.

**Figure 126.1** A general model for an electrical communication system.



We focus here on two elements of the communication system shown in Fig. 126.1—namely, the transmitter and the receiver. In most communication systems the primary purpose of the transmitter is **modulation**, and the primary purpose of the receiver is **demodulation** (or **detection**).

*Modulation* refers to the impression of the message signal onto a **carrier** signal—that is, onto a signal that is well-suited to carry the message through the channel. There are several reasons why it may be desirable to impress the message signal onto a carrier. For example, in radio channels the carrier signal lies in a frequency range that is more easily radiated than is the message signal. Also, different carriers can be used to assign different messages to different channels, or modulation can be used to place multiple messages into the same channel. A further purpose of modulation is to reduce the effects of noise and interference by creating a transmitted signal that is less susceptible to these effects than is the message signal. Carrier signals are almost always sinusoids or signals related to sinusoids, and thus modulation involves modifying the amplitude, phase, or frequency of a sinusoidal carrier in concert with the message signal. *Demodulation* or *detection* refers to the extraction of the message signal from the modulated carrier after it has passed through the channel.

The following paragraphs give an overview of the basic techniques used for modulation and demodulation in electrical communication systems. This treatment is necessarily very brief, and a number of more detailed treatments to which the reader may refer are listed at the end of the chapter.

## 126.1 Analog Modulation and Detection

Many basic electrical communication systems involve *continuous-wave* or *analog* message signals. Such signals have amplitudes that can take on values in a continuum of real numbers, and they are defined for continuous time. In this section we consider modulation techniques and corresponding detection techniques for transmitting such signals via sinusoidal carriers. Thus, we consider a message signal  $m(t)$  that we wish to impress onto a sinusoidal carrier signal

$$A_c \sin(2\pi f_c t + \phi_c) \quad (126.1)$$

to produce a transmitted signal  $x(t)$ . In order to accomplish this transmission, the message signal can be used to modulate (i.e., to change) either the amplitude,  $A_c$ , the frequency,  $f_c$ , or the phase,  $\phi_c$ , of the carrier. The first of these three possibilities is termed **linear modulation** and the second two are termed **angle modulation** (or *exponential modulation*).

Before describing these methods, we first digress briefly to provide some necessary background material.

## Linear Filtering and Fourier Transforms

Analog modulation and demodulation techniques are often most easily described in the frequency domain via the *Fourier transform*, which is a way of representing a temporal signal in terms of sinusoids. The Fourier transform  $S(f)$  of a temporal signal  $s(t)$  is defined by

$$S(f) = \int_{-\infty}^{\infty} s(t) e^{-j\omega\pi ft} dt \quad (126.2)$$

where  $j$  denotes the imaginary unit  $\sqrt{-1}$ . The value of  $S(f)$  represents the component of  $s(t)$  at frequency  $f$ . This transform can be inverted via the formula

$$s(t) = \int_{-\infty}^{\infty} S(f) e^{j2\pi ft} df \quad (126.3)$$

which displays the representation of  $s(t)$  in terms of the (continuous) superposition of the sinusoids  $\{e^{j2\pi ft}\}$ . A key property of the Fourier transform is that the Fourier transform of a *real-valued* signal is even symmetric about  $f = 0$ ; that is,  $S(f) = S(-f)$ . In general, the Fourier transform is a complex-valued function. The magnitude and phase of  $S(f)$  as functions of frequency are known as the *amplitude spectrum* and *phase spectrum*, respectively, of  $s(t)$ .

The Fourier transform is particularly useful in describing the effects of time-invariant linear filtering on signals. In particular, if a signal  $s(t)$  is passed through a time-invariant linear filter with impulse response  $h(t)$  to form the output signal

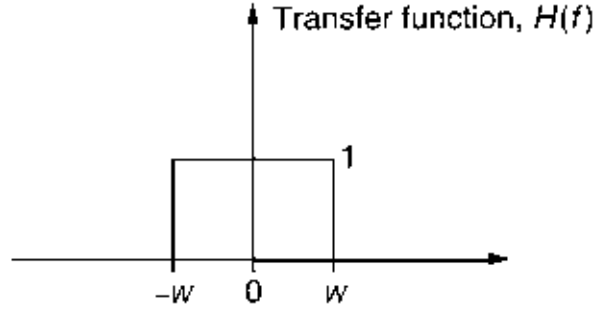
$$s_o(t) = \int_{-\infty}^{\infty} h(t - \tau) s(\tau) d\tau \quad (126.4)$$

then the Fourier transform of the output is given simply in terms of the  $S(f)$  and  $H(f)$  [the Fourier transform of  $h(t)$ ] via

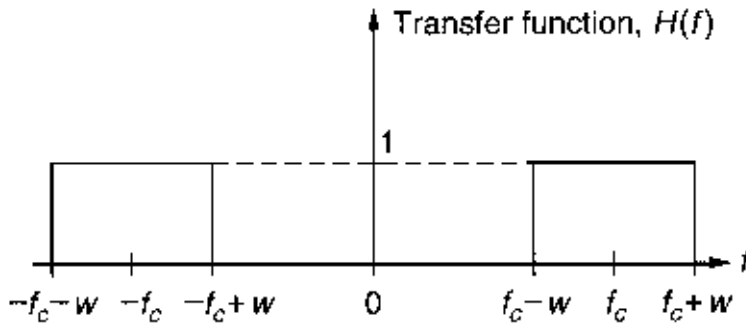
$$S_o(f) = H(f) S(f) \quad (126.5)$$

The function  $H(f)$  is known as the *transfer function* of the filter. [Figure 126.2](#) illustrates two types of filters of interest: an ideal *low-pass filter* of bandwidth  $2W$ , and an ideal *band-pass filter* of bandwidth  $2W$  and center frequency  $f_c$ . Note that these filters are idealizations of actual filters that can be used in practice. Practical filters must have a smoother transition between the *passband* [i.e., the set of frequencies for which  $|H(f)| = 1$ ] and the *stopband* [the set of frequencies for which  $|H(f)| = 0$ ].

**Figure 126.2** Transfer functions of idealized filters.



(a) Ideal lowpass filter with bandwidth  $2w$



(b) Ideal bandpass filter with bandwidth  $2w$  and center frequency  $f_c$

## Linear Modulation

There are several ways in which the amplitude of a sinusoidal carrier can be modulated by a message signal. The two most basic of these are:

*Double-sideband (DSB) modulation:*

$$x_{\text{DSB}}(t) = A_c m(t) \sin(2\pi f_c t + \phi_c) \quad (126.6)$$

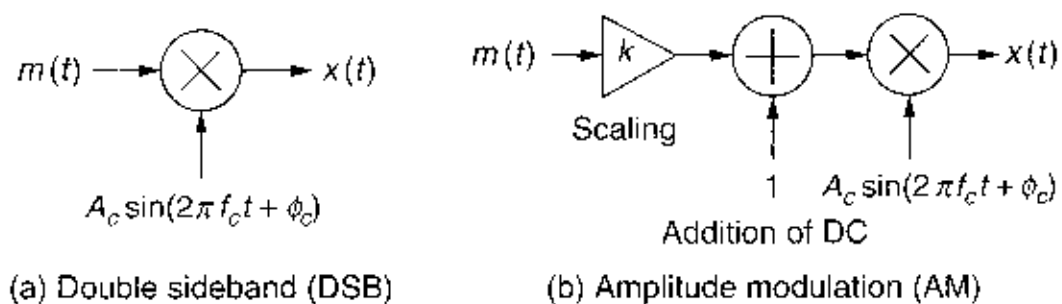
*Amplitude modulation (AM):*

$$x_{\text{AM}}(t) = A_c [1 + km(t)] \sin(2\pi f_c t + \phi_c) \quad (126.7)$$

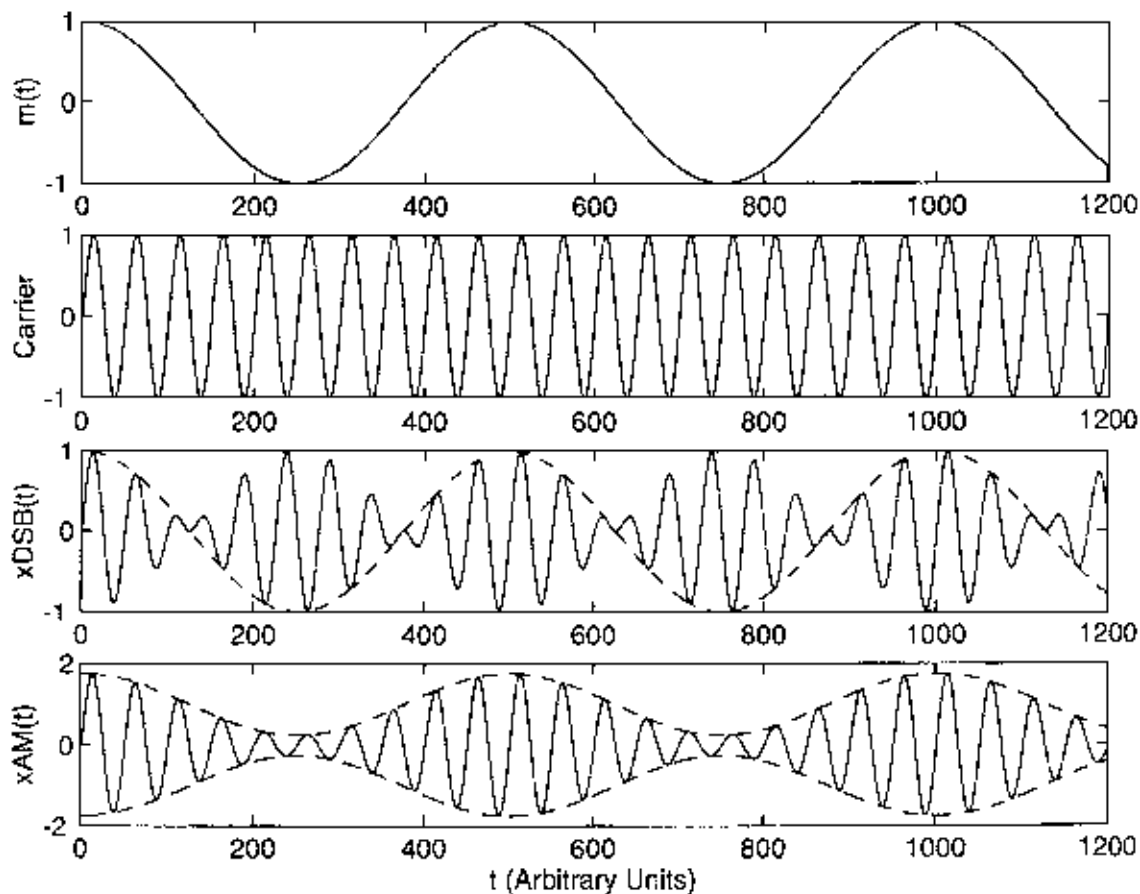
where the *modulation index*,  $k$ , satisfies  $0 < k \leq 1$  and where  $m(t)$  has been scaled to satisfy  $|m(t)| \leq 1$ .

These modulation techniques are illustrated in Fig. 126.3, and the resulting signals are illustrated in Fig. 126.4 for the case in which the message signal is also a sinusoid. (In this illustration  $k = 0.75$ .)

**Figure 126.3** Basic linear analog modulation techniques.



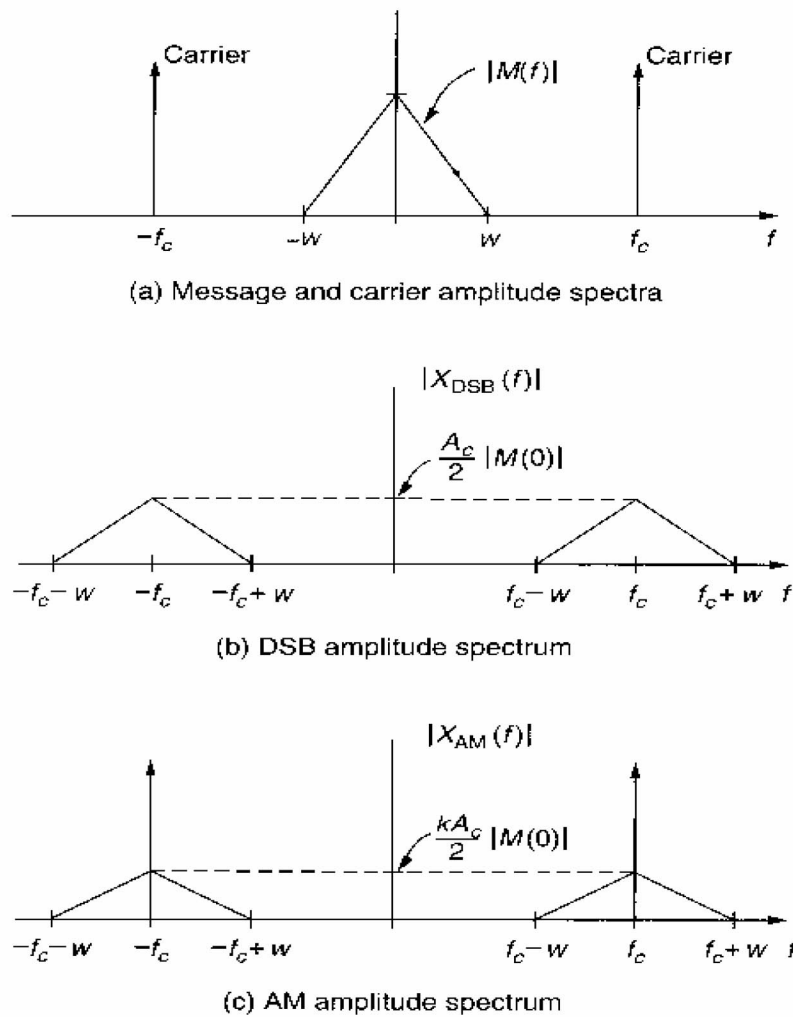
**Figure 126.4** Illustration of waveforms arising in DSB and AM.



The multiplication of the message signal by the sinusoidal carrier has the effect of translating the *baseband* power of the message signal (i.e., the power concentrated around zero frequency) up to

power centered around the carrier frequency. This is illustrated in Fig. 126.5, which shows the Fourier transforms of a baseband message signal and of DSB and AM signals created from this message signal. An impulse in the Fourier transform (represented by a vertical arrow) denotes the transform of a sinusoid. Thus, we note that the AM signal contains a residual of the carrier, as evidenced by the impulse in its transform at the carrier frequency. A basic DSB does not have this residual carrier, although a carrier is sometimes transmitted with DSB to aid in demodulation. For this reason the basic form of DSB defined in Eq. (126.6) is sometimes known as *suppressed carrier* DSB.

**Figure 126.5** Amplitude spectra in linear modulation.



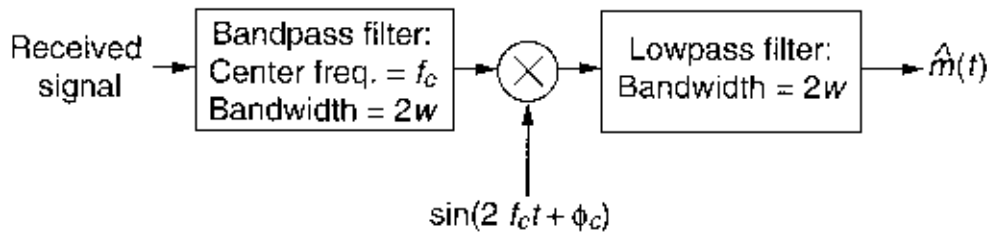
**Figure 126.5** Amplitude spectra in linear modulation.

DSB can be demodulated by multiplying the received signal by a replica of the carrier and then

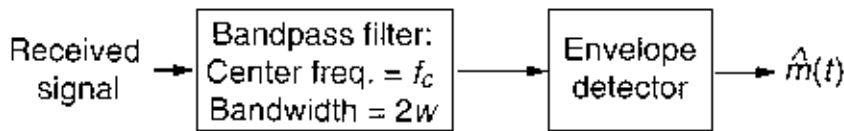


passing the result through a low-pass filter whose bandwidth includes the modulated signal. This type of demodulator—known as a *coherent* detector because it uses an exact replica of the carrier—is shown in Fig. 126.6(a). [Here,  $2W$  denotes the bandwidth of the message signal  $m(t)$ .] The essential difference between DSB and AM is that AM places the message signal in the *envelope* (i.e., in the magnitude of the instantaneous amplitude) of the carrier, and thus it can be demodulated without producing a replica of the carrier. In particular, an AM demodulator is illustrated in Fig. 126.6(b). The block marked **envelope detector** refers to a circuit that outputs the envelope of its input signal. Envelope detectors are usually implemented with a half-wave rectifier (e.g., a diode) followed by a low-pass filter.

**Figure 126.6** Demodulators of linearly modulated signals.



(a) Coherent demodulator for DSB



(b) Noncoherent demodulator for AM

Note that AM could also be demodulated coherently by using a DSB demodulator followed by a circuit that eliminates the direct current (DC) or constant part of the the linear modulation,  $[1 + km(t)]$ . However, the fact that AM can be demodulated via an envelope detector is its chief virtue, since it is less efficient than other forms of linear modulation. In particular, the addition of DC to the message signal before impressing it onto the carrier results in only a fraction of the electrical power contained in an AM signal being due to the message itself. This fraction, known as the *efficiency* of the modulation waveform, is given by

$$E = \frac{k^2 \langle m^2(t) \rangle}{1 + k^2 \langle m^2(t) \rangle} \quad (126.8)$$

where

$$\langle m^2(t) \rangle = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} [m(t)]^2 dt \quad (126.9)$$

is the average power in the message signal. The AM efficiency can never be greater than 1/2, and it can achieve this maximum only for very limited types of message signals.

Each of the demodulators of Fig. 126.6 is preceded by a band-pass filter, which passes only those frequencies contained in the modulated waveform. The purpose of this filter is to eliminate extraneous noise from the receiver, an issue that will be discussed later.

Two further types of linear modulation of interest are described as follows.

### Single-Sideband (SSB) Modulation

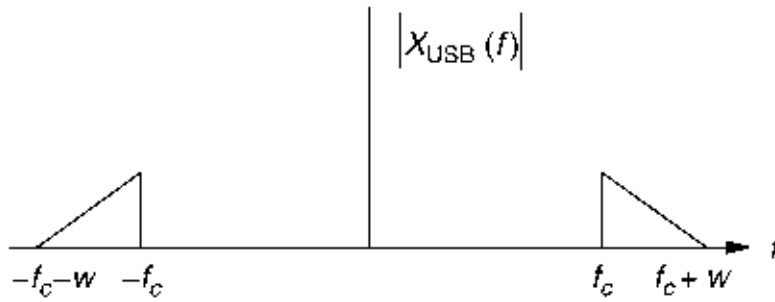
$$x_{\text{SSB}}(t) = \frac{A_c}{2} m(t) \cos(2\pi f_c t + \phi_c) \pm \frac{A_c}{2} \hat{m}(t) \sin(2\pi f_c t + \phi_c) \quad (126.10)$$

where  $\hat{m}(t)$  is the *Hilbert transform* of the message signal:

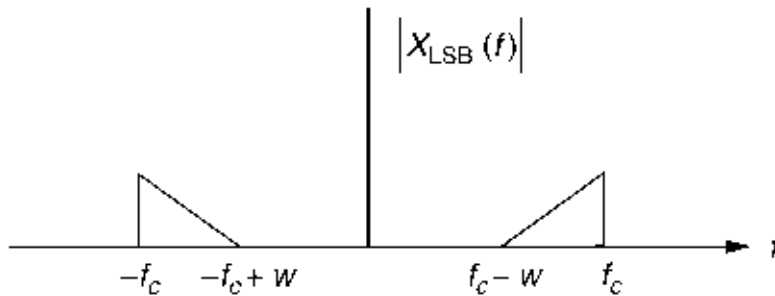
$$\hat{m}(t) = \int_{-\infty}^{\infty} \frac{m(\tau)}{t - \tau} d\tau \quad (126.11)$$

Taking the + in  $\pm$  in SSB yields lower-sideband (LSB) modulation, and taking the – yields upper-sideband (USB) modulation. The Fourier transforms of LSB and USB signals corresponding to the example of Fig. 126.5 are shown in Fig. 126.7. The basic motive behind SSB modulation is to remove redundancy that occurs in DSB due to inclusion of both positive and negative frequencies of the message signal. Since the message signal is real valued, the symmetry of its Fourier transform renders these two sidebands redundant. Thus, SSB can transmit the same information as DSB can while using only half as much bandwidth.

**Figure 126.7** SSB spectra.



(a) USB amplitude spectrum



(b) LSB amplitude spectrum

### Vestigial Sideband (VSB) Modulation

As seen from Fig. 126.7, SSB signals are DSB signals from which one of the sidebands has been removed. It is difficult to implement circuitry for removing these sidebands completely; an alternative is *VSB modulation*, in which a vestige of the unused sideband is also included in the modulated signal. This approach results in less stringent requirements on the hardware needed, while retaining most of the advantages of SSB modulation.

### Angle Modulation

There are two basic forms of continuous angle modulation.

*Phase modulation (PM):*

$$x_{\text{PM}}(t) = A_c \sin[2\pi f_c t + k_d m(t)] \quad (126.12)$$

Here, as before,  $A_c$  and  $f_c$  are the amplitude and frequency of carrier, and  $m(t)$  is the message signal. The constant  $k_d$  is a modulation index.

*Frequency modulation (FM):*

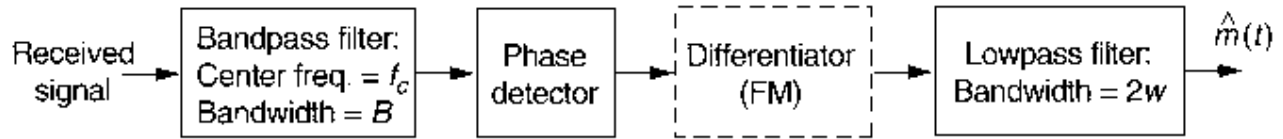
$$x_{\text{FM}}(t) = A_c \sin \left\{ 2\pi [f_c t + f_d \int^t m(u) du] \right\} \quad (126.13)$$

Here, the constant  $f_d$  is the *frequency deviation constant*, which quantifies the amount of frequency excursion that the carrier undergoes per unit of message signal amplitude.

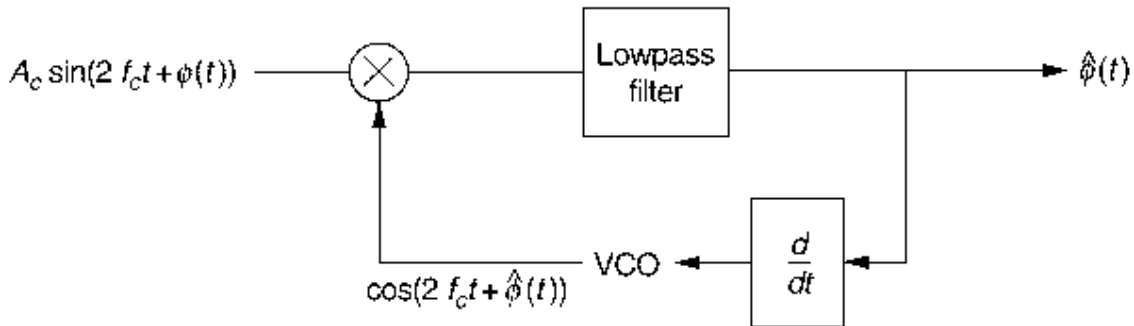
Note that PM and FM are essentially the same type of modulation, except that the frequency deviation of the carrier is proportional to the time derivative of the message signal in PM, and it is directly proportional to the message signal in FM.

A generic demodulator structure for angle modulation is shown in Fig. 126.8(a). The bandwidth  $B$  of the band-pass filter is the bandwidth of the angle-modulated signal, which can be significantly larger than the message bandwidth. The block labeled **phase detector** refers to a type of circuit that detects the instantaneous phase of its input. One such circuit is a *phase-lock loop* (PLL), which is shown in Fig. 126.8(b). The block labeled VCO refers to a *voltage-controlled oscillator*, which is a type of device that produces a sinusoidal signal whose frequency is proportional to the input voltage. Essentially, a PLL tracks the phase of a modulated sinusoid by applying feedback through a VCO. The low-pass filter in the PLL has bandwidth approximately equal to that of the phase signal  $\phi(t)$ , and the output signal  $\hat{\phi}(t)$  is an approximation of  $\phi(t)$ .

**Figure 126.8** Angle demodulation.



(a) Generic demodulator for angle-modulated signals



(b) Phase-lock loop

The Fourier transforms of angle-modulated signals are difficult to depict in generality because of the nonlinear way in which the message signal affects the carrier. However, it is generally true that,

for a given message bandwidth, angle-modulated signals typically require greater bandwidth than do linearly modulated signals. An empirical rule of thumb for the bandwidth required by FM is *Carson's rule*:

$$B \cong 2W(D + 1) \quad (126.14)$$

where  $B$  is the bandwidth of the modulated signal,  $2W$  is the bandwidth of the message signal, and  $D$  is the *deviation ratio*, given by

$$D = \frac{f_d}{W}M \quad (126.15)$$

where  $M$  is the maximum absolute value of the message signal  $m(t)$ . If  $D \ll 1$ , then  $B \cong 2W$  and the required bandwidth is approximately the same as that for DSB. Modulation of this type is called *narrowband* angle modulation. For large deviation ratios, however, angle modulation requires considerably more bandwidth than does linear modulation for the same message bandwidth. Angle modulation has advantages in defeating the nonideal effects of the channel. In particular, the fact that the transmission quality of angle modulation is unaffected by amplitude fluctuations is an advantage for channels in which there is amplitude fading. Also, the ability to spread the message over a larger bandwidth for transmission is useful in mitigating the effects of noise. This issue will be discussed further in the following section.

## Signal-to-Noise Ratio Analysis

In the preceding sections we commented on the bandwidth requirements and ease of demodulation of several modulation schemes. A further criterion differentiating such schemes is their performance in noisy channels. A measure of the quality of a waveform that consists of useful signal plus noise can be measured by the **signal-to-noise ratio (SNR)**, which is the ratio of the power contained in the useful signal to the power contained in the noise. SNR is typically expressed in *decibels* (dB), units that refer to  $10 \log_{10}$  of the actual ratio of powers.

The effectiveness of various modulation/demodulation schemes in combating noise can be compared by considering the SNR at the output of the demodulator relative to the *baseband SNR*, which is the SNR of a system that involves no modulation and demodulation. In order to compare modulation schemes on this basis, it is common to assume that the channel is corrupted by **additive white Gaussian noise (AWGN)**. AWGN is an additive random noise process that delivers a constant density (i.e., power per unit frequency) of power across all frequencies and whose amplitude statistics obey a normal (or Gaussian) probability distribution. This assumption provides a useful model for many types of noise-corrupting communication signals.

Relative SNRs for various demodulators in AWGN channels are shown in [Table 126.1](#). The values shown in this table for PM, FM, and envelope-detected AM are approximations valid only for large SNRs. From this table we see that DSB and SSB are superior to AM, whose relative SNR is twice the AM efficiency (and hence is never greater than 1). Also, DSB and SSB have

equivalent performance even though SSB requires only half the bandwidth required by DSB. It is also interesting that coherently detected AM and envelope-detected AM have identical performance for sufficiently large received SNR (greater than 10 dB). For received SNRs below 10 dB, the relative SNR of envelope detection is very poor. Similarly, note that for large received SNRs, the quality of demodulated FM signals can be made arbitrarily large at the expense of increased transmitted bandwidth (i.e., increased deviation ratio  $D$ ). As with AM, the performance of PM and FM degrades dramatically if the received SNR is not sufficiently high.

**Table 126.1** SNRs Relative to Baseband for Analog Modulation/Demodulation Techniques

Modulator/Demodulator	Relative SNR
DSB/SSB	1
Coherent AM	$\frac{k^2 \langle m^2(t) \rangle}{1 + k^2 \langle m^2(t) \rangle}$
Envelope-detected AM (large SNR)	$\frac{k^2 \langle m^2(t) \rangle}{1 + k^2 \langle m^2(t) \rangle}$
PM (large SNR)	$k_p^2 \langle m^2(t) \rangle$
FM (large SNR)	$3 \left( \frac{f_d}{W} \right)^2 \langle m^2(t) \rangle$

## 126.2 Digital Modulation and Detection

It is often advantageous to transmit messages as digital data. This approach allows greater flexibility in protecting the data from noise and other nonideal effects of the channel through the use of *error-control* (or channel) coding (see **Chapter 127**), and it also often allows greater efficiency in terms of usage of channel bandwidth through the use of *data compression* (or source coding). Some types of message sources, such as computer data or text, are inherently digital in nature. However, continuous information sources such as speech and images can be converted into digital data for transmission. This process involves *analog-to-digital* (A/D) conversion—that is, *time sampling* to convert the continuous-time message signal to a discrete-time message—and *amplitude quantization* to map the continuous amplitudes of the message signal into a finite set of values. In order to preserve the fidelity of the source, it is necessary that the sampling rate be sufficiently high to prevent loss of information and that the quantization be sufficiently fine to prevent undue distortion. [A minimum sampling rate to capture the message in discrete time is twice the bandwidth of the message (a rate known as the *Nyquist rate*).] Once the message is in digital form it can be converted to a sequence of binary words for transmission by encoding its discrete amplitude values. Modulation schemes that transmit analog messages in this way are known as *pulse code modulation* (PCM) schemes.

A number of forms of modulation can be used to transmit data through a communication channel. The most basic of these transmit a single binary digit (1 or 0) in each of a sequence of symbol intervals of duration  $T$ , where  $T$  is the reciprocal of the data rate (expressed in bits per second). To use such a scheme, the data sequence of binary words can be converted to a sequence

of bits via parallel-to-serial conversion. Most binary transmission schemes can be described in terms of two waveforms,  $x^{(0)}(t)$  and  $x^{(1)}(t)$ , where the transmitted waveform  $x(t)$  equals  $x^{(0)}(t)$  in a given symbol interval if the corresponding data bit is a 0, and  $x(t)$  equals  $x^{(1)}(t)$  if the corresponding data bit is a 1.

As with analog modulation, the basic manner in which most binary modulation procedures couple the data sequence to the channel is to impress it onto a sinusoidal carrier. Thus, we can have digital modulation schemes based on amplitude, phase, or frequency modulation, and the waveforms  $x^{(0)}(t)$  and  $x^{(1)}(t)$  are chosen accordingly to modify a basic carrier waveform  $A_c \sin(2\pi f_c t + \phi_c)$ , where  $A_c$ ,  $f_c$ , and  $\phi_c$  are, respectively, the amplitude, frequency, and phase of the carrier. Several fundamental techniques of this type are described in the following paragraphs.

## On-Off Keying (OOK)

OOK is the simplest type of binary modulation. It transmits the signal

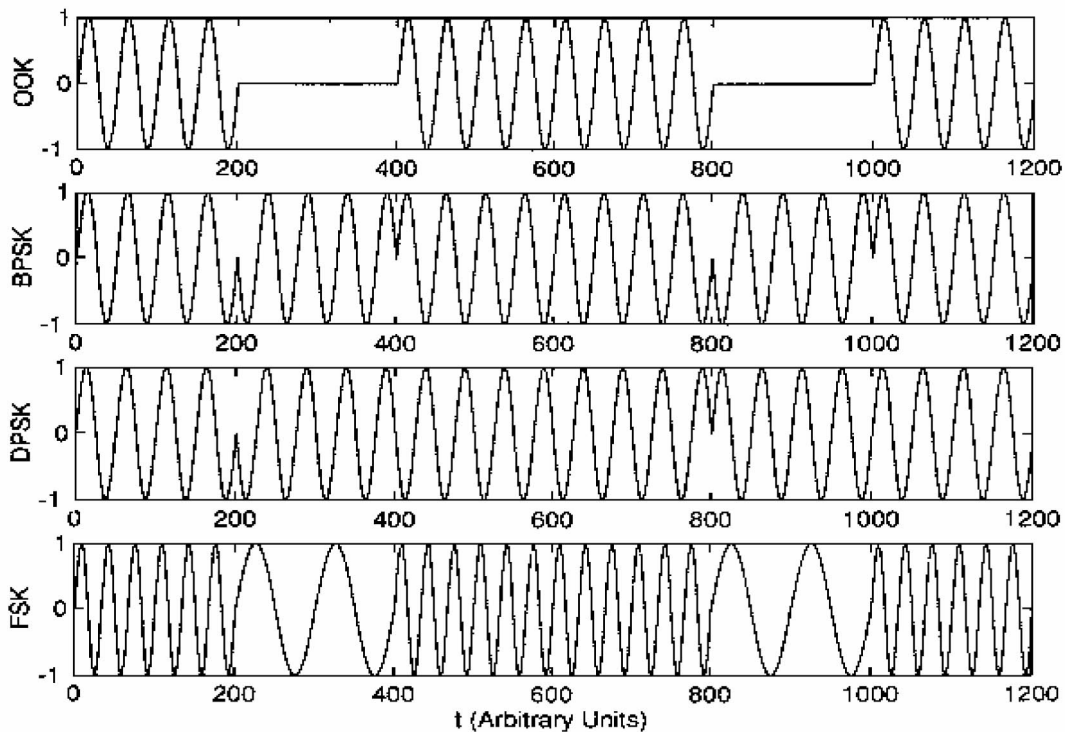
$$x_{\text{OOK}}^{(1)}(t) = A_c \sin(2\pi f_c t + \phi_c) \quad (126.16)$$

in a given symbol interval if the corresponding data bit is a 1, and it transmits nothing, that is,

$$x_{\text{OOK}}^{(0)}(t) = 0 \quad (126.17)$$

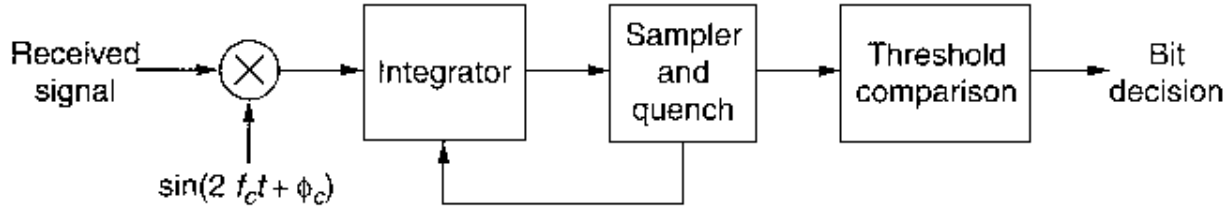
if the corresponding data bit is a 0. An OOK waveform is illustrated in [Fig. 126.9](#). OOK is a form of *amplitude-shift keying* (ASK) since it "keys" (i.e., modulates) the carrier by shifting its amplitude by an amount depending on the polarity of the data bit. ASK waveforms other than the "on-off" version described here can also be used.

**Figure 126.9** Digital modulation waveforms for transmitting the bit sequence 101101 ( $T = 200$ ).

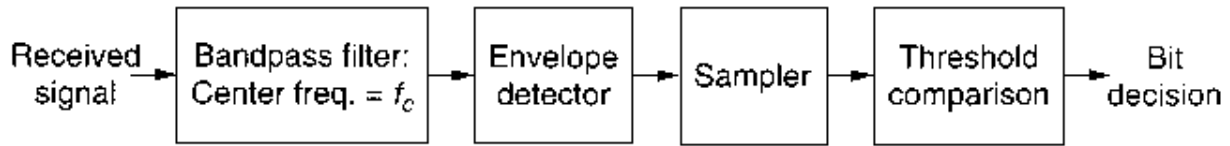


OOK can be demodulated either coherently (i.e., with knowledge of the carrier phase), as shown in Fig. 126.10 (a), or noncoherently (i.e., without knowledge of the carrier phase) as shown in Fig. 126.10(b). In either case the output of the detector (coherent or noncoherent) is sampled at the end of each symbol interval and compared with a threshold. If this output exceeds the threshold for a given sampling time, then the corresponding data symbol is detected as a 1; otherwise, this symbol is detected as a 0. (In a coherent detector, the integrator is "quenched"—that is, reset to 0—as it is sampled.) Errors occur in these systems because the noise in the channel can move the output of the detector to the incorrect side of the threshold. The proper choice of the threshold to minimize this effect, and the corresponding rate of bit errors, are discussed below.

**Figure 126.10** Demodulation of OOK.



(a) Coherent demodulator for OOK (and BPSK)



(b) Noncoherent demodulator for OOK

## Phase-Shift Keying (PSK)

As its name suggests, PSK uses the phase of the carrier to encode the binary data to be transmitted. The basic forms of PSK are described in the following sections.

### Binary PSK (BPSK)

This form of modulation uses the following waveforms:

$$x_{\text{BPSK}}^{(1)}(t) = A_c \sin(2\pi f_c t + \phi_c) \quad (126.18)$$

$$x_{\text{BPSK}}^{(0)}(t) = A_c \sin(2\pi f_c t + \phi_c + \pi) \equiv -A_c \sin(2\pi f_c t + \phi_c) \quad (126.19)$$

A BPSK waveform is illustrated in Fig. 126.9. This type modulation uses antipodal signaling [i.e.,  $x^{(0)}(t) = -x^{(1)}(t)$ ]. Note that BPSK is also a form of ASK, in which the two amplitudes are  $\pm 1$ .

Since the information in BPSK is contained in the carrier phase, it is necessary to use coherent detection in order to accurately demodulate BPSK. The demodulator of Fig. 126.10(a) can be used to perform this demodulation. In the case of BPSK the decision threshold can be taken to be zero due to the antipodal nature of the signaling.

### Differential PSK (DPSK)

The necessity of knowing the carrier for demodulation of BPSK is a disadvantage that can be overcome by the use of DPSK. In a given bit interval (say the  $k$ th one), DPSK uses the following



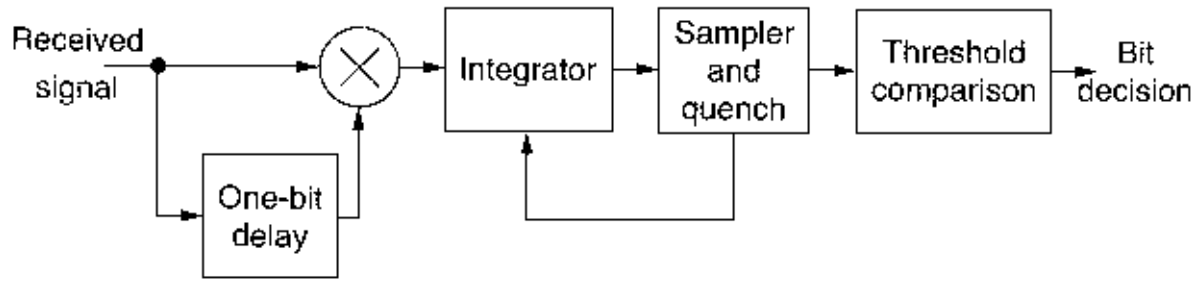
waveforms,

$$x_{\text{DPSK}}^{(1)}(t) = A_c \sin(2\pi f_c t + \phi_{k-1}) \quad (126.20)$$

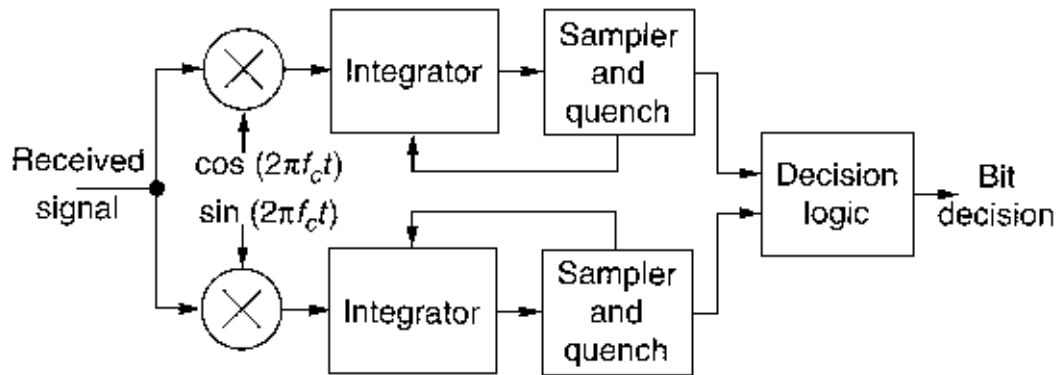
$$x_{\text{DPSK}}^{(0)}(t) = A_c \sin(2\pi f_c t + \phi_{k-1} + \pi) \quad (126.21)$$

where  $\phi_{k-1}$  denotes the phase transmitted in the *preceding* bit interval (i.e., the  $(k - 1)$  th bit interval). Thus, the information is encoded in the difference between the phases in succeeding bit intervals rather than in the absolute phase, as illustrated in Fig. 126.9. (DPSK requires an initial reference bit, which is taken to be 1 in the illustration.) This step allows for noncoherent demodulation of DPSK, as shown in Fig. 126.11.

**Figure 126.11** Demodulation of DPSK.



(a) Suboptimum demodulator for DPSK



(b) Optimum demodulator for DPSK

Note that the demodulator in Fig. 126.11(a) does not require knowledge of the carrier phase or frequency, whereas that in Fig. 126.11(b) requires knowledge of the carrier frequency but not its phase. The block marked "decision logic" in Fig. 126.11(b) makes each bit decision based on two successive pairs of outputs of the two channels that provide its inputs. In particular, the  $k$ th bit is

demodulated as a 1 if  $p_k p_{k-1} + q_k q_{k-1} > 0$  and as a 0 otherwise, where  $p_k$  and  $p_{k-1}$  are the outputs of the upper channel (known as *in-phase* channel) at the end of the  $k$ th and  $(k - 1)$  th bit intervals, respectively, and where  $q_k$  and  $q_{k-1}$  are the corresponding outputs of the lower channel (the *quadrature* channel). The second of these demodulators is actually optimum for demodulating DPSK and, as such, exhibits performance advantages over the first. This performance comes in exchange for the obvious disadvantage of requiring carrier-frequency reference signals at the receiver.

### Quadrature PSK (QPSK)

The bandwidth efficiency of BPSK can be improved by taking advantage of the fact that there is another pair of antipodal signals, namely,

$$A_c \cos(2\pi f_c t + \phi_c) \quad (126.22)$$

$$A_c \cos(2\pi f_c t + \phi_c + \pi) \quad (126.23)$$

that have the same frequency as the two signals used in BPSK [i.e.,  $x_{\text{BPSK}}^{(1)}(t)$  and  $x_{\text{BPSK}}^{(0)}(t)$  of Eqs. (126.18) and (126.19)] while being completely orthogonal to those signals. By using all four of these signals, two bits can be sent in each symbol interval, thereby doubling the transmitted bit rate. Such a signaling scheme is known as *QPSK* because it involves the simultaneous transmission of two BPSKs in quadrature (i.e., 90 degrees out of phase). Although the performance in terms of bit-error of rate of QPSK is the same as that for BPSK, QPSK has the advantage of requiring half the bandwidth needed by BPSK to transmit at the same bit rate. This situation is directly analogous to that involving DSB and SSB analog modulation, the latter of which uses two quadrature signals to transmit the same information as the former does, while using only half the bandwidth.

Note that QPSK is a form of  $M$ -ary signaling, in which  $\log_2 M$  bits are transmitted in each symbol interval by selecting among  $M$  possible waveforms,  $x^{(0)}(t), x^{(1)}(t), \dots, x^{(M-1)}(t)$ . In QPSK we have the case  $M = 4$  and the waveforms are four sinusoids having the same frequency and having phases separated by 90 degrees. QPSK can be generalized by allowing more than four separate phases and also by allowing modulation types other than PSK. Such techniques are widely used in modern digital communication systems such as high-speed data modems.

### Frequency-Shift Keying (FSK)

FSK transmits binary data by sending one of two distinct frequencies in each bit interval, depending on the polarity of the bit to be transmitted. This scheme can be described in terms of the two signaling waveforms,

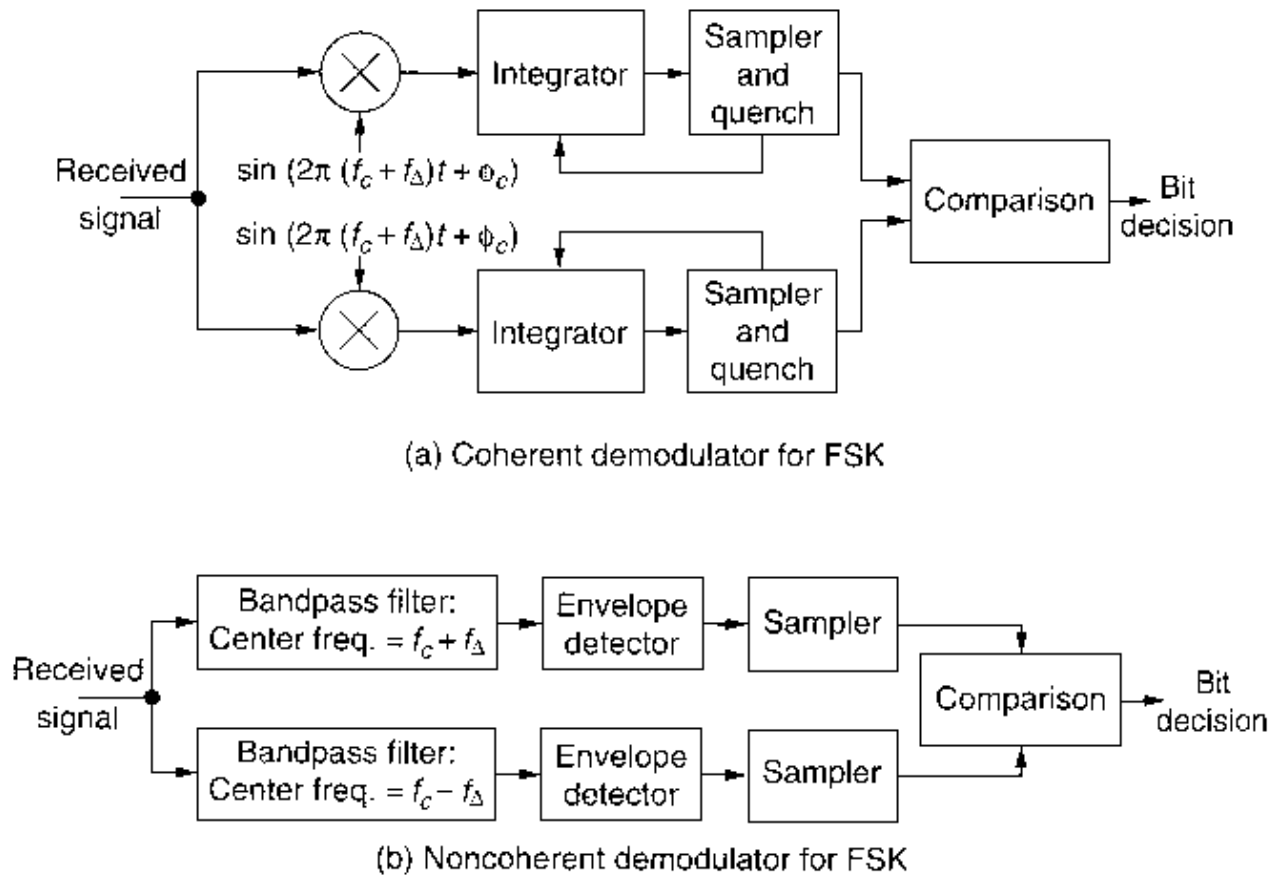
$$x_{\text{FSK}}^{(1)}(t) = A_c \sin(2\pi(f_c + f_\Delta)t + \phi_c) \quad (126.24)$$

$$x_{\text{FSK}}^{(0)}(t) = A_c \sin(2\pi(f_c - f_\Delta)t + \phi_c) \quad (126.25)$$

where  $f_\Delta$  is a constant. An FSK waveform is illustrated in Fig. 126.9.

FSK can be demodulated either coherently or noncoherently, as shown in Fig. 126.12. In these demodulators the block marked "comparison" chooses the bit decision as a 1 if the upper-channel output is larger than the lower-channel output and as a 0 otherwise. Note that, when noncoherent demodulation is to be used (as is very commonly the case), it is not necessary that the carrier phase be maintained from bit interval to bit interval. This simplifies the design of the modulator and makes noncoherent FSK one of the simplest types of digital modulation. It should be noted, however, that FSK generally requires greater bandwidth than do the other forms of digital modulation described above.

**Figure 126.12** Demodulation of FSK.



## Bit-Error Rates and Bandwidth Efficiency

The various binary modulation/demodulation types described above can be compared by analyzing

their **bit-error rates (BERs)** or bit-error probabilities—that is, the probabilities with which they incur errors in receiving bits. As in the analysis of analog modulation/demodulation, this comparison is commonly done by assuming that the channel is corrupted by AWGN. In this case and under the further assumption that the symbols 0 and 1 are equally likely to occur in the message, expressions for the bit-error probabilities of the various schemes described above are shown in [Table 126.2](#). These results are given as functions of the SNR parameter,  $E_b/N_0$ , where  $E_b$  is the signal energy received per bit and  $N_0/2$  is the spectral density of the AWGN. In some cases the expressions involve the function  $Q$ , which denotes the tail probability of a standard normal probability distribution:

$$Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-y^2/2} dy$$

The expressions for OOK assume that the decision threshold in the demodulators have been optimized. This optimization requires knowledge of the received SNR, which makes OOK the only one of these techniques that requires this information for demodulation. The expression for noncoherent OOK is an approximation that is valid for large SNRs. The result for DPSK corresponds to the optimum demodulator depicted in [Fig. 126.11\(b\)](#). The suboptimum DPSK demodulator of [Fig. 126.11\(a\)](#) requires approximately 2 dB higher values of  $E_b/N_0$  in order to achieve the same performance as the optimum demodulator.

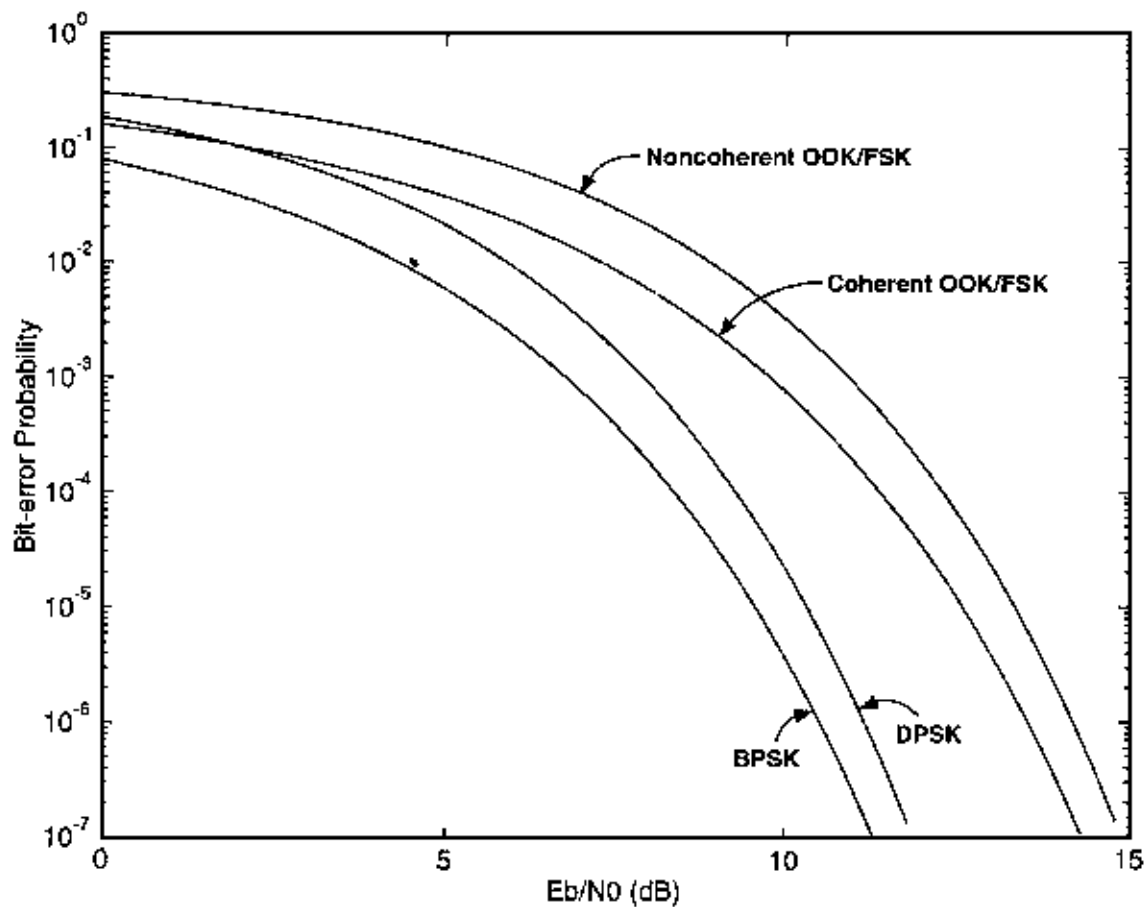
**Table 126.2** Bit-Error Probabilities and Bandwidth Efficiencies for Digital Modulation/Demodulation Techniques

Modulator/Demodulator	Bit-Error Probability	Bandwidth Efficiency
BPSK	$Q(\sqrt{2E_b/N_0})$	1/2 (bits/s)/hertz
QPSK	$Q(\sqrt{2E_b/N_0})$	1
Optimum DPSK	$(1/2)e^{-E_b/N_0}$	1/2
Coherent OOK	$Q(\sqrt{E_b/N_0})$	1/2
Coherent FSK	$Q(\sqrt{E_b/N_0})$	1/3
Noncoherent OOK	$(1/2)e^{-E_b/2N_0}$	1/2
Noncoherent FSK	$(1/2)e^{-E_b/2N_0}$	1/3

The quantities of [Table 126.2](#) are plotted in [Fig. 126.13](#). From this figure we see that BPSK is the best performing of these schemes, followed, in order, by DPSK, coherent OOK and FSK, and noncoherent OOK and FSK. The superiority of BPSK is due to its use of antipodal signals, which can be shown to be an optimum choice in this respect for signaling through an AWGN channel. DPSK exhibits a small loss relative to BPSK, which is compensated for by its simpler demodulation. OOK and FSK are both examples of *orthogonal* signaling schemes (i.e., schemes in which  $\int x^{(0)}(t)x^{(1)}(t) dt = 0$ , where the integration is performed over a single bit interval), which explains why they exhibit the same performance. Orthogonal signaling is less efficient than antipodal signaling, which is evident from [Fig. 126.13](#). Finally, note that there is a small loss in performance for these two orthogonal signaling schemes when they are demodulated

noncoherently. Generally speaking, DPSK and noncoherent FSK are seen from this comparison to be quite utilitarian. They each require less than a single dB more of received SNR to achieve the same performance as the best antipodal and orthogonal signaling schemes, respectively, while allowing relatively simple demodulation. For these reasons these two schemes are quite commonly used in practice.

**Figure 126.13** Bit-error probabilities for digital communication systems.



In addition to bit-error rate, digital communication systems can also be compared in terms of bandwidth efficiency, which is often quantified in terms of the number of bits per second that can be transmitted per hertz of bandwidth. Bandwidth efficiencies for the various signaling schemes are also shown in [Table 126.2](#).

## 126.3 Further Issues

There are many types of modulation and demodulation used in practice other than those described

in the preceding sections. Some of these are discussed briefly in the following paragraphs.

## Pulse Modulation

A general class of modulation techniques that are used to transmit analog messages in discrete time are known as *pulse modulation* schemes. Such techniques convert an analog message signal into a discrete-time one by time sampling, and then transmit these time samples by using them to modulate the properties of a train of pulses. Techniques in this category include *pulse-amplitude modulation* (PAM), *pulse-width modulation* (PWM), and *pulse-position modulation* (PPM), whose characteristics can be inferred from their names.

Recall that we have already mentioned *pulse-code* modulation, in which the amplitudes of the time-sampled message are quantized and encoded for digital transmission. Two forms of PCM that are commonly used in practice are *delta* modulation (DM) and *differential* PCM (DPCM), which encode message samples by digitally representing the *change* in the message from sampling time to sampling time. This change can be measured in terms of a direct difference or, more commonly, as the difference between the sample at one time and a value of that sample predicted from past values through the use of a prediction algorithm. These techniques generally require fewer data bits to represent each message sample, at the expense of somewhat higher sampling rate requirements. Differential encoding techniques are particularly effective in transmitting speech signals, which contain considerable redundancy that can be removed through such encoding.

## Multiplexing and Multiple Access

As noted early in the chapter, one purpose of modulation is *multiplexing*, which refers to the simultaneous transmission of multiple message signals through a single channel (i.e., on a single carrier frequency). One form of multiplexing is *quadrature* multiplexing (QM), such as that employed in QPSK, in which two orthogonal carriers of the same frequency are employed. In addition to this form of multiplexing, there are two other basic types: *frequency-division* multiplexing (FDM) and *time-division* multiplexing (TDM). FDM involves the modulation of the message signals to be multiplexed onto distinct low-frequency carriers, which are then added to form a bulk baseband signal that is modulated onto the main carrier. The low-frequency carriers, known as *subcarriers*, must be spaced sufficiently far apart in frequency to prevent overlap of the various message signals when the bulk signal is formed. TDM involves, instead, the interlacing in time of samples from a group of (discrete-time) message signals to form a bulk baseband signal to be modulated onto the main carrier.

A process related to multiplexing is *multiple-access* transmission. Like multiplexing, multiple-access transmission involves the sharing of channel resources by multiple message sources, or *users*. However, whereas *multiplexing* usually refers to the situation in which the multiple messages are combined centrally for transmission, multiple-access transmission allows each message source to access the channel remotely and independently of the other sources. (An example of an application of multiple-access communications is a cellular telephony system.) Commonly used multiple-access techniques include *frequency-division* multiple access (FDMA)

and *time-division* multiple access (TDMA), which divide the channel among the users on the basis of frequency and time, respectively (analogous to FDM and TDM). A further basic type of multiple-access technique is code-division multiple access (CDMA), which divides the channel by encoding each user in a way that a receiver can distinguish it from the other users.

## Spread-Spectrum Signaling

CDMA is usually implemented through the use of *spread-spectrum signaling*, which refers to signaling techniques in which the bandwidth of the signal transmitted by a given user is much larger than the bandwidth required for the user's message signal. There are two basic ways in which spread spectrum can be implemented: *direct-sequence* spread spectrum (DSSS) and *frequency-hopping* spread spectrum (FHSS).

In DSSS the message signal to be transmitted by a given user is multiplied (before modulation onto the carrier) by a signal (known as a *pseudonoise* or PN signal) whose Fourier transform is constant over the entire bandwidth of the shared channel and whose amplitude is always  $\pm 1$ . The resulting transmitted signal uniformly occupies the entire bandwidth of the shared channel. A receiver knowing the PN signal can recover the original message signal by removing the carrier and multiplying the remaining signal by the PN signal. Since the amplitude of the PN signal is always 1, this multiplication removes the PN signal from the message signal, allowing the latter to be detected. In this way all users can in principle share all of the bandwidth of the channel all of the time. Interference among users is kept at a minimum by assigning a unique PN signal to each user in the channel and by choosing this set of PN signals appropriately. Although the use of this type of modulation does not, of course, create new bandwidth for message transmission, it does permit effective use of bandwidth in channels (such as cellular telephony channels) in which the traffic is sporadic or in which there is significant amplitude fading. DSSS signaling also offers other advantages over narrowband communication systems, including a greater degree of privacy and greater immunity to certain types of noise and interference.

As can be inferred from its name, FHSS performs spectrum spreading by changing the carrier frequency used by an individual user in a regular pattern while the user is transmitting its message. Each user's frequency "hops" from frequency to frequency in a distinct pattern, and the choice of the set of "hopping patterns" is used to prevent undue interference among the active users in the channel.

## Sequences of Symbols

The digital modulation and demodulation techniques described thus far involve the modulation and demodulation of a single symbol at a time. In some circumstances, though, it is desirable (or necessary) to modulate or to demodulate strings of symbols taken as a group. This can be done at the transmitter for the purposes of coding (resulting in *coded modulation*) or at the receiver for the purposes of correcting nonideal effects of the channel. Among techniques falling into this latter category is *equalization*, which is applied to undo the effects of symbol mixing in the channel [known as *intersymbol interference* (ISI)]. Discussions of these and other advanced methods can

be found in the sources cited at the end of the chapter.

## Implementation Issues

A further important issue that has not been discussed in this chapter, is the implementation of the various devices that comprise the modulators and demodulators described herein. For example, in general, none of the received carrier parameters (amplitude, frequency, and phase) can be assumed to be known at the receiver without special provision for determining them, since none of these parameters may coincide with their transmitted counterparts. Certainly the amplitude and phase of the carrier will be modified by the channel simply by attenuation and time delay, and the frequency may also be modified by Doppler shift if the transmitter and receiver are moving relative to one another. Likewise, the digital transmission the timing of the symbol sequence (which must be known for choosing the correct sampling times to detect the data) is not generally known in advance by the receiver. Thus, in order to employ any of this information in demodulation, it is necessary to provide means for acquiring it. Moreover, devices for performing carrier generation and multiplication, phase detection, envelope detection, and so forth require special consideration.

## Defining Terms

**Additive white Gaussian noise (AWGN):** An additive random noise process that delivers a constant density (i.e., power per unit frequency) of power across all frequencies and whose amplitude statistics obey a normal (or Gaussian) probability distribution.

**Angle modulation:** The impression of a message onto a sinusoidal carrier by varying the phase of the carrier. Types of angle modulation include phase modulation (PM), frequency modulation (FM), phase-shift keying (PSK), and frequency-shift keying (FSK).

**Bandwidth:** For a *channel*, the property of the channel that describes the maximum rate of change in the input to which the channel can respond. This quantity limits the signaling rate that the channel can accommodate. For a *signal*, the width of its amplitude spectrum in the frequency domain. For a *filter*, the width of its passband.

**Bit-error rate (BER):** The rate at which a digital communication system incurs bit errors.

**Carrier:** A signal onto which a message signal is impressed for the purposes of transmitting the message signal through a communication channel.

**Channel:** The medium through which a message is to be transmitted by a communication system.

**Demodulation/detection:** The process of extracting a message signal from a modulated carrier.

**Envelope detector:** A device whose output is the instantaneous magnitude of an input sinusoidal signal.

**Linear modulation:** The impression of a message onto a sinusoidal carrier by varying the amplitude of the carrier. Types of linear modulation include double sideband (DSB),



amplitude modulation (AM), single sideband (SSB), vestigial sideband (VSB), amplitude-shift keying (ASK), and on-off keying (OOK).

**Message signal:** A signal containing the information to be transmitted by a communication system.

**Modulation:** The process of impressing a message signal onto a carrier.

**Phase detector:** A device whose output is the instantaneous phase of an input sinusoidal signal.

**Receiver:** An element of a communication system that extracts the message signal from the signal received at the output of the channel.

**Signal-to-noise ratio (SNR):** A measure of the quality of a noisy signal, defined as the ratio of the power contained in the useful part of the signal to the power contained in the noise. SNR is typically expressed in *decibels* (dB), whose units refer to  $10 \log_{10}$  of the actual ratio of powers.

**Transmitter:** An element of a communication system that couples the message signal to the channel.

## References

- Carlson, A. B. 1986. *Communication Systems: An Introduction to Signals and Noise in Electrical Communication*, 3rd ed. McGraw-Hill, New York.
- Couch, L. W., II. 1990. *Digital and Analog Communication Systems*, 3rd ed. Macmillan, New York.
- Gibson, J. D. 1993. *Principles of Digital and Analog Communications*, 2nd ed. Macmillan, New York.
- Haykin, S. 1994. *Communication Systems*, 3rd ed. John Wiley & Sons, New York.
- Proakis, J. G. 1983. *Digital Communications*. McGraw-Hill, New York.
- Stark, H., Tuteur, F. B., and Anderson, J. B. 1988. *Modern Electrical Communications: Analog, Digital and Optical Systems*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.
- Viterbi, A. J. and J. K. Omura. 1979. *Principles of Digital Communication and Coding*. McGraw-Hill, New York.
- Ziener, R. E. and Tranter, W. H. 1990. *Principles of Communications: Systems, Modulation and Noise*, 4th ed. Wiley, New York.

## Further Information

In addition to these textbooks, a number of journals describe recent advances in the field of modulation and detection. Among the most widely circulated of these are the *IEEE Transactions on Communications*, the *IEEE Journal on Selected Areas in Communications*, and the *IEEE Transactions on Information Theory*, all published by the Institute of Electrical and Electronics Engineers, New York; and the *IEE Proceedings<sup>3/4</sup> Communications* published by the Institution of Electrical Engineers, London.

Miller, S. L., Couch II, L. W. "Coding"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 127

## Coding

---

### 127.1 Block Codes

### 127.2 Convolutional Codes

### 127.3 Trellis-Coded Modulation

**Scott L. Miller**

*University of Florida*

**Leon W. Couch II**

*University of Florida*

Error correction coding involves adding redundant symbols to a data stream in order to allow for reliable transmission in the presence of channel errors. Error correction codes fall into two main classes: block codes and convolutional codes. In block codes the code word is only a function of the current data input, whereas in convolutional codes the current output is a function of not only the current data input but also of previous data inputs. Thus a convolutional code has memory, whereas a block code is memoryless. Codes can also be classified according to their code rate, symbol alphabet size, error detection/correction capability, and complexity. The code rate determines the extra bandwidth needed and also the reduction in energy per transmitted symbol relative to an uncoded system. The alphabet size is usually either two (binary) or a power of two. When the alphabet size is binary, the symbol is called a bit. Codes can be made either to detect or to correct errors or some combination of both and also can be made to detect/correct either random errors or burst errors. The complexity of a code is usually a function of the decoding procedure that is used, so the most popular codes are those that can be decoded in a relatively simple manner and still provide good error correction capability.

## 127.1 Block Codes

---

An  $(n, k)$  block code is a mapping of a  $k$  symbol information word into an  $n$  symbol code word. The code is said to be systematic if the first  $k$  symbols in the code word are the information word. In an  $(n, k)$  code,  $n - k$  symbols are redundant and are often referred to as *parity symbols*. The code rate,  $R$ , is defined as  $R = k/n$ . An important parameter of a block code is its minimum distance,  $d_{\min}$ , which is defined as the minimum Hamming distance between any two code words. (Hamming distance is the number of places in which two code words disagree.) A code with a minimum distance of  $d_{\min}$  can guarantee to correct any pattern of  $t = \lfloor (d_{\min} - 1)/2 \rfloor$  or fewer

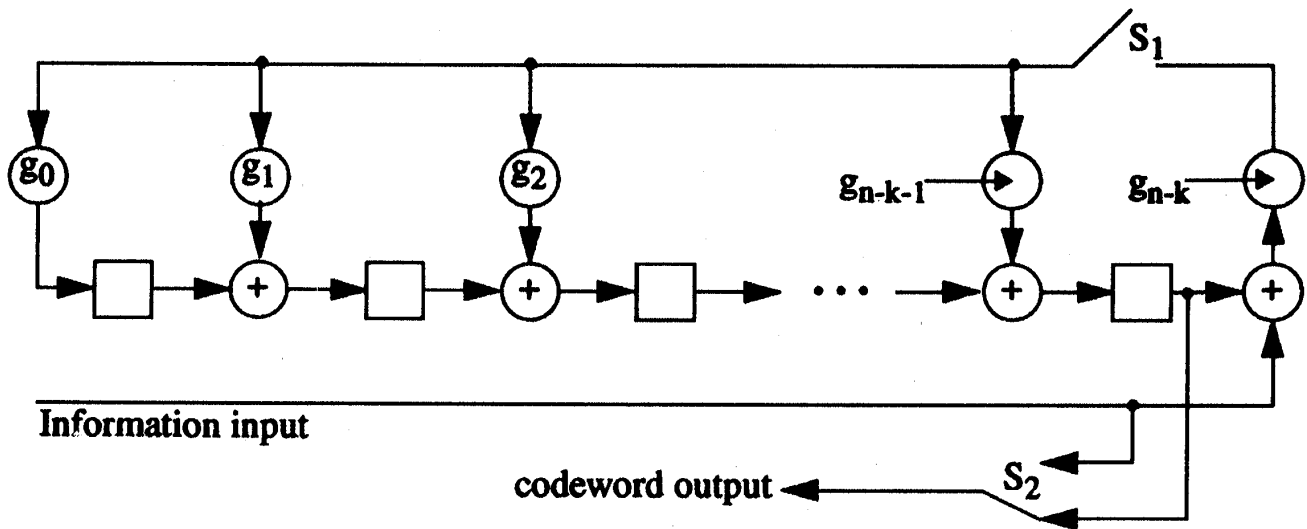
errors. Alternatively, the code can guarantee to detect any error pattern consisting of  $s = d_{\min} - 1$  or fewer errors. If simultaneous error detection and correction is desired, the minimum distance of the code must be such that  $d_{\min} \geq s + t + 1$ .

Most block codes used in practice are linear (or group) codes. In a linear code the set of code words must form a mathematical group. Most importantly, the sum of any two code words must itself be a code word. Encoding of linear block codes is accomplished by multiplying the  $k$  symbol information word by an  $n \times k$  matrix (known as a generator matrix) forming an  $n$  symbol code word. To decode, a received word is multiplied by an  $n \times (n - k)$  matrix (known as a *parity check matrix*) to form an  $n - k$  symbol vector known as the *syndrome*. A look-up table is then used to find the error pattern that corresponds to the given syndrome and that error pattern is subtracted from the received word to form the decoded word. The parity (redundant) symbols are then removed to form the decoded information word. In general, unless the number of parity symbols ( $n - k$ ) of a linear code is very small, syndrome decoding is too complicated and so codes with a simpler decoding procedure are used.

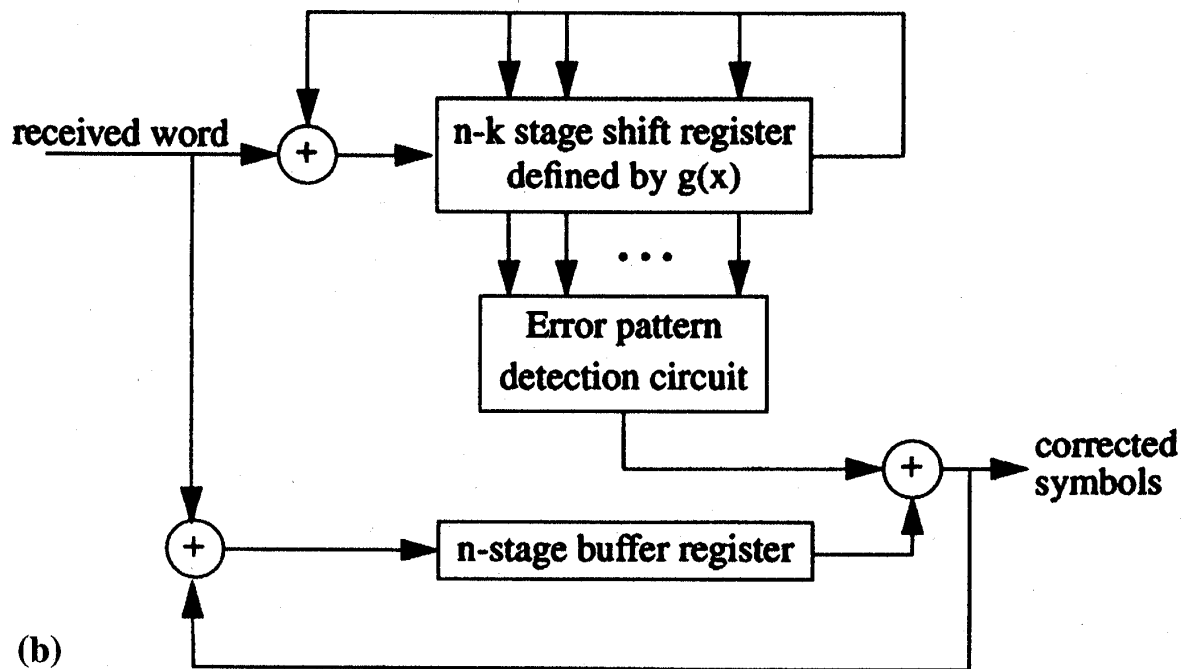
A cyclic code is a group code that has the additional property that any cyclic shift of a code word produces another code word. A cyclic code is specified by a generator polynomial, which is a polynomial of order  $n - k$  of the form  $g(x) = g_0 + g_1x + \cdots + g_{n-k}x^{n-k}$ , or by its parity check polynomial,  $h(x)$ , which is related to  $g(x)$  by  $g(x)h(x) = x^n - 1$ . The code is encoded with a division circuit (a linear feedback shift register) as shown in Fig. 127.1(a). In that circuit, switches  $S_1$  and  $S_2$  are in the closed and up positions, respectively, until the  $k$  data symbols are clocked in. After that, the two switches are flipped and the parity symbols are read out. The received sequence can be decoded using a similar circuit known as the *Meggitt decoder*, shown in Fig. 127.1(b). Most of the complexity of the decoder is in the error pattern detector, which must recognize all correctable error patterns with errors in the last position. For a single-error correcting code this can be implemented with a multiple-input AND gate, but for codes with large error correction capability the number of error patterns that must be recognized becomes unfeasible.

A technique known as *error-trapping decoding* can be used to simplify the error pattern detector of the Meggitt decoder, but again, this technique only works for short codes with low error correcting capability. Error trapping is quite effective for decoding burst error correcting codes and also works nicely on the well-known Golay code. When longer codes with high error correction capability are needed, it is common to consider various subclasses of cyclic group codes that lead to implementable decoding algorithms. One such subclass that is extremely popular comprises the BCH codes. Since BCH codes are cyclic codes, they can be encoded with the same general encoder shown in Fig. 127.1(a). In addition, a well-known technique called the *Berlekamp algorithm* can be used to decode BCH codes. The Berlekamp algorithm is fast and easy to implement, but it is not a maximum likelihood decoding technique, and so it does not always take advantage of the full error correcting capability of the code. In addition, it is often desirable for a demodulator to feed soft decisions on to the decoder. This will usually enable the decoder to make a more reliable decision about which symbols are in error. A demodulator is said to make soft decisions when it quantizes its output into an alphabet size larger than the size of the transmitted symbol set. All of the standard techniques for decoding block codes suffer from the fact that they cannot be used on soft decisions; this is the main reason why convolutional codes are often preferred.

**Figure 127.1** (a) Encoder for a cyclic group code; (b) Meggitt decoder for cyclic group codes.



(a)



(b)

For a memoryless channel with an error rate of  $p$ , the probability of decoding error,  $P_d$ , for an  $(n, k)t$ - error correcting code is given by

$$P_d = \sum_{i=t+1}^n \binom{n}{i} p^i (1-p)^{n-i} = 1 - \sum_{i=0}^t \binom{n}{i} p^i (1-p)^{n-i}$$

Finding exact expressions for the probability of a symbol error,  $P_s$  (i.e., probability of bit error or bit-error rate for the case of binary symbols), is generally not tractable, but the following bounds work quite well:

$$\frac{d_{\min}}{n} P_d \leq P_s \leq \sum_{i=t+1}^n \binom{n}{i} \frac{i+t}{n} p^i (1-p)^{n-i}$$

Another method of specifying the performance of an error correction coding scheme focuses on coding gain. Coding gain is the reduction in dB of required signal-to-noise ratio needed to achieve a given error rate relative to an uncoded system. Specifically, if an uncoded system needs an energy per bit of  $E_u(P_s)$  to achieve a symbol error rate of  $P_s$ , and a coded system requires  $E_c(P_s)$ , the coding gain is

$$\gamma(P_s) = 10 \log \{ [E_u(P_s)] / [E_c(P_s)] \} \text{ dB}$$

The asymptotic coding gain is the coding gain measured in the limit as  $P_s$  goes to zero and is given for a Gaussian noise channel by

$$\gamma = \lim_{P_s \rightarrow 0} \gamma(P_s) = 10 \log \left( \frac{k}{n} (t+1) \right) \text{ dB}$$

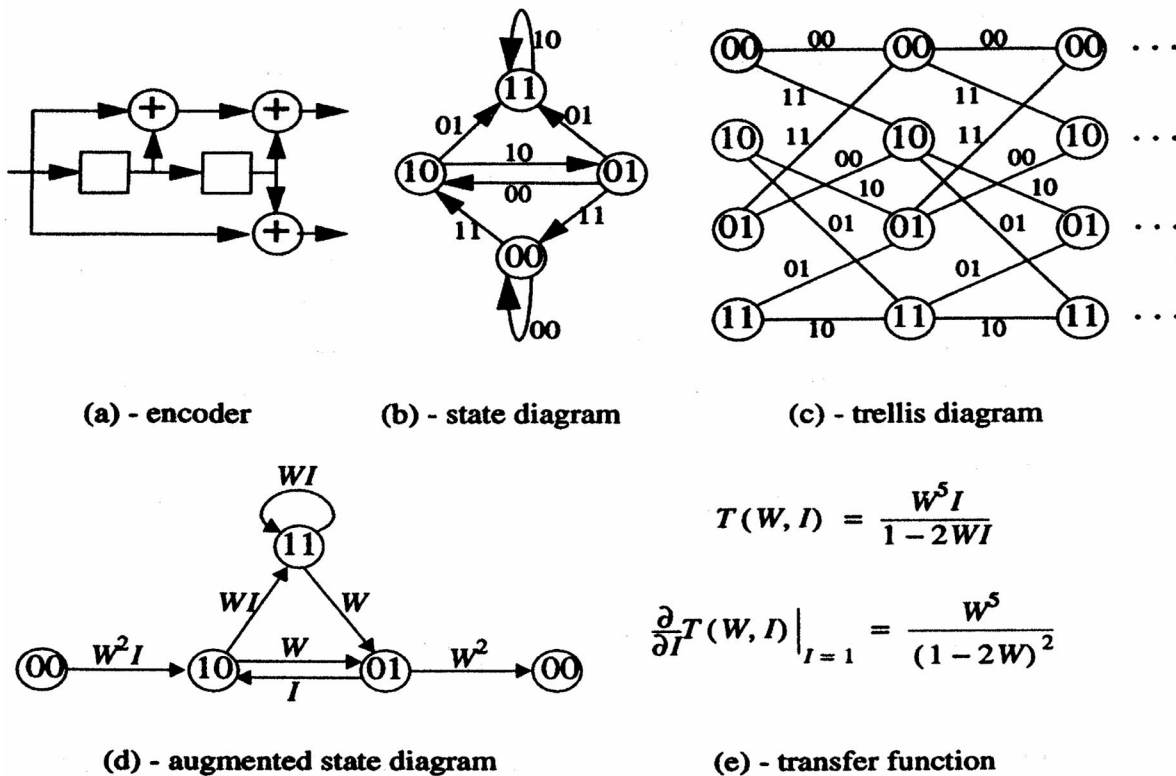
## 127.2 Convolutional Codes

---

In a convolutional code, as in a block code, for each  $k$  information symbols,  $n - k$  redundant symbols are added to produce  $n$  code symbols and the ratio of  $k/n$  is referred to as the code rate,  $R$ . The mapping, however, is not memoryless, and each group of  $n$  code symbols is a function of not only the current  $k$  input symbols but also the previous  $K - 1$  blocks of  $k$  input symbols. The parameter  $K$  is known as the *constraint length*. The parameters  $n$  and  $k$  are usually small and the code word length is arbitrary (and potentially infinite). As with block codes, the ability to correct errors is determined by  $d_{\min}$ , but in the case of convolutional codes this quantity is generally referred to as the *free distance*,  $d_{\text{free}}$ . Although the error correction capability,  $t$ , has the same relationship to  $d_{\text{free}}$  as in block codes, this quantity has little significance in the context of convolutional codes.

A convolutional code is generated using a feed-forward shift register. An example of the structure for the convolutional encoder is shown in Fig. 127.2(a) for a code with  $R = 1/2$  and  $K = 3$ . In general, for each group of  $k$  symbols input to the encoder,  $n$  symbols are taken from the output. A convolutional code can be described graphically in terms of a state diagram or a trellis diagram, as shown in Figs. 127.2(b) and 127.2(c) for the encoder of Fig. 127.2(a). The state diagram represents the operation of the encoder, whereas the trellis diagram illustrates all the possible code sequences. The states in both the state and trellis diagrams represent the contents of the shift register in the encoder, and the branches are labeled with the code output corresponding to that state transition.

**Figure 127.2** A rate  $1/2$ ,  $K = 3$  convolutional encoder.



The optimum (maximum likelihood) method for decoding convolutional codes is the Viterbi algorithm, which is nothing more than an efficient way to search through the trellis diagram for the most likely transmitted code word. This technique can be used equally well with either hard or soft channel decisions and has a complexity that is exponential in the constraint length. As a result the Viterbi algorithm can be used only on codes with a moderate constraint length (currently about  $K \leq 10$ ). For most applications this limit on  $K$  is not too severe and so Viterbi decoding is used almost exclusively. If a larger constraint length is needed, several suboptimal decoding algorithms exist whose complexities do not depend on the constraint length. Sequential decoding techniques such as the Stack and Fano algorithms are generally nearly optimum in terms of performance and faster than Viterbi decoding, whereas majority logic decoding is very fast, but generally gives substantially worse performance.

The performance of a convolutionally encoded system is specified in terms of the decoded symbol error probability or in terms of a coding gain, as defined in section 127.1. An exact result for the error probability cannot be easily found but a good upper bound is provided through the use of transfer functions. To find the transfer function for a given code, an augmented state diagram is

produced from the original state diagram of the code by splitting the all-zero state into a starting and ending state and labeling each branch with  $W^w I^i$ —where  $w$  is the Hamming weight of the output sequence on that branch,  $i$  is the Hamming weight of the input that caused that transition, and  $W$  and  $I$  are dummy variables. The transfer function of the code is then given by applying Mason's gain formula to the augmented state diagram. The augmented state diagram and the transfer function for the encoder in Fig. 127.2(a) are shown in Figs. 127.2(d) and 127.2(e). The probability of decoded symbol error can be found from the transfer function by

$$P_e \leq \frac{1}{k} \frac{\partial}{\partial I} T(W, I) \Big|_{I=1, W=Z}$$

The parameter  $Z$  that appears in this equation is a function of the channel being used. For a binary symmetric channel (BSC),  $Z = \sqrt{4p(1-p)}$ , where  $p$  is the crossover probability of the channel; for an additive white Gaussian noise (AWGN) channel with antipodal signaling,  $Z = \exp(-E_s/N_o)$ , where  $E_s$  is the average energy per symbol and  $N_o/2$  is the two-sided power spectral density of the noise.

Although the given bound is convenient because it is general, it is often not very tight. A tighter bound can be obtained for various channel models. For example, for the AWGN channel, a tighter bound is

$$P_e \leq \frac{1}{k} \exp\left(\frac{d_{\text{free}} E_s}{N_o}\right) Q\left(\sqrt{\frac{2d_{\text{free}} E_s}{N_o}}\right) \frac{\partial}{\partial I} T(W, I) \Big|_{I=1, W=\exp(-E_s/N_o)}$$

This bound generally gives a good estimate of the true error probability, at least for high values of  $E_s/N_o$ . For complex codes (long constraint length) calculating the transfer function can be tedious. In that case it is common to use a lower bound that involves just the first term in the Taylor series expansion of the transfer function, resulting in

$$P_e \geq \frac{B_{d_{\text{free}}}}{k} Q\left(\sqrt{\frac{2d_{\text{free}} E_s}{N_o}}\right)$$

where  $B_{d_{\text{free}}}$  is the sum of the input Hamming weights of all paths with an output Hamming weight of  $d_{\text{free}}$ . In many cases  $B_{d_{\text{free}}} = 1$ , and this can always be used as a lower bound. This expression can also be used to obtain the expression for asymptotic coding gain:

$$\gamma = 10 \log \left( \frac{k d_{\text{free}}}{2n} \right) \text{ dB}$$

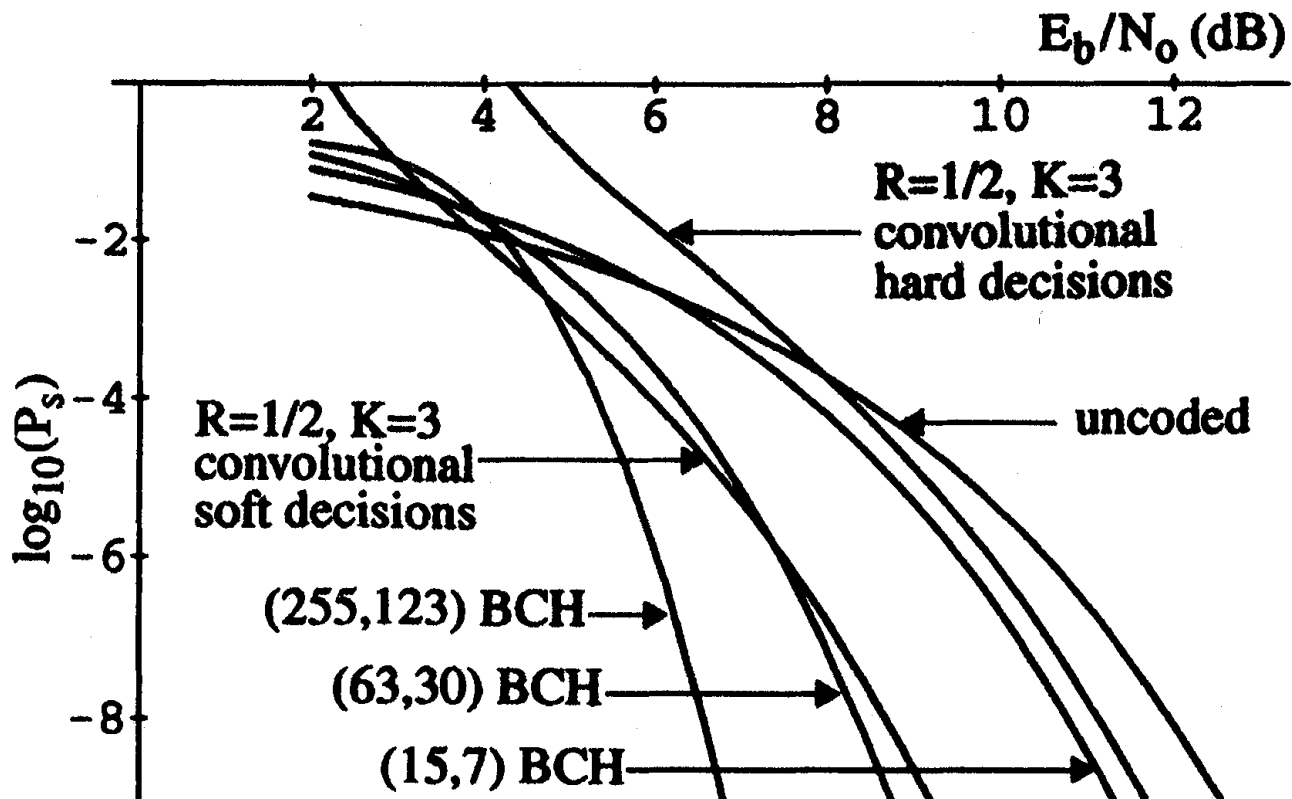
This result is very similar to the expression for asymptotic coding gain for block codes.

Finally, Fig. 127.3 shows a performance comparison of the rate  $1/2$ ,  $K = 3$  convolutional code depicted in Fig. 127.2 with several block codes whose code rates are nearly equal to  $1/2$ . Bit-error rate is plotted as a function of  $E_b/N_o$  (energy per bit divided by noise spectral density). For all



cases the assumed modulation is BPSK. Note that even this simple convolutional code with soft decisions can perform as well as a fairly long (and hence complex) block code.

**Figure 127.3** Bit-error-rate comparison of block and convolutional codes.



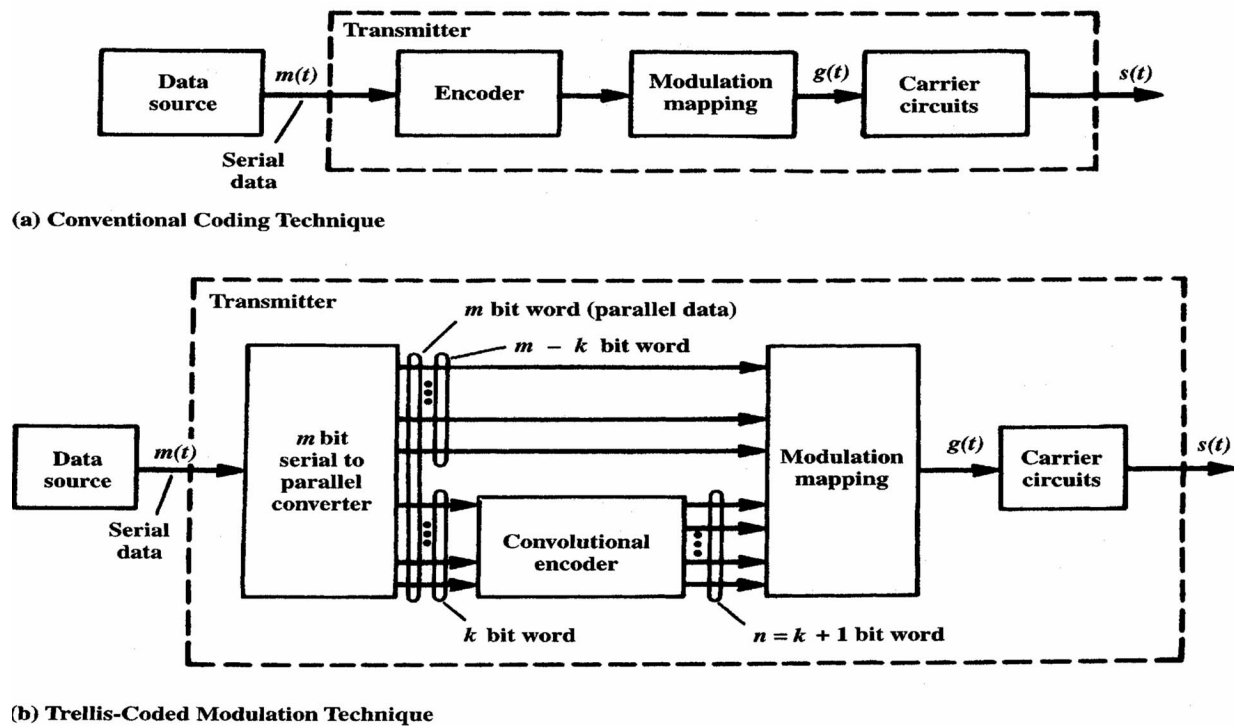
### 127.3 Trellis-Coded Modulation

In both block and convolutional coding a savings in power (coding gain) is obtained at the cost of extra bandwidth. In order to provide the redundancy, extra symbols are added to the information stream. Accommodating these symbols without sacrificing information rate requires a shorter symbol duration or, equivalently, a larger bandwidth. The bandwidth expansion relative to an uncoded system is equal to  $1/R = n/k$ . Thus, coding is used when power is at a premium but bandwidth is available—that is, a power-limited environment. When power savings is not so crucial and bandwidth is at a premium, (a bandwidth-limited environment), spectrum can be conserved by using a multilevel signaling scheme (some form of  $M$ -ary PSK or  $M$ -ary quadrature amplitude modulation is popular in this case). When both bandwidth and power are tightly constrained, trellis-coded modulation (TCM) provides power savings with no bandwidth expansion. Redundancy is added to the data stream through the expansion of the size of the signal set. For example, binary data could be sent using a QPSK signal set. Since a QPSK constellation has four points and thus could represent two bits, each QPSK signal would represent one data bit

and one parity bit (a rate  $1/2$  code). Alternatively, one could use 8-PSK with a rate  $2/3$  code and send two bits of information per symbol. As a general rule of thumb, if it is desired to send  $m$  bits of information per channel symbol, the size of the signal set is doubled and a rate  $m/(m+1)$  code is used.

A general block diagram of a TCM encoder is shown in Fig. 127.4. Data bits are grouped into  $m$  bit blocks. These blocks are then input to a rate  $m/(m+1)$  binary convolutional encoder. The  $m+1$  output bits are then used to select a signal from an  $M = 2^{m+1}$ -ary constellation. The constellation is usually restricted to being one- or two-dimensional in order to save bandwidth. The encoding is often done in a systematic manner as indicated in Fig. 127.4, where  $m-k$  of the input bits are left uncoded and the remaining  $k$  bits are encoded using a rate  $k/(k+1)$  encoder. As with convolutional codes, the TCM scheme is designed to maximize the free distance of the code, but in this case distance is measured as Euclidean distance in the signal constellation. Each path through the trellis diagram now represents a sequence of signals from the constellation, and performance is mainly determined by the minimum Euclidean distance between any two distinct sequences of signals.

**Figure 127.4** Block diagram of transmitter for TCM. (Source: Couch, L. W. 1993. *Digital and Analog Communication Systems*, 4th ed. Macmillan, New York. With permission.)



At the receiver, the received signal is demodulated using soft decisions and then decoded using the Viterbi algorithm. Performance is determined in a manner very similar to that for a convolutional code. Bounds on the bit-error rate can be found using transfer function techniques, although the method to obtain these transfer functions can be more complicated if the signal constellation does not exhibit a great deal of symmetry. Also, a coding gain relative to an uncoded system can be calculated as

$$\gamma = 20 \log \left( \frac{d_{\text{free}}}{d_{\text{uncoded}}} \right) \text{ dB}$$

**Figure 127.5** Free distance of binary convolution codes with 4-PSK modulation, and TCM with a variety of two-dimensional modulation schemes. (Source: Ungerboeck, G. 1987. Trellis-coded modulation with redundant signal sets, Part I: Introduction. *IEEE Comm. Magazine*. 25(2):5–11. ©1987 IEEE. With permission.)



# Example 10. Register

Fig. 1

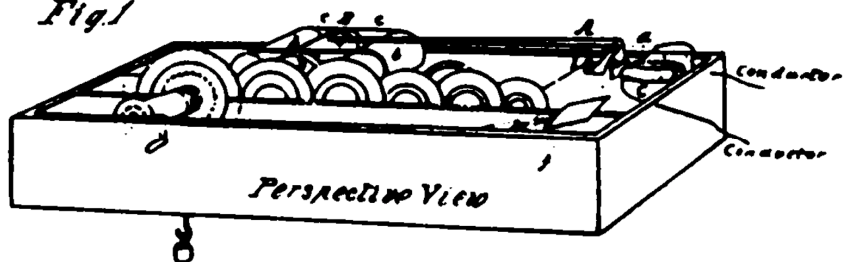


Fig. 2.

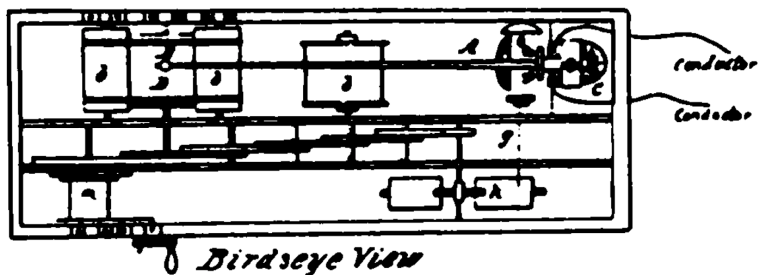


Fig. 3.

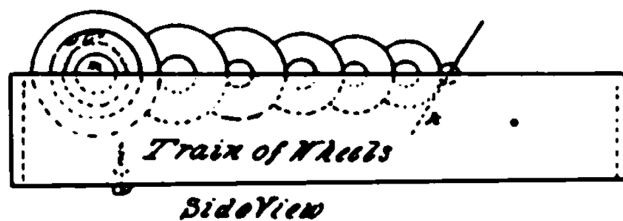
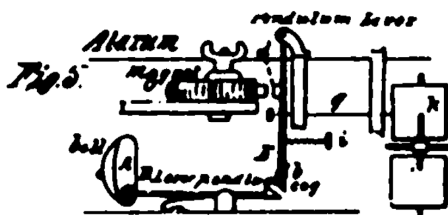
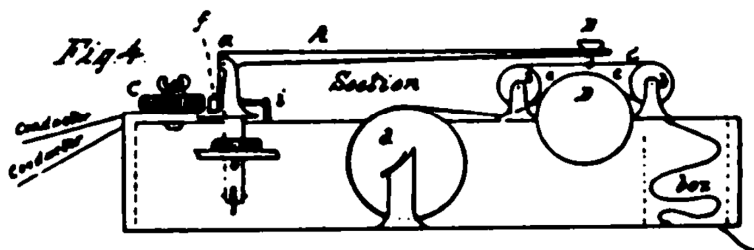


Fig. 4.



## TELEGRAPH SIGNS

*Samuel F. B. Morse*

*Patented June 20, 1840*

*#1,647*

An excerpt:

*To all whom it may concern:*

Be it known that I, the undersigned, SAMUEL F. B. MORSE, of the city, county, and State of New York, have invented a new and useful machine and system of signs for transmitting intelligence between distant points by the means of a new application and effect of electro-magnetism in producing sounds and signs, or either, and also for recording permanently by the same means, and application, and effect of electro-magnetism any sign thus produced and representing intelligence, transmitted as before named between distant points; and I denominate said invention the "American Electro-Magnetic Telegraph,"...

This was the Telegraph that opened interstate and international commerce as well as the frontier west. The compact system of dots and dashes representing letters and numerals known as the Morse Code was also developed and is still used in radio communications. (©1992, DewRay Products, Inc. Used with permission.)

## Defining Terms

**Block code:** A memoryless mapping from  $k$  input symbols to  $n$  output symbols.

**Code rate:** The ratio of the number of input symbols to the number of output symbols for a block or convolutional encoder.

**Constraint length:** The number of input blocks that affect a current output block in a convolutional encoder. It is also a measure of the complexity of the code.

**Convolutional code:** A mapping from  $k$  input symbols to  $n$  output symbols that is not memoryless. The current output depends on current and past inputs.

**Cyclic code:** A block code for which every cyclic shift of a code word produces another code word.

**Group Code:** A block code for which any linear combination of two code words produces another code word. Also known as a *linear code*.

**Trellis-coded modulation:** A combination of traditional modulation and convolutional coding techniques whereby redundancy is added through the expansion of the size of the signal constellation.

## References

Bhargava, V. K. 1983. Forward error correction schemes for digital communications. *IEEE Comm. Mag.* 21(1):11–19.

Biglieri, E., Divsalar, D., McLane, P. J., and Simon, M. K. 1991. *Introduction to Trellis-Coded Modulation with Application*. Macmillan, New York.

Couch, L. W. 1993. *Digital and Analog Communication Systems*, 4th ed. Macmillan, New York.

- Lin, S. and Costello, D. J. 1983. *Error Control Coding: Fundamentals and Applications*. Prentice Hall, Englewood Cliffs, NJ.
- Ungerboeck, G. 1987. Trellis-coded modulation with redundant signal sets. *IEEE Comm. Mag.* 25(2):5–21.

## **Further Information**

A good tutorial presentation of traditional coding techniques can be found in the article by Bhargava. For full details the book by Lin and Costello gives a good presentation of both block and convolutional coding techniques.

For information on trellis-coded modulation the reader is referred to the tutorial article by Ungerboeck, which is quite readable. The text by Biglieri *et al.* is the only one known to the authors completely devoted to TCM.

For the latest developments in the area of coding, the reader is referred to the *IEEE Transactions on Information Theory* and the *IEEE Transactions of Communications*.

Daigle, J. N. "Computer Communication Networks"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Computer Communication Networks

---

Adapted from Daigle, J. N. 1993. *The Electrical Engineering Handbook*, ed. R. C. Dorf, pp. 1447–1460. CRC Press, Boca Raton, FL.

- 128.1 General Networking Concepts
- 128.2 Computer Communication Network Architecture
- 128.3 Local-Area Networks and Internets
- 128.4 Some Additional Recent Developments

### John N. Daigle

*University of Mississippi*

The last several years have indeed been exciting times in the brief history of computer communication networks. Over this period of time we have seen the explosive growth of the Internet, a worldwide interconnection of computer communication networks that allows low-latency person-to-person communications on a global basis. So-called **firewall** technology has been developed to the point where more private companies are able to connect to the internet with significantly reduced fear of compromising their important private information. Significant steps have been taken to extend networking services to the mobile information user.

The potential for using networking technology as a vehicle for delivering multimedia—voice, data, image, and video—presentations, and, indeed, multiparty, multimedia conferencing service, is being demonstrated, and the important problems that must be solved in order to realize this potential are rapidly being defined and focused upon. And, perhaps more importantly, user-friendly applications that facilitate navigation within the **World Wide Web** have been developed and made available to networking users on a nonfee basis via network servers, thus facilitating virtually instantaneous search and retrieval of information on a global basis.

By definition, a **computer communication network** is a collection of applications hosted on separate machines and interconnected by an infrastructure that provides communications among the communicating entities. Although the applications are generally understood to be computer programs, the generic model includes the human being as an application.

This chapter summarizes the major characteristics of computer communication networks. Its objective is to provide a concise introduction that will allow the reader to gain an understanding of the key distinguishing characteristics of the major classes of networks that exist today and some of the issues involved in the introduction of emerging technologies.

There are a significant number of well-recognized books in this area. Among these are the excellent texts by Schwartz [1987] and Spragins [1991], which have enjoyed wide acceptance both by students and practicing engineers and cover most of the general aspects of computer communication networks. Other books that have been found especially useful by many practitioners are those by Rose [1990] and Black [1991].



The latest developments are, of course, covered in the current literature, conference proceedings, and the notes of standards meetings. A pedagogically oriented magazine that specializes in computer communications networks is *IEEE Network*, but *IEEE Communications* and *IEEE Computer* often contain interesting articles in this area. *ACM Communications Review*, in addition to presenting pedagogically oriented articles, often presents very useful summaries of the latest standards activities. Major conferences that specialize in computer communications include the IEEE INFOCOM and ACM SIGCOMM series, which are held annually. It is becoming common at this time to have more and more discussion about personal communication systems, and the mobility issues involved in communication networks are often discussed in *IEEE Network* and a new magazine, *IEEE Personal Communication Systems*.

We begin our discussion with a brief statement of how computer networking came about and a capsule description of the networks that resulted from the early efforts. Networks of this generic class, called **wide-area networks** (WANs) are broadly deployed today, and there are still a large number of unanswered questions with respect to their design. The issues involved in the design of those networks are basic to the design of most networks, whether wide area or otherwise. In the process of introducing these early systems, we describe and contrast three basic types of communication switching: circuit, message, and packet.

We next turn to a discussion of computer communication **architecture**, which describes the structure of communication-oriented processing software within a communication-processing system. Our discussion is limited to the **International Standards Organization/Open Systems Interconnection (ISO/OSI) reference model** (ISORM) because it provides a framework for discussion of some of the modern developments in communications in general and communication networking in particular. This discussion is necessarily simplified in the extreme, thorough coverage requiring on the order of several hundred pages, but we hope our brief description will enable the reader to appreciate some of the issues.

Having introduced the basic architectural structure of communication networks, we next turn to a discussion of an important variation on this architectural scheme: the **local-area network** (LAN). Discussion of this topic is important because it helps to illustrate what the reference model is and what it is not. In particular, the architecture of LANs illustrates how the ISO/OSI reference model can be adapted for specialized purposes. Specifically, early network architectures anticipate networks in which individual node pairs are interconnected via a single link, and connections through the network are formed by concatenating node-to-node connections.

LAN architectures, on the other hand, anticipate all nodes being interconnected in some fashion over the same communication link (or medium). This, then, introduces the concept of adaption layers in a natural way. It also illustrates that, if the services provided by an architectural layer are carefully defined, the services can be used to implement virtually any service desired by the user, possibly at the price of some inefficiency.

We conclude with a brief discussion of the status of two recent developments in communication networking: frame relay and asynchronous transfer mode (ATM) technology, which is a part of the larger **broadband integrated services digital network** (BISDN) effort. These technologies are likely to be important building blocks for the computer communication networks of the future.

## 128.1 General Networking Concepts

---

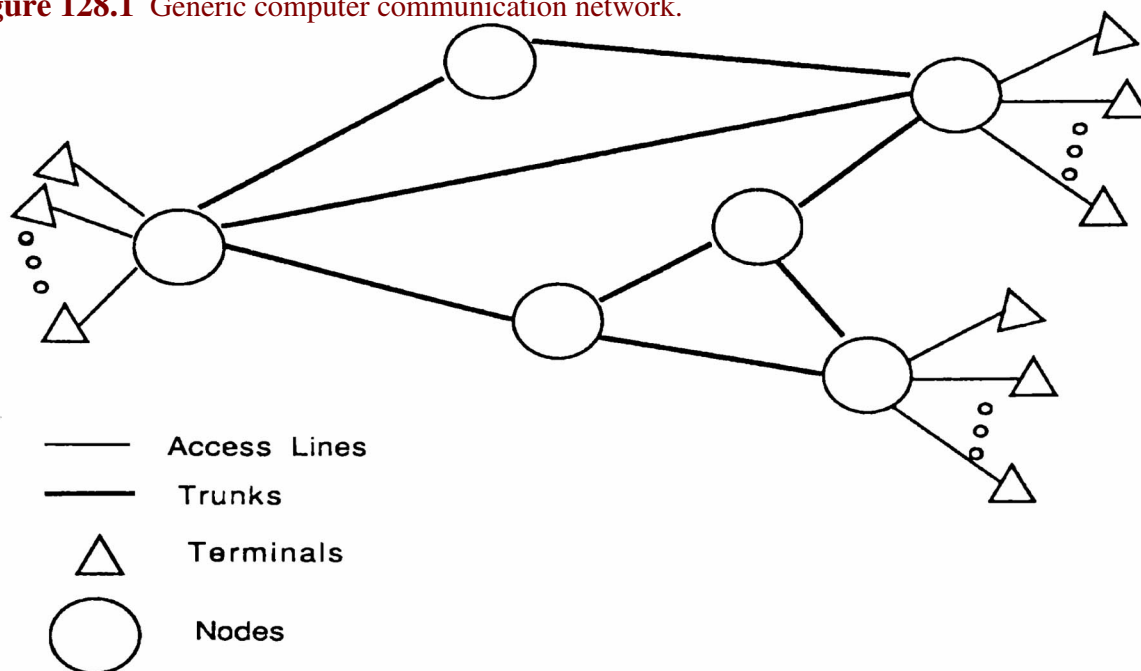
Data communication networks have existed since about 1950. The early networks existed primarily for the purpose of connecting users of a large computer to the computer itself, with additional capability to provide communications between computers of the same variety and

having the same operating software. The lessons learned during the first 20 or so years of operation of these types of networks have been valuable in preparing the way for modern networks. For the purposes of our current discussion, however, we think of communication networks as being networks whose purpose is to interconnect a set of applications that are implemented on hosts manufactured by possibly different vendors and managed by a variety of operating systems. Networking capability is provided by software systems that implement standardized interfaces specifically designed for the exchange of information among heterogeneous computers.

The earliest effort to develop large-scale, general purpose networking capability based on packet switching was led by the Advanced Research Projects Agency (ARPA) of the Department of the Army in the late 1960s; this effort resulted in the computer communication network called the ARPANET. The end results of the ARPA networking effort, its derivatives, and the early initiatives of many companies such as AT&T, DATAPOINT, DEC, IBM, and NCR have been far reaching in the extreme. We will concentrate on the most visible product of these efforts, which is a collection of programs that allows applications running in different computers to intercommunicate. Before turning to our discussion of the software, however, we shall provide a brief description of a generic computer communication network.

Figure 128.1 shows a diagram of a generic computer communication network. The most visible components of the network are the *terminals*, the **access lines**, the **trunks**, and the **switching nodes**. Work is accomplished when the users of the network, the terminals, exchange messages over the network.

**Figure 128.1** Generic computer communication network.



The terminals represent the set of communication-terminating equipment communicating over the network. Equipment in this class includes, but is not limited to, user terminals, general purpose computers, and database systems. This equipment, either through software or through human interaction, provides the functions required for information exchange between pairs of application programs or between application programs and people. The functions include, but are not limited to, call setup, session management, and message transmission control. Examples of applications include electronic mail transfer, terminal-to-computer connection for time sharing or other purposes, and terminal-to-database connections.

Access lines provide for data transmission between the terminals and the network switching nodes. These connections may be set up on a permanent basis or they may be switched connections, and there are numerous transmission schemes and protocols available to manage these connections. The essence of these connections, however, from our point of view, is a channel that provides data transmission at some number of bits per second (bps), called the *channel capacity*,  $C$ . The access line capacities may range from a few hundred bps to in excess of millions of bps, and they are usually not the same for all terminating equipments of a given network. The actual information-carrying capacity of the link depends upon the protocols employed to effect the transfer; the interested reader is referred to Bertsekas and Gallager [1987], especially chapter 2, for a general discussion of the issues involved in transmission of data over communication links.

Trunks, or internodal trunks, are the transmission facilities that provide for transmission of data between pairs of communication switches. These are analogous to access lines and, from our point of view, they simply provide a communication path at some capacity, specified in bps.

There are three basic switching paradigms: circuit, message, and packet switching. **Circuit switching** and **packet switching** are transmission technologies while **message switching** is a service technology. In circuit switching a call connection between two terminating equipments corresponds to the allocation of a prescribed set of physical facilities that provide a transmission path of a certain bandwidth or transmission capacity. These facilities are dedicated to the users for the duration of the call. The primary performance issues, other than those related to quality of transmission, are related to whether or not a transmission path is available at call setup time and how calls are handled if facilities are not available.

**Message switching** is similar in concept to the postal system. When a user wants to send a message to one or more recipients, the user forms the message and addresses it. The message-switching system reads the address and forwards the complete message to the next switch in the path. The message moves asynchronously through the network on a message switch-to-message switch basis until it reaches its destination. Message-switching systems offer services such as mailboxes, multiple-destination delivery, automatic verification of message delivery, and bulletin boards. Communication links between the message switches may be established using circuit or packet-switching networks, as is the case with most other networking applications. Examples of message-switching protocols that have been used to build message-switching systems are Simple Mail Transfer Protocol (SMTP) and the International Telegraph and Telephone Consultative Committee (CCITT) X.400 series. The former is much more widely deployed, whereas the latter has significantly broader capabilities, but its deployment is plagued by having two incompatible versions (1984 and 1988) and other problems. Many commercial vendors offer message-switching services based either on one of the above protocols or a proprietary protocol.

In the circuit-switching case, there is a one-to-one correspondence between the number of trunks between nodes and the number of simultaneous calls that can be carried. That is, a trunk is a facility between two switches that can service exactly one call, and it does not matter how this transmission facility is derived. Major design issues include the specification of the number of trunks between node pairs and the routing strategy used to determine the path through a network in order to achieve a given call-blocking probability. When blocked calls are queued, the number of calls that may be queued is also a design question.

A packet-switched communication system exchanges messages between users by transmitting sequences of packets comprising the messages. That is, the sending terminal equipment partitions a message into a sequence of packets, the packets are transmitted across the network, and the receiving terminal equipment reassembles the packets into messages. The transmission facility interconnecting a given node pair is viewed as a single trunk, and the transmission capacity of this trunk is shared among all users whose packets traverse both nodes. Whereas the trunk capacity is

specified in bps, the packet-handling capacity of a node pair depends on both the trunk capacity and the nodal processing power.

In many packet-switched networks the path traversed by a packet through the network is established during a call setup procedure, and the network is referred to as a *virtual circuit packet-switching network*. Other networks provide datagram service, a service that allows users to transmit individually addressed packets without the need for call setup. Datagram networks have the advantage of not having to establish connections before communications take place, but have the disadvantage that every packet must contain complete addressing information. Virtual circuit networks have the advantage that addressing information is not required in each packet, but have the disadvantage that a call setup must take place before communications can occur. Datagram is an example of **connectionless service**, whereas virtual circuit is an example of **connection-oriented service**.

Prior to the late 1970s signaling for circuit establishment was in-band. That is, in order to set up a call through the network, the call setup information was sent sequentially from switch to switch using the actual circuit that would eventually become the circuit used to connect the end users. In an extreme case this process amounted to trying to find a path through a maze, sometimes having to retrace steps before finally emerging at the destination or simply giving up when no path could be found. This system had two negative characteristics: First, the rate of signaling information transfer was limited to the circuit speed, and, second, the circuits that could have been used for accomplishing the end objective were being consumed simply to find a path between the end points. These limitations resulted in tremendous bottlenecks on major holidays, which were solved by virtually disallowing alternate routes through the toll switching network.

An alternate out-of-band signaling system, usually called **common channel interoffice signaling** (CCIS), was developed primarily to solve this problem. Signaling now takes place over a signaling network that is partitioned from the network that carries the user traffic. This principle is incorporated into the concept of integrated services digital networks (ISDNs), which is described thoroughly in Helgert [1991]. The basic idea of ISDN is to offer to the user some number of 64 kbps access lines, plus a 16 kbps access line through which the user can describe to an ISDN how the user wishes to use each of the 64 kbps circuits at any given time. The channels formed by concatenating the access lines with the network interswitch trunks having the requested characteristics are established using an out-of-band signaling system, the most modern of which is Signaling System #7 (SS#7).

In either virtual circuit or *datagram networks*, packets from a large number of users may simultaneously need transmission services between nodes. Packets arrive at a given node at random times. The switching node determines the next node in the transmission path and then places the packet in a queue for transmission over a trunk facility to the next node. Packet arrival processes tend to be bursty—that is, the number of packet arrivals over fixed-length intervals of time has a large variance. Because of the burstiness of the arrival process, packets may experience significant delays at the trunks. Queues may also build due to the difference in transmission capacities of the various trunks and access lines, and delays result. Processing is also a source of delay; the essence of packet-switching technology is to trade delay for efficiency in resource utilization.

Protocol design efforts, which seek to improve network efficiencies and application performance, are frequent topics of discussion at both general conferences in communications and those specialized to networking. The reader is encouraged to consult the proceedings of the conferences mentioned earlier for a better appreciation of the range of issues and the diversity of the proposed solutions to the issues.

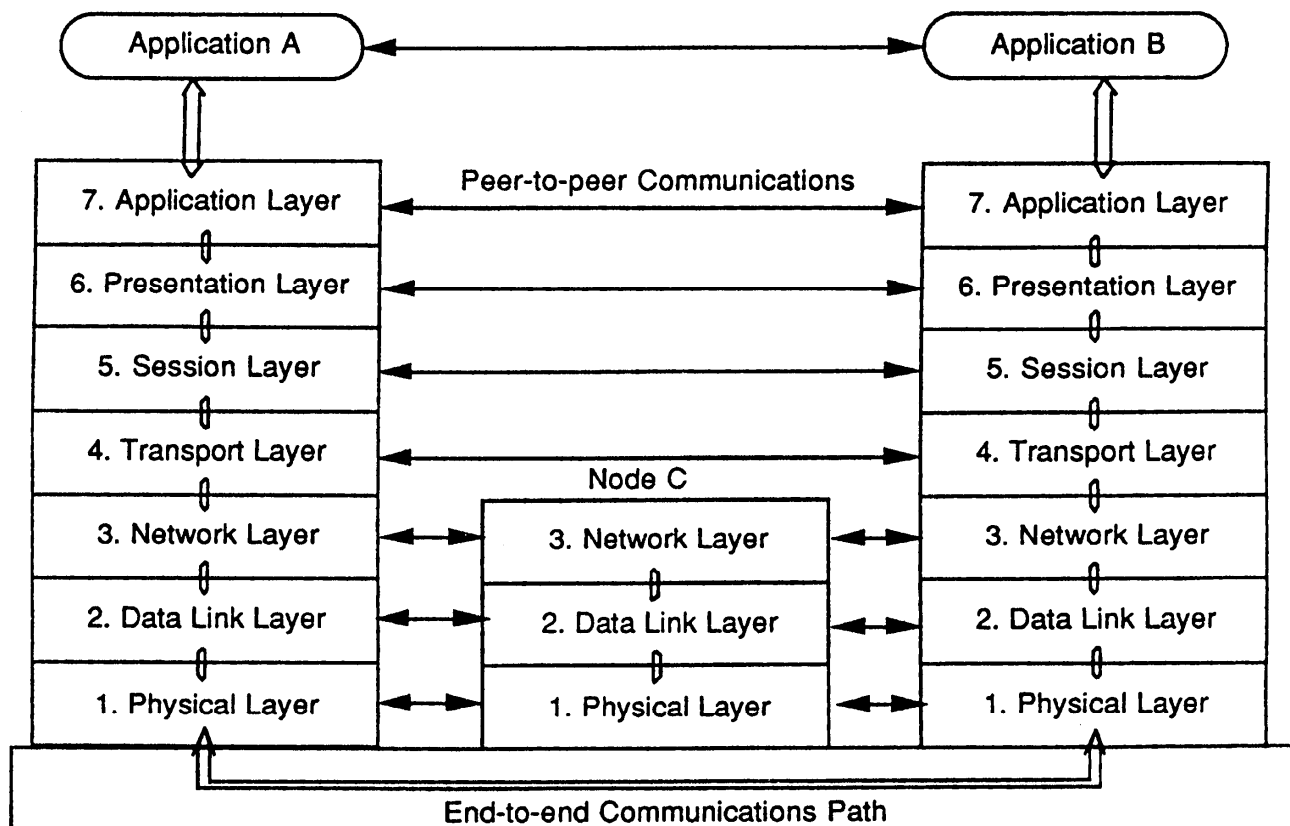
## 128.2 Computer Communication Network Architecture

In this section we begin with a brief, high-level definition of the ISORM, which is discussed in significant detail in Black [1991]. There is significant debate over whether the efforts of the ISO/OSI community are leading to the best standards, but we choose to base our discussion on the ISORM because it is very useful for discussing network architecture principles, and these principles apply across the board.

The reference model has seven layers, none of which can be bypassed conceptually. In general, a layer is defined by the types of services it provides to its users and the quality of those services. For each layer in the ISO/OSI architecture the user of a layer is the next layer up in the hierarchy—except for the highest layer, for which the user is an application. Clearly, when a layered architecture is implemented under this philosophy, the quality or service obtained by the end user, the application, is a function of the quality of service provided by all of the layers.

Figure 128.2, adapted from Spragins [1991], shows the basic structure of the OSI architecture and how this architecture is envisaged to provide for exchange of information between applications. As shown in the figure, there are seven layers: application, presentation, session, transport, network, data link, and physical. Brief definitions of the layers are now given, but the reader should bear in mind that substantial further study will be required to develop an understanding of the practical implications of the definitions.

**Figure 128.2** Layered architecture for ISO/OSI reference model. (Adapted from Spragins, J. D. 1991. *Telecommunications: Protocols and Design*. Addison-Wesley, Reading, MA.)



*Physical layer.* Provides electrical, functional, and procedural characteristics to activate, maintain, and deactivate physical data links that transparently pass the bit stream for communication between data link entities.

*Data link layer.* Provides functional and procedural means to transfer data between network entities. Provides for activation, maintenance, and deactivation of data link connections; character and frame synchronization; grouping of bits into characters and frames; error control, media access control; and flow control.

*Network layer.* Provides switching and routing functions to establish, maintain, and terminate network layer connections and transfer data between transport layers.

*Transport layer.* Provides host-to-host, cost-effective, transparent transfer of data; end-to-end flow control; and end-to-end quality of service as required by applications.

*Session layer.* Provides mechanisms for organizing and structuring dialogues between application processes.

*Presentation layer.* Provides for independent data representation and syntax selection by each communicating application and for conversion between selected contexts and the internal architecture standard.

*Application layer.* Provide applications with access to the ISO/OSI communication stack and certain distributed information services.

As mentioned previously, a layer is defined by the types of services it provides to its users. In the

case of a request or a response, these services are provided via invocation of **service primitives** of the layer in question by the layer that wants the service performed. In the case of an indication or a confirm, these services are provided via invocation of service primitives of the layer in question by the same layer that wants the service performed.

This process is not unlike a user of a programming system calling a subroutine from a scientific subroutine package in order to obtain a service, for instance, matrix inversion or memory allocation. For example, a request is analogous to a CALL statement in a FORTRAN program, and a response is analogous to the RETURN statement in the subroutine that has been CALLED. The requests for services are generated asynchronously by all of the users of all of the services, and these join (typically prioritized) queues along with other requests and responses while awaiting servicing by the processor or other resource such as a transmission line.

The service primitives fall into four basic types, which are as follows: request, indication, response, and confirm. These types are defined as follows:

*Request.* A primitive sent by layer  $(N + 1)$  to layer  $N$  to request a service.

*Indication.* A primitive sent by layer  $N$  to layer  $(N + 1)$  to indicate that a service has been requested of layer  $N$  by a separate layer  $(N + 1)$  entity.

*Response.* A primitive sent by  $(N + 1)$  to layer  $N$  in response to an *indication primitive*.

*Confirm.* A primitive sent by layer  $N$  to layer  $(N + 1)$  to indicate that a response to an earlier *request* primitive has been received.

In order to be more specific about how communication takes place, we now turn to a brief discussion of layer 2, the data link layer. The primitives provided by the ISO data link layer are as follows [Stallings, 1990a].

- DL\_CONNECT.request
- DL\_RESET.request
- DL\_CONNECT.indication
- DL\_RESET.indication
- DL\_CONNECT.response
- DL\_RESET.response
- DL\_CONNECT.confirm
- DL\_RESET.confirm
- DL\_DATA.request
- DL\_DISCONNECT.request
- DL\_DATA.indication
- DL\_DISCONNECT.indication
- DL\_DATA.response
- DL\_UNITDATA.request
- DL\_DATA.confirm
- DL\_UNITDATA.indication

Each *primitive* has a set of **formal parameters**, which are analogous to the formal parameters of a procedure in a programming language. For example, the parameters for the DLCONNECT.request primitive are the called address, the calling address, and the quality-of-service parameter set. These three parameters are used in the establishment of data link



(DL) connections. The called address and the calling address are analogous to the telephone numbers of two parties of a telephone call, whereas the quality-of-service parameter set allows for the negotiation of various agreements, such as throughput (measured in bits per second).

All four DLCONNECT primitives are used to establish a data link. An analogy to an ordinary phone call is now drawn so that the basic idea of the primitives can be better appreciated. DLCONNECT.request is equivalent to picking up the phone and dialing. The phone ringing at the called party's end is represented by DLCONNECT.indication. DLCONNECT.response is equivalent to the called party lifting the receiver and answering, and DLCONNECT.confirm is equivalent to the calling party hearing the response of the called party.

In general, communications take place between peer layer protocols by the exchange of **protocol data units** (PDUs), which contain all of the information required for the receiving protocol **entity** to provide the required service. In order to exchange PDUs entities at a given layer use the services of the next lower layer. The data link primitives listed include both connection-mode primitives and connectionless mode primitives. For connection mode communications a connection must be established between two peer entities before they can exchange PDUs.

For example, suppose a network layer entity in host A wishes to be connected to a network layer entity in host B, as shown in [Fig. 128.2](#). Then the connection would be accomplished by the concatenation of two data link connections: one between A and C, and one between C and B. In order to establish the connection, the network layer entity in host A would issue a DLCONNECT.request to its associated data link entity, providing the required parameters. This data link entity would then transmit this request to a data link entity in C, which would issue a DLCONNECT.indication to a network entity in C. The network entity in C would then analyze the parameters of the DLCONNECT.indication and realize that the target destination is B. This network layer entity would then reissue the DLCONNECT.request to its data link entity, which would transmit the request to a data link entity in B. The data link entity in B would send a DLCONNECT.indication to a network layer entity in B, and this entity would issue a DLCONNECT.response back to the data link entity in B. This DLCONNECT.response would be relayed back to the data link entity in A following the same sequence of events as in the forward paths. Eventually, this DLCONNECT.response would be converted to a DLCONNECT.confirm by the data link entity in A and passed to the network entity in A, thus completing the connection.

Once the connection is established, data exchange between the two network layer entities can take place; that is, the entities can exchange PDUs. For example, if a network layer entity in host A wishes to send a PDU to a network layer entity in host B, the network layer entity in host A would issue a DLDATA.request to the appropriate data link layer entity in host A. This entity would package the PDU together with appropriate control information into a data link service data unit (DLSDU) and send it to its peer at C. The peer at C would deliver it to the network entity at C, which would forward it to the data link entity in C providing the connection to host B. This entity would then send the DLSDU to its peer in host B, and this data link entity would pass the PDU to host B network entity via a DLDATA.indication. Now, network layer PDUs are called *packets* and DL layer PDUs are called *frames*. But, the data link layer does not know that the information it is transmitting is a packet; to the DL layer entity, the packet is simply user information. From the perspective of a data link entity, it is not necessary to have a network layer. The network layer exists to add value for the user of the network layer to the services provided by the DL layer. In the



example given, value was added by the network layer by providing a relaying capability since hosts A and C were not directly connected. Similarly, the DL layer functions on a hop-by-hop basis, each hop being completely unaware that there are any other hops involved in the communication. We will see later that the data link need not be limited to a single physical connection.

The philosophy of the ISO/OSI architecture is that, in addition to the software being layered, implementations are not allowed to bypass entire layers; that is, every layer must appear in the implementation. This approach was developed after the approach defined for the ARPANET project, which is hierarchical, was fully developed. In the hierarchical approach the layer interfaces are carefully designed, but any number of layers of software can be bypassed by any application (or other higher-layer protocol) that provides the appropriate functionality. These two approaches have been hotly debated for a number of years, but as the years pass, the approaches are actually beginning to look more and more alike for a variety of reasons that will not be discussed here.

The ISO/OSI layered architecture described would appear to be very rigid, not allowing for any variations in underlying topology or variations in link reliability. However, as we shall see, this is not necessarily the case. For example, ISO 8348, which developed as a result of the X.25 project, provides only connection-oriented service, and it was originally intended as the only network layer standard for ISO/OSI. However, ISO 8473, or ISO-IP—which is virtually identical to the Department of Defense (DoD) internet protocol (DoD-IP) developed in the ARPANET project—has since been added to the protocol suite to provide connectionless service as well as internet service.

The ISO/OSI protocol suite is in a constant state of revision as new experience reveals the need for additional capabilities and flexibility. Some of this additional flexibility and functionality is being provided through the use of so-called **adaption sublayers**, which enhance the capabilities of a given layer so that it can use the services of a lower layer with which it was not specifically designed for compatibility.

Interestingly, the use of adaption sublayers is only a short step away from using adaption layers that would allow applications to directly interface with any ISO layer. This would result in a hierarchical rather than layered architecture; to wit, ISORM becomes DoDRM. Meanwhile, the severe addressing limitations of DoD-IP—together with the emergence of high-performance multimedia networking applications, which require service quality guarantees—have led to a significant effort in the internet community to rethink the service requirements and implementation features of the internetworking layer. Many features of the ISO protocols have been considered for incorporation into the next generation of IP. Fundamental changes in the national (and worldwide) communications infrastructure appear to be leading naturally in the hierarchical direction, and this fact has been recognized by the U.S. government, which has recently relaxed its requirements for compliance with ISO standards in networking products.

We now turn to a discussion of LANs, which have inherent properties that make the use of sublayers particularly attractive.

## 128.3 Local-Area Networks and Internets

---

In this section we discuss the organization of communications software for LANs. In addition, we

introduce the idea of **internets**, which were brought about to a large extent by the advent of LANs. We discuss the types of networks only briefly and refer the reader to the many excellent texts on the subject. Layers 4 and above for local-area communications networks are identical to those of wide-area networks. However, because the hosts communicating over an LAN share a single physical transmission facility, the routing functions provided by the network layer, layer 3, are not necessary. Thus, the functionality of a layer 3 within a single LAN can be substantially simplified without loss of utility. On the other hand, a DL layer entity must now manage many simultaneous DL layer connections because all connections entering and leaving a host on a single LAN do so over a single physical link. Thus, in the case of connection-oriented communications, the software must manage several virtual connections over a single physical link.

There were several basic types of transmission schemes in use in early local-area networks. Three of these received serious consideration for standardization: the **token ring**, **token bus**, and **carrier sense multiple access** (CSMA). All three of these access methods became IEEE standards (IEEE 802) and eventually became ISO standards (ISO 8802 series) because all merited standardization. On the other hand, all existed for the express purposes of exchanging information among peers, and it was recognized at the outset that the upper end of the data link layer could be shared by all three access techniques. The decision to use a common logical link control (LLC) sublayer for all of the LAN protocols and develop a separate sublayer for each of the different media apparently ushered in the idea of adaption sublayers, in this case, the **media access control** (MAC) sublayer. This idea has proven to be valuable as new types of technologies have become available. For example, the new fiber-distributed digital interface (FDDI) uses the LLC of all other LAN protocols, but its MAC is completely different from the token ring MAC even though FDDI is a token ring protocol. A thorough discussion of FDDI and related technologies is given in Jain [1994].

One of the more interesting consequences of the advent of local-area networking is that many traditional computer communication networks became internets overnight. LAN technology was used to connect stations to a host computer, and these host computers were already on a WAN. It was then a simple matter to provide a relaying, or bridging, service at the host in order to provide wide-area interconnection of station to LANs to each other. In short, the previously established WANs became networks for interconnection of LANs; that is, they were interconnecting networks rather than stations. Internet performance suddenly became a primary concern in the design of networks; a new business developed vending specialized equipment, routers, to provide interconnection among networks.

Metropolitan-area networks (MANs) have been deployed for the interconnection of LANs within a metropolitan area. The primary media configuration for MANs is a dual-bus configuration, and it is implemented via the distributed queue, dual-bus (DQDB) protocol, also known as IEEE 802.6. The net effect of this protocol is to use the dual-bus configuration to provide service approaching the FCFS service discipline to the traffic entering the FDDI network, which is remarkable considering that the LANs being interconnected are geographically dispersed. Interestingly, DQDB concepts have recently also been adapted to provide wide-area communications. Specifically, structures have been defined for transmitting DQDB frames over standard DS-1 (1.544 megabits per second) and DS-3 (6.312 megabits per second) facilities, and these have been used as the basis for a service offering called *switched multimegabit data*

services (SMDS).

Over the last few years, wireless LANs—which are local-area networks in which radio or photonic links serve as cable replacements—have been marketed. Wireless LAN technology is viewed by many as crucial to the evolution of personal communication networks, but deployment has been far less than anticipated up to this time. Part of the reason for this fact is the continued improvement in Ethernet innovations, such as Ethernet bridges, which continue to add value to the installed base. Thus, in order to achieve a market niche, added values of wireless LANs must continue to be improved. This is, of course, typical in product life cycle.

## 128.4 Some Additional Recent Developments

---

In this section we describe two recent developments of significant interest in communication networking: *frame relay* (FR) technology, which is described in Braun [1994], and **asynchronous transfer mode** (ATM) technology, which is described in McDysan and Spohn [1994]. We also comment briefly on personal communication services.

As mentioned previously, there is really no requirement that the physical media between two adjacent data link layers be composed of a single link. In fact, if a path through the network is initially established between two data link entities, there is no reason that DLC protocols need to be executed at intermediate nodes. Through the introduction of adaption layers and an elementary routing layer at the top of the DL layer, DLC frames can be relayed across the physical links of the connection without performing the error checking, flow control, and retransmission functions of the DLC layer on a link-by-link basis. The motivation is that, since link transmission is becoming more reliable, extensive error checking and flow control is not needed across individual links; an end-to-end check should be sufficient. Meanwhile, the savings in processing due to not processing at the network layer can be applied to frame processing, which allows interconnection of the switches at higher line speeds. Since bit-per-second costs decrease with increased line speed, service providers can offer savings to their customers through FRNs. Significant issues are frame loss probability and retransmission delay. Such factors will determine the retransmission strategy deployed in the network. The extensive deployment of FR technology at this time suggests that this technology provides improvements over standard packet technology.

Another recent innovation is the ATM, usually associated with BISDN. The idea of ATM is to partition a user's data into many small segments, called cells, for transmission over the network. Independent of the data's origin, the cell size is 53 octets, of which five octets are for use by the network itself for routing and error control. Users of the ATM are responsible for segmentation and reassembly of their data. Any control information required for this purpose must be included in the 48 octets of user information in each cell. In the usual case these cells would be transmitted over networks that would provide users with 135 Mbps and above data transmission capacity (with user overhead included in the capacity).

The segmentation of units of data into cells introduces tremendous flexibility for handling various types of information—such as voice, data, image, and video—over a single transmission facility. As a result, there has been a tremendous investment in developing implementation agreements that will enable a large number of vendors to independently develop interoperable equipment. This effort is focused primarily in the ATM Forum, a private, not-for-profit consortium

of over 500 companies, of which more than 150 are principal members and active contributors.

LANs, WANs, and MANs based on the ATM paradigm are being designed and, indeed, deployed. A significant portion of the early deployment activity was part of a national test bed program under joint sponsorship of the National Science Foundation (NSF) and the Advanced Research Projects Agency (ARPA). But significant quantities of ATM equipment are already deployed in various private test bed efforts and, indeed, in various commercial networks. For example, ATM is an excellent technology for providing worldwide Ethernet LAN interconnection at the full rate of ten megabits per second. Not surprisingly, numerous vendors are planning to have ATM capabilities at the back plane in much the same way that Ethernet is provided today.

There are numerous possibilities for connection of hosts to ATM networks, but they all share a common architecture, which consists of three sublayers: the ATM adaption layer (AAL), the ATM layer, and the physical media-dependent (PMD) layer. It is convenient, for the purposes of this discussion, to think of services as falling into two categories: circuit mode and packet mode, where a circuit mode service, such as voice, is a service that is naturally implemented over a circuit-switched facility, and a packet-mode service, such as E-mail, is a service that is more naturally implemented over a packet-switched connection. From many perspectives it is natural to implement circuit-mode services directly over ATM, whereas it is more natural to implement packet-mode services at the internet (or packet) layer.

The implication of this partitioning of service types is that any service that has been developed for deployment over an IP network could naturally be deployed over an ATM network by simply using the ATM network as a packet delivery network. Each packet would traverse the network as a sequence of cells over an end-to-end virtual connection. If the flow control and resource management procedures can be worked out, the net effect of this deployment strategy would be, for example, that an application designed to be deployed over an Ethernet segment could be deployed on a nationwide (or even global) network without a noticeable performance degradation. The implications of this type of capability are obvious as well as mind-boggling.

At the present time, end-to-end connections at the ATM level are expected to be connection oriented. As cells traverse the network, they are switched on a one-by-one basis, using information contained in the five ATM overhead octets to follow the virtual path established during the ATM call setup. Typically, cells outbound on a common link are statistically multiplexed, and, if buffers are full, cells are dropped. In addition, if one or more errors are found in a cell, then the cell is dropped.

In the case of data transmission a lost cell will result in an unusable frame unless the data are encoded to guard against cell loss prior to transmission. Coding might be provided by the AAL, for example. The trade-offs involved in coding and retransmission and their impact upon network throughput, delay, and complexity are not well understood at the time of this writing. Part of the reason for this uncertainty is that cell loss probability and the types of traffic that are likely to use the network are not thoroughly understood at this time. Resolution of these issues accounts for a significant portion of the research activity in computer communication networking at this time. The relevant ANSI and CCITT documents are frequently updated to include the results.

The accomplishment of deploying ATM technology would be unexciting in terms of reaching the potential pointed out in the previous paragraph if application development for communication networking were at a standstill. Fortunately, this is far from true. In addition, the last few years

have witnessed a tremendous increase in attention to the problem of providing access to networking service to the mobile user, and this has resulted in significant advances both in wireless data-networking technology and in telecommunication network control infrastructure. Computer communication networking, pronounced dead about ten years ago, has never been more alive than it is today.

## Defining Terms

**Access line:** A communication line that connects a user's terminal equipment to a switching node.

**Adaption sublayer:** Software that is added between two protocol layers to allow the upper layer to take advantage of the services offered by the lower layer in situations where the upper layer is not specifically designed to interface directly to the lower layer.

**Architecture:** The set of protocols defining a computer communication network.

**Asynchronous transfer mode (ATM):** A mode of communication in which communication takes place through the exchange of tiny units of information called *cells*.

**Broadband integrated services digital network (B-ISDN):** A generic term that generally refers to the future network infrastructure that will provide ubiquitous availability of integrated voice, data, imagery, and video services.

**Carrier sense multiple access:** A random access method of sharing a bus-type communications medium in which a potential user of the medium listens before beginning to transmit.

**Circuit switching:** A method of communication in which a physical circuit is established between two terminating equipments before communication begins to take place. This is analogous to an ordinary phone call.

**Common channel interoffice signaling:** Use of a special network, dedicated to signaling, to establish a path through a communication network, which is dedicated to the transfer of user information.

**Computer communication network:** Collection of applications hosted on different machines and interconnected by an infrastructure that provides intercommunications.

**Connectionless service:** A mode of packet switching in which packets are exchanged without first establishing a connection. Conceptually, this is very close to message switching, except that if the destination node is not active, then the packet is lost.

**Connection-oriented service:** A mode of packet switching in which a call is established prior to any information exchange taking place. This is analogous to an ordinary phone call, except that no physical resources need be allocated.

**Entity:** A software process that implements a part of a protocol in a computer communication network.

**Fast packet networks:** Networks in which packets are transferred by switching at the frame layer rather than the packet layer. Such networks are sometimes called *frame relay networks*. At this time it is becoming popular to think of frame relay as a service, rather than transmission, technology.

**Firewall:** Computer communication network hardware and software introduced into an internet at the boundary of a public network and a private network for the purpose of protecting the confidential information and network reliability of the private network.

**Formal parameters:** The parameters passed during the invocation of a service primitive; similar to the arguments passed in a subroutine call in a computer program.

**International Standards Organization reference model:** A model, established by ISO, that organizes the functions required by a complete communication network into seven layers.

**Internet:** A network formed by the interconnection of networks.

**Local-area networks:** A computer communication network spanning a limited geographic area, such as a building or college campus.

**Media access control:** A sublayer of the link layer protocol whose implementation is specific to the type of physical medium over which communication takes place and controls access to that medium.

**Message switching:** A service-oriented class of communication in which messages are exchanged among terminating equipments by traversing a set of switching nodes in a store-and-forward manner. This is analogous to an ordinary postal system. The destination terminal need not be active at the same time as the originator for the message exchange to take place.

**Metropolitan-area networks:** A computer communication network spanning a limited geographic area, such as a city; sometimes features interconnection of LANs.

**Packet switching:** A method of communication in which messages are exchanged between terminating equipments via the exchange of a sequence of fragments of the message called *packets*.

**Protocol data unit:** The unit of exchange of protocol information between entities. Typically, a protocol data unit (PDU) is analogous to a structure in C or a record in Pascal; the protocol is executed by processing a sequence of PDUs.

**Service primitive:** The name of a procedure that provides a service; similar to the name of a subroutine or procedure in a scientific subroutine library.

**Switching node:** A computer or computing equipment that provides access to networking services.

**Token bus:** A method of sharing a bus-type communications medium that uses a token to schedule access to the medium. When a particular station has completed its use of the token, it broadcasts the token on the bus, and the station to which the token is addressed takes control of the medium.

**Token ring:** A method of sharing a ring-type communications medium that uses a token to schedule access to the medium. When a particular station has completed its use of the token, it transmits the token on the bus, and the station that is physically next on the ring takes control.

**Trunk:** A communication line between two switching nodes.

**Wide-area network:** A computer communication network spanning a broad geographic area, such as a state or country.

**World Wide Web:** A collection of hypertext-style servers interconnected via internet services.

## References

Bertsekas, D. and Gallager, R. 1987. *Data Networks*, 2nd. Ed. Prentice Hall, Englewood Cliffs, NJ.

- Black, U. D. 1991. *OSI: A Model for Computer Communication Standards*. Prentice Hall, Englewood Cliffs, NJ.
- Braun, E. 1994. *The Internet Directory*. Fawcett Columbine, New York.
- Hammond, J. L. and O'Reilly, P. J. P. 1986. *Performance Analysis of Local Computer Networks*. Addison-Wesley, Reading, PA.
- Helgert, H. J. 1991. *Integrated Services Digital Networks*. Addison-Wesley, Reading, MA.
- Jain, Raj. 1994. *Handbook: High-Speed Networking Using Fiber and Other Media*. Addison-Wesley, Reading, MA.
- McDysan, D. E. and Spohn, D. E. 1994. *ATM: Theory and Application*. McGraw-Hill, New York.
- Rose, M. 1990. *The Open Book*. Prentice Hall, Englewood Cliffs, NJ.
- Schwartz, M. 1987. *Telecommunications Networks: Protocols, Modeling and Analysis*. Addison-Wesley, Reading, MA.
- Spragins, J. D. 1991. *Telecommunications: Protocols and Design*. Addison-Wesley, Reading, MA.
- Stallings, W. 1990. *Handbook of Computer-Communications Standards: The Open Systems Interconnection (OSI) Model and OSI-Related Standards*. Macmillan, New York.

## Further Information

There are many conferences and workshops that provide up-to-date coverage in the computer communications area. Among these are the IEEE INFOCOM and ACM SIGCOMM conferences and the IEEE Computer Communications Workshop, which are specialized to computer communications and are held annually. In addition, IEEE GLOBCOM (annual), IEEE ICC (annual), IFIPS ICC (biannual), and the International Telecommunications Congress (biannual) regularly feature a substantial number of paper and panel sessions in networking.

The *ACM Communications Review*, a quarterly, specializes in computer communications and often presents summaries of the latest standards activities. *IEEE Network*, a bimonthly, specializes in tutorially oriented articles across the entire breadth of computer communications and includes a regular column on books related to the discipline. Additionally, *IEEE Communications* and *IEEE Computer*, monthly magazines, frequently have articles on specific aspects of networking. Also, see *IEEE Personal Communication Systems*, a quarterly magazine, for information on wireless networking technology.

For those who wish to be involved in the most up-to-date activities, there are many interest groups on the internet, a worldwide TCP/IP-based network, that specialize in some aspect of networking. Searching for information on the Internet has become greatly simplified with the advent of the World Wide Web and publicly available software to support access. *The User's Directory of Computer Networks*, Digital Press, T. L. LaQuey (ed.) provides an excellent introduction to the activities surrounding internetworking and advice on how to obtain timely information.

Fordyce, S.W., Wu, W. W. "Satellites and Aerospace"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



129.1 Communications Satellite Services and Frequency Allocation

129.2 Information Transfer and Link Margins—Ground to Space (Up-Link)

129.3 Communication Satellite Orbits

129.4 Launch Vehicles

129.5 Spacecraft Design

129.6 Propagation

129.7 Earth Stations

**Samuel W. Fordyce**

*Consultare Technology Group*

**William W. Wu**

*Consultare Technology Group*

## 129.1 Communications Satellite Services and Frequency Allocation

---

An agency of the United Nations, the International Telecommunications Union (ITU) issues the radio regulations that have treaty status among the ITU members. The services of interest in satellite communications include: the fixed satellite service, the mobile satellite service (including maritime, aeronautical, and land mobile vehicles), broadcast satellite service, and intersatellite service. Most commercial communications satellites operating in the fixed satellite service use the C band (6 GHz up, 4 GHz down) or the Ku band (14 GHz up, 12 GHz down). The mobile satellite services operate primarily in the L band (1.6 GHz up, 1.5 GHz down). Exact frequencies and permitted signal characteristics are contained in the radio regulations and tables of frequency allocations issued by the ITU in Geneva, Switzerland.

## 129.2 Information Transfer and Link Margins<sup>3/4</sup>Ground to Space (Up-Link)

---

A transmitter with an output power of  $P_t$ , transmitting through an antenna with a gain of  $G_t$ , will provide a power-flux density,  $\phi$ , at a range of  $r_u$ , according to the formula

$$\phi = \frac{P_t G_t}{4\pi r_u^2} \quad (129.1)$$

The satellite receiver antenna with an effective aperture of  $A_r$  located at a range from the ground transmitter will have a received power level,  $P_r$ , given by

$$P_r = \phi A_r = \frac{\phi G_u \lambda_u^2}{4\pi} \quad (129.2)$$

where  $\lambda_u$  is the wavelength of the up-link, and  $G_u$  is the gain of the receiving antenna.

The signal is received in the presence of thermal noise from the receiver plus external noise. The total noise power density is  $N_0 = kT_s$ , where  $k = 1.38 \cdot 10^{-23}$  J/K (joules per kelvin) is Boltzmann's constant and  $T_s$  is the system temperature. The noise power in the radio frequency (rf) transmission bandwidth,  $B$ , is  $N = N_0 B$ .

The up-link signal-to-noise power ratio is

$$P_u/N_u = (P_t G_t G_u / kT_s B) (\lambda / 4\pi r_u)^2 \quad (129.3)$$

In decibels,

$$\begin{aligned} P_u/N_u = & 10 \log P_t G_t - 20 \log (\lambda / 4\pi r_u) - 10 \log B \\ & + 10 \log (G/T) - 10 \log k \end{aligned} \quad (129.4)$$

The terms of this equation represent the earth station *effective isotropic radiated power* (EIRP) in dBW, the *free space loss* in dB, the rf channel bandwidth in dB Hz, the satellite  $G/T$  ratio in dB/K, and Boltzmann's constant, which is  $-228.6$  dBW/Hz K.

Similarly, the signal-to-noise power ratio of the space-to-earth link, or *down-link* ( $P_d/N_d$ ), yields

$$P_d/N_d = (P_s G_s G_e / kT_e B) (\lambda_d / 4\pi r_d)^2 \quad (129.5)$$

where  $P_s$  is the power of the satellite transmitter,  $G_s$  is the gain of the satellite transmitting antenna,  $G_e$  is the gain of the earth station receiving antenna,  $k$  is Boltzmann's constant,  $T_e$  is the temperature of the earth station receiver,  $B$  is the bandwidth,  $\lambda_d$  is the wavelength of the down-link, and  $r_d$  is the range of the down-link.

The overall system power-to-noise ratio ( $P_s/N_s$ ) is given by

$$\frac{1}{P_s/N_s} = \frac{1}{P_u/N_u} + \frac{1}{P_d/N_d} \quad (129.6)$$

Signals relayed via communications satellites include voice channels (singly, or in groups or supergroups), data channels, and video channels. Signals from multiple sources are multiplexed

onto a composite baseband signal. The techniques used to modulate these signals so that they can be kept separate use a physical domain such as frequency, time, space (separate antenna beams), or encoding. Access can be preassigned or assigned on demand. The three principal modes of multiple access include the following:

*Frequency-demand multiple access* (FDMA) isolates signals by filtering different frequencies.

*Time-Demand Multiple access* (TDMA) isolates signals by switching time slots.

*Code-Division Multiple access* (CDMA) isolates signals by correlation.

## 129.3 Communication Satellite Orbits

---

Earth-orbiting satellite trajectories are conic sections with the earth's center of mass located at one focus. The simplest orbit is circular, wherein the centrifugal force on the satellite is balanced by the gravitational force:

$$F_c = F_g \quad (129.7)$$

$$\frac{mv^2}{r} = \frac{GMm}{r^2} \quad (129.8)$$

where  $v$  is the satellite's velocity,  $r$  is the distance from the earth's center to the satellite,  $m$  is the satellite's mass,  $M$  is the earth's mass, and  $G$  is the constant of universal gravitation. Solving for  $v$  gives

$$v = \sqrt{\frac{GM}{r}} \quad (129.9)$$

$$(GM = 3.9858 \cdot 10^5 \text{ km}^3/\text{s}^2) \quad (129.10)$$

If the satellite is in an elliptical orbit, with the semimajor axis of the ellipse given by  $a$ , the velocity at any point is given by the "vis-viva" equation:

$$v = \sqrt{GM \left( \frac{2}{r} - \frac{1}{a} \right)} \quad (129.11)$$

The maximum distance from the earth's center is known as the *apogee radius* ( $A$ ), and the minimum distance is known as the *perigee radius* ( $P$ ). The eccentricity ( $e$ ) of the ellipse is given by

$$e = \frac{A - P}{A + P} \quad (129.12)$$

The velocity at apogee  $v_A$  is given by

$$v_A = \sqrt{\frac{GM}{A}(1 - e)} \quad (129.13)$$

and at perigee

$$v_P = \sqrt{\frac{GM}{P}(1 + e)} \quad (129.14)$$

The orbit period ( $T$ ) is given by

$$T = 2\pi\sqrt{\frac{a^3}{GM}} \quad (129.15)$$

Most communications satellites are in geostationary orbits, which are in the equatorial plane with a period equal to one (sidereal) day, which is approximately four minutes shorter than a solar day. These satellites have an altitude of approximately 36000 km.

Geostationary satellites appear to be stationary to observers (and antennas) on earth. The antennas are not required to track such a stationary target. Low-gain antennas with broad beams do not need to track satellites, and many of the satellites designed to operate with small mobile terminals use orbits with lower altitudes than the geostationary satellites. These orbits usually have high inclinations to the equatorial plane to provide coverage of the earth's surface to high latitudes.

## 129.4 Launch Vehicles

---

Launch services that were once the province of government organizations have become commercial enterprises. Practically all of the commercial launches to date have been to the geostationary orbit.

Recently, the most popular launch provider has been Arianespace, with launches from Kourou, French Guiana, on the Atlantic coast of South America. Three launch vehicles used for commercial launches in the U.S. are McDonnell Douglas's Delta and Martin Marietta's Atlas and Titan. In the U.S. all eastward launches are conducted from Cape Canaveral; the polar and near-polar launches use Vandenberg Air Force Base in California. Russia provides launches using the Proton and Zenit launch vehicles from Baikonur (Tyuratam) in Kazakhstan and from Plesetsk in Russia. China's Great Wall Trading Co. provides launches on the Long March launch vehicle from Xichang, China. Japan launches from Tanegashima, Japan, using the H-1 and H-2 launch vehicles. Locations of these launch sites are given in [Table 129.1](#).

**Table 129.1** Launch Site Locations

Site	Latitude	Longitude
Kourou, French Guiana	5°N	53°W
Cape Canaveral, Florida	28°N	81°W
Vandenberg Air Force Base, California	35°N	121°W
Baikonur (Tyuratam), Kazakhstan	46°N	63°E
Plesetsk, Russia	63°N	35°E
Xiachang, People's Republic of China	28°N	102°E
Tanegashima, Japan	30°N	131°E

The velocity increment ( $v$ ) gained by a launch vehicle when a propellant is burned to depletion is given by the equation

$$v = c \log_e M \quad (129.16)$$

The term  $c$  is the characteristic velocity of the propellants and is often expressed as

$$c = I_{sp}g \quad (129.17)$$

where  $I_{sp}$  is the specific impulse of the propellants (typically,  $I_{sp} = 300$  s for lox/kerosene propellants in space) and  $g$  is the acceleration due to gravity. The mass ratio ( $M$ ) can be expressed as

$$M = \frac{\text{Propellents} + \text{structure} + \text{payload}}{\text{Structure} + \text{payload}} \quad (129.18)$$

This velocity increment assumes that the vehicle doesn't change altitude appreciably during the propellant burn or experience significant aerodynamic drag.

## 129.5 Spacecraft Design

The primary subsystem in a communications satellite is the payload, which is the communications subsystem. The supporting subsystems include structure; electrical power; thermal control; attitude control; propulsion; and the telemetry, tracking, and control (TT&C).

The communications subsystem includes the receiving antennas, receivers, transponders, and transmitting antennas. This payload makes up 35 to 60% of the mass of the spacecraft. The communications capability of the satellite is measured by the antenna gain ( $G$ ), receiver sensitivity ( $T_r$ ), transmit power ( $P_t$ ), and signal bandwidth ( $B$ ). The sensitivity is often expressed as  $G_r/T_r$ , and the radiated power (EIRP) as  $G_t P_t$ .

Early models of communications satellites used antennas with broad beams and little directivity. These beams were spread over the whole earth and required large, highly directional antennas on

the ground to pick up the weak satellite signals.

Improvements in the satellite attitude control subsystems and in launch vehicle capabilities have enabled satellites to carry large antennas with multiple feeds to provide narrow beams that "spotlight" the desired coverage area ("footprint") on earth. These footprints can be focused and contoured to cover designated areas on earth. The narrow beams also permit multiple reuse of the same frequency allocations.

Antenna gain  $G$  is given by

$$G = 4\pi A\eta/\lambda^2 \quad (129.19)$$

where  $A$  is the effective cross-sectional area,  $\eta$  is the antenna efficiency (typically 55 to 80%), and  $\lambda$  is the wavelength. At the Ku-band down-link (12 GHz), where  $\lambda = 2.5$  cm, a 1 m diameter antenna can provide a gain of 40 dB. The half-power beam width in degrees  $\theta$  is given approximately by

$$\theta \approx 21/fD \quad (129.20)$$

where  $f$  is the frequency in gigahertz and  $D$  is the antenna diameter in meters. In this example the beam width  $\theta = 1.75^\circ$ . Using the same antenna reflector on the Ku band up-link (14 GHz), the beam width would be slightly narrower ( $1/5^\circ$ ) if the reflector is fully illuminated.

Transponders are satellite-borne microwave repeaters. A typical C-band transponder is composed of filters and low-noise amplifiers to select and amplify the received (6 GHz) signal. Local oscillators are fed into mixers along with the incoming signal to produce intermediate frequency (IF) signals, which are further amplified before mixing with another local oscillator to produce the down-link signal. This signal is fed to a high-power amplifier (HPA). Traveling wave tubes (TWTs) were used originally as HPAs but have been replaced by solid-state power amplifiers (SSPAs) at C-band. The amplified signal is fed to the transmitting antenna for transmission to earth on the down-link (4GHz).

Geostationary communications satellites operating at C band usually carry 24 transponders with 40 MHz separation between them. Dual polarizations permit double use of the 500 MHz frequency allocation.

Without onboard processing the transponders are usually "transparent," in that they receive the incoming (up-link) signals, amplify them, and change the carrier frequency before transmitting the down-link signals. Transparent transponders provide no signal processing other than amplification and changing the carrier frequency (heterodyning). These transponders can relay any signals that are within the transponders' bandwidth. Other types of transponders can demodulate the incoming signals and remodulate them on the down-link carrier frequencies.

The *spacecraft structure* must support the payload and the subsystems through the propulsion phases as well as in orbit. The accelerations can be as high as 6–8 g during the launch phases. On orbit the structure must permit deployment and alignment of solar cell arrays and antenna reflectors. The spacecraft utilization factor,  $U$ , is defined as

$$U = \frac{M_u}{M} \quad (129.21)$$

where  $M_u$  is the mass of the communications payload and power subsystems, and  $M$  is the total spacecraft mass in orbit.

Representative values of  $U$  range from 0.35 to 0.60. Typical large geostationary communications satellites have a mass of 2000 kg on orbit.

The *electric power subsystem* uses solar cell arrays to provide electric power. The transmitters of the communications subsystem consume most of this power. Increasing demands for electric power led to large deployable solar arrays composed of silicon solar cells. Typically, power demand is for 2 kW of 28 V DC power. Power-conditioning units maintain the voltage levels within prescribed specifications. Rechargeable batteries (nickel-cadmium or nickel-hydrogen) provide power in emergencies and during solar eclipses.

The *thermal control subsystem* must maintain the specified temperature for all components of the spacecraft. Heat sources include incident sunlight and internal electrical heat sources (especially the HPAs). The only way to eliminate heat is to radiate it to space. The Stefan-Boltzmann law of radiation gives the radiated heat ( $q$ ) as  $q = \varepsilon A \sigma T^4$  where  $\varepsilon$  is the emissivity, which is between 0 and 1 (a black body has  $\varepsilon = 1$ ),  $A$  is the surface area;  $\sigma$  is the Stefan-Boltzmann constant ( $= 5.760 \pm 0.007 \cdot 10^{-8} \text{ W/m}^2 \text{ K}^4$ ); and  $T$  is the absolute temperature.

The temperature of a passive black sphere in geostationary orbit around the earth in sunlight is between 275 and 280 K. The incident solar flux density ( $S$ ) is approximately  $1.37 \text{ kW/m}^2$ . To prevent absorption of this heat energy requires the spacecraft to have a low absorptivity ( $\alpha$ ). A designer can control the temperature by controlling the absorptivity/emissivity ratio and by using spacecraft radiators.

The satellite's average temperature is given by

$$T = \frac{1}{\sigma} \left( \frac{\alpha a S}{\varepsilon A} + \frac{Q}{\varepsilon A} \right) \quad (129.22)$$

where  $a$  is the projected area (facing the sun),  $A$  is the total surface area, and  $Q$  is the internal heat dissipation.

Satellites need *attitude control and station-keeping subsystems* to maintain their orientation so that the antenna beams will illuminate the desired coverage areas on earth, so that the solar arrays will intercept the sun's rays, and so that velocity increments from the onboard propulsion subsystem will keep the satellite in its desired location.

Onboard sensors can be used to detect the earth's horizon, the sun, and reference stars. Radio-frequency sensors are also used to detect beacons from earth stations.

Inertial measurements from on-board gyroscopes are used to detect attitude changes. These changes can be made using onboard reaction wheels, which transfer angular momentum between the spacecraft and the wheel, or by the reaction jets on the attitude control system.

Communications satellites are usually held to accuracies of approximately  $0.1^\circ$  or less in all three

axes (yaw, pitch, and roll). The onboard propulsion system can provide velocity for the initial insertion in orbit and subsequent station-keeping velocity increments, as well as attitude control.

A typical communications satellite has propellant tanks containing monopropellant hydrazine ( $\text{N}_2\text{H}_4$ ). On command, this fuel will flow through a valve to a thruster that contains a catalyst bed, a combustion chamber, and a nozzle. The resulting force from the thruster is used for the velocity corrections needed for station keeping or for attitude control. This force is given by

$$F = \dot{W} I_{\text{sp}} \quad (129.23)$$

where  $\dot{W}$  is the propellant weight flow rate, and  $I_{\text{sp}}$  is the specific impulse, which is about 230 s for the example described.

By using electrically heated thrusters and bipropellant systems, the specific impulse can be raised to 300 s or more. Some communications satellites have such large propellant tanks that velocity increments of several km/s can be achieved.

*Telemetry, tracking, and command (TT&C) subsystems* are provided for satellites in order to determine the status, performance, position, and velocity of the satellite; to control it; and to provide commands. Data from onboard sensors are transmitted by the telemetry transmitters to the ground control for monitoring the status of the satellite. Beacons are tracked in range and angle to determine the position and velocity. Ground commands are received, demodulated, and processed by the command receivers onboard the satellite. TT&C subsystems include high-power transmitters and omnidirectional antennas that can provide communications even if attitude control is lost and the spacecraft is tumbling. Link margins are similar to those described for the communications payload, but with more robust signal-to-noise ratios.

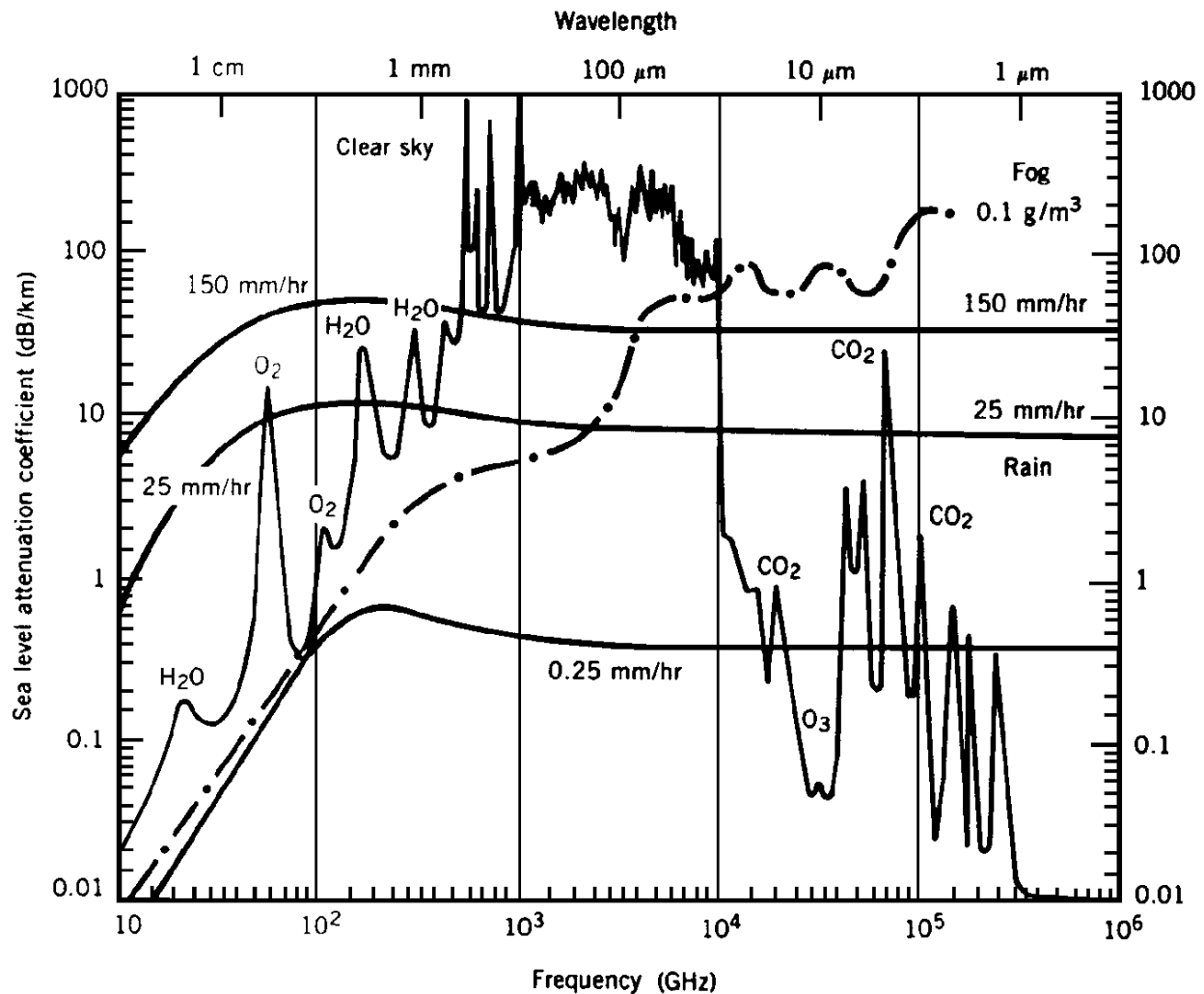
## 129.6 Propagation

---

Free space is transparent to electromagnetic waves; however, waves undergo refraction and absorption when passing through the troposphere and the ionosphere. The most important effect upon satellite communications is the clear sky attenuation caused by the molecular resonance absorption bands. The clear sky attenuation per km is plotted versus frequency in [Fig. 129.1](#). The resonance of the water vapor molecules reaches 0.15 dB/km at 25 GHz. During heavy rainstorms, this attenuation can increase significantly, as shown in [Fig. 129.1](#).



**Figure 129.1** Atmospheric path losses at sea level.

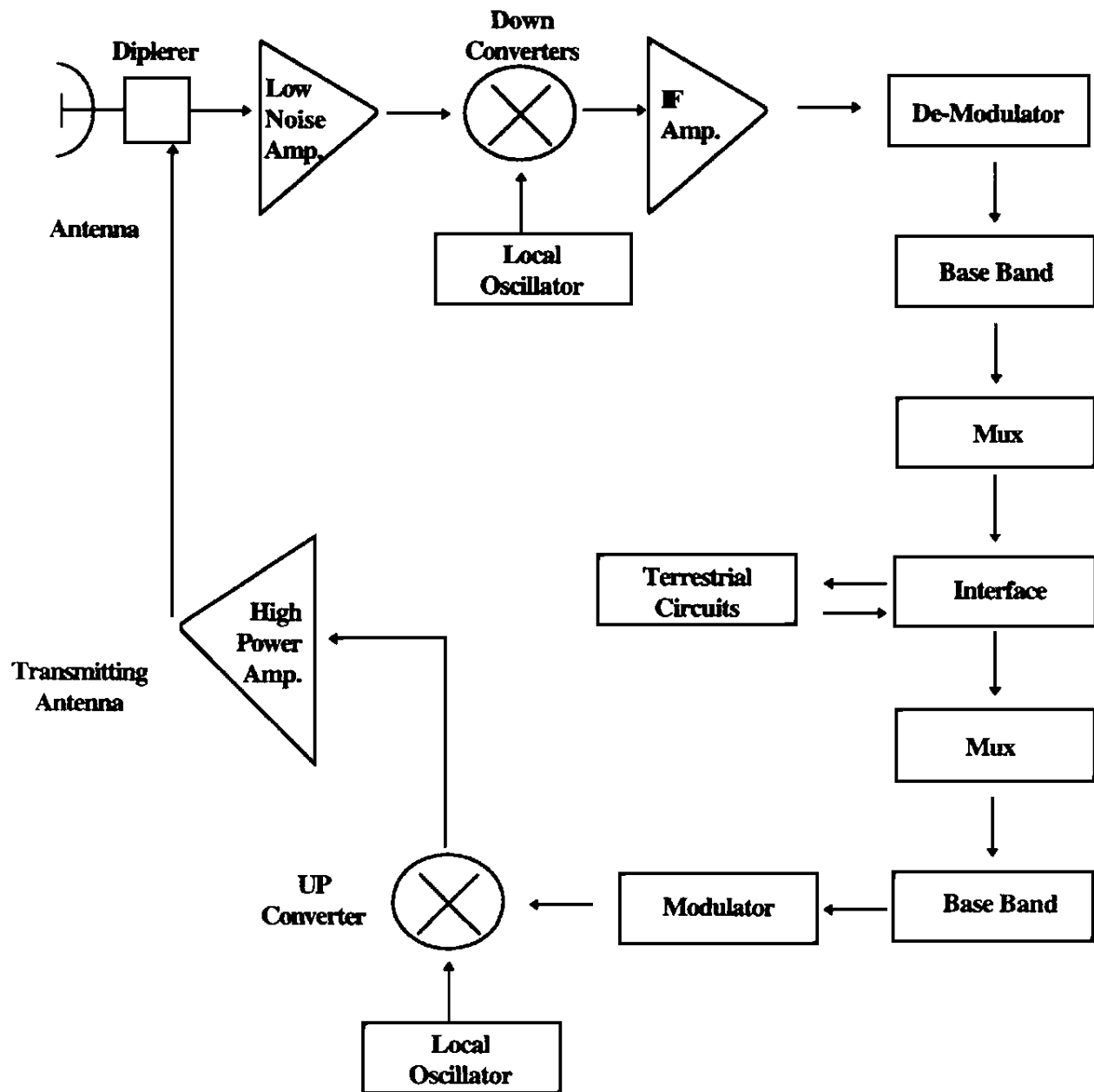


Below 10 GHz the atmosphere is transparent, but the resonant peaks shown for various molecules in the atmosphere can cause serious attenuation in the K band. The use of this band may be hampered by heavy rainstorms, which can cause depolarization as well as attenuation. One technique used to overcome this problem is to use diversity ground stations, spaced 5 to 10 km apart, to avoid transmission through localized thunderstorms.

## 129.7 Earth Stations

Earth stations are located on the ground segment of satellite communications systems. They provide communications with the satellites and interconnections to the terrestrial communications systems. A simplified block diagram is shown in [Fig. 129.2](#).

**Figure 129.2** Earth station block diagram.



The antennas, the most prominent subsystems, focus the beams to enhance the sensitivity of the receivers, to enhance the radiated power of the transmitters, and to discriminate against radio interference from transmitters located outside the narrow antenna beam.

## Further Information

More complete explanations of this material are contained in

*An Introduction to the Mathematics and Methods of Astrodynamics*, by Richard H. Battin,

AIAA, 1633 Broadway, New York, NY 10019.

*Reference Guide to Space Launch Systems*, by Stephan J. Isakowitz in a 1991 AIAA Publication, 370 L'Enfant Promenade SW, Washington, DC 20024-2518.

*Satellite Communications*, by Timothy Pratt and Charles Bostian, published by John Wiley & Sons, New York.

*Communications Satellite Handbook*, by Walter Morgan and Gary Gordon, also published by John Wiley & Sons, New York.

*Satellite Communications System Engineering*, by Wilbur Pritchard, Henri Suyderhoud, and Robert Nelson, Prentice Hall, Englewood Cliffs, NJ 07632.

*Space Vehicle Design*, by Michael D. Griffin and James R. French, AIAA, 370 L'Enfant Promenade SW, Washington, DC 20024-2518

*Radiowave Propagation in Satellite Communications*, by Dr. Louis Ippolito Jr., published by Van Nostrand–Reinhold, New York.

Developments in satellite communications are discussed in many symposia, including those sponsored by the IEEE (Institute of Electrical and Electronic Engineers) and the AIAA (American Institute of Aeronautics and Astronautics).

Rappaport, T. S., et al. "Mobile and Cellular Radio Communications"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 130

## Mobile and Cellular Radio Communications

---

130.1 Paging Systems

130.2 Cordless Telephone Systems

130.3 Cellular Telephone Systems

130.4 Personal Communications System (PCS)

130.5 The Cellular Concept and System Fundamentals

Frequency Reuse • Channel Assignment and Handoff Strategies

130.6 System Capacity and Performance of Cellular Systems

Radio Interference and System Capacity • Grade of Service

130.7 Mobile Radio Systems Around the World

**Theodore S. Rappaport**

*Virginia Polytechnic Institute & State University*

**Rias Muhamed**

*Virginia Polytechnic Institute & State University*

**Michael Buehrer**

*Virginia Polytechnic Institute & State University*

**Anil Doradla**

*Virginia Polytechnic Institute & State University*

Mobile radio communication systems transport information using electromagnetic waves to provide a viable communication link between a transmitter and receiver, either or both of which may be in motion or stationary at arbitrary locations. A variety of mobile radio systems are in use today, and the industry is experiencing rapid growth. In the U.S. alone, for example, the number of cellular telephone users grew from 25000 in 1984 to about 15 million in 1994. In Sweden, cellular telephones are already used by over 10% of the population. This growth is expected to continue worldwide at an even greater pace during the next decade.

Mobile radio transmission systems may be classified as *simplex*, *half-duplex*, or *full-duplex*. In simplex systems communication is possible in only one direction. Paging systems, in which

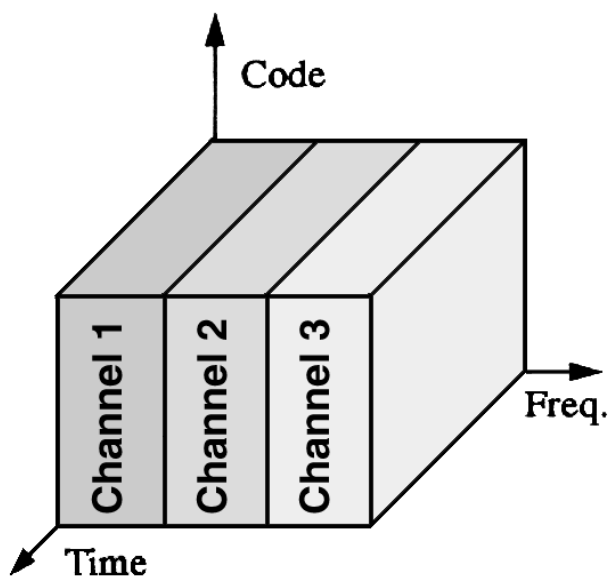
messages are received but not acknowledged, are simplex systems. Half-duplex radio systems allow two-way communication but use the same radio channel for both transmission and reception. This means that at any given time a user can only transmit or receive information. Constraints like "push to talk" and "release to listen" are fundamental features of half-duplex systems. Full-duplex systems, on the other hand, allow simultaneous radio transmission and reception by providing two separate channels or time slots for communication to and from the user. The channel used to convey traffic to the mobile user is called the *forward channel*, whereas the channel used to carry traffic from the mobile user is called the *reverse channel*. Full-duplex mobile radio systems provide many of the capabilities of the standard telephone, with the added convenience of mobility.

Modern mobile communication systems are networked through a central switching system, called a *mobile telephone switching office* (MTSO) or a *mobile switching center* (MSC), and typically provide coverage throughout a large metropolitan area. Many mobile radio systems are connected to the *public switched telephone network* (PSTN), which enables the full-duplex connection of mobile users to any telephone throughout the world. In order to utilize the available spectrum and radio equipment efficiently, these networked systems rely on **trunking**, so that a limited number of frequencies can accommodate a large number of mobile users on a statistical demand basis. Trunking exploits the fact that not every user requires service at a particular time; hence it is possible for a large number of users to share a relatively small number of radio channels. When a subscriber in a trunked system engages in a call, the system searches for a free channel and allocates it to the subscriber. If all the channels are already in use, the user is *blocked*, or denied access to the system. The process of allocating a channel in a trunked radio system requires a dedicated control channel, called the *control* or *call setup* channel.

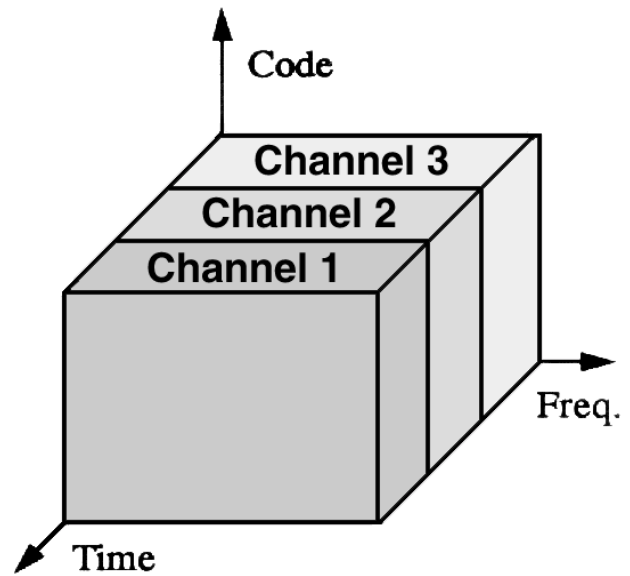
Multiple-access techniques such as *frequency-division multiple access* (FDMA), *time-division multiple access* (TDMA), and *code-division multiple access* (CDMA) are employed to provide simultaneous access to many users without causing mutual interference. In FDMA separate users are allocated separate frequency bands (channels) for their exclusive use for the entire duration of a call. In TDMA several users share the same radio channel but are assigned unique time slots in which they communicate. In CDMA every user transmits at the same frequency and at the same time, and interference is avoided by the use of specialized codes that are unique to each user and uncorrelated between users. [Figure 130.1](#) illustrates the three multiple-access techniques.

Handheld walkie-talkies, paging receivers (pagers), cordless telephones, and cellular telephones are examples of mobile radio systems that are commonly used today. The complexity, performance, required infrastructure, and types of services offered by each of these systems are vastly different. In this chapter a variety of modern mobile radio systems are described with an emphasis placed on modern cellular radio systems. The chapter concludes with a summary of all major mobile radio system standards in use throughout the world.

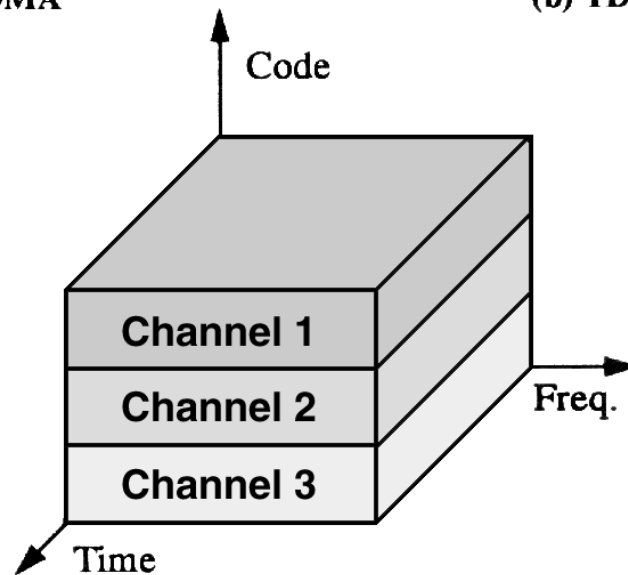
**Figure 130.1** Illustration of the three multiplexing techniques, and how multiple channels are provided without interfering with each other: (a) FDMA, (b) TDMA, and (c) CDMA.



(a) FDMA



(b) TDMA



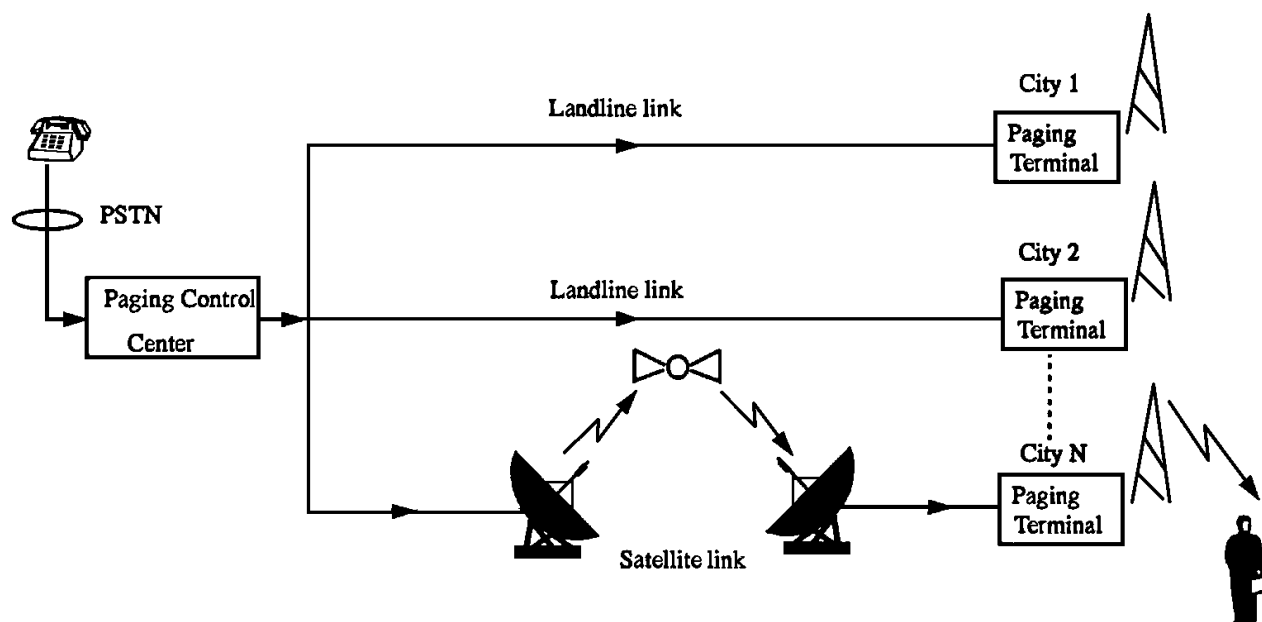
(c) CDMA

## 130.1 Paging Systems

Paging systems are simplex communication systems that can be used to send brief messages to a subscriber. Depending on the type of service, the message can be a numeric message, an alphanumeric message, or a voice message. Paging systems are typically used to notify a subscriber of the need to call a particular telephone number or to travel to a known location to receive further instructions. Anyone may send a message to a paging subscriber by dialing the paging system number (usually a toll-free telephone number) and using a telephone keypad or modem to issue a message, called a *page*. The paging system then transmits the page throughout the service area using base stations which broadcast the page on a radio carrier.

Paging systems vary widely in their complexity and coverage area. Whereas simple paging systems may only cover a limited range of 2–5 km, wide-area paging systems may provide worldwide coverage. Though paging receivers are simple and inexpensive, the transmission system required can be quite sophisticated. Wide-area paging systems consist of a network of telephone lines, large radio towers, satellite links, and powerful transmitters and simultaneously dispatch a page to many transmitters (this is called *simulcasting*), which may be located within the same service area or in different cities or countries. Paging systems are designed to provide reliable communication to subscribers wherever they are—whether inside buildings, driving on a highway, or in an airplane. This necessitates large transmitter powers (on the order of kilowatts) and low data rates (a few thousand bits per second) for maximum coverage. Figure 130.2 shows a diagram of a wide-area paging system.

**Figure 130.2** Example of a wide-area paging system. The paging control center dispatches pages received from the PSTN throughout several cities at the same time.





## 130.2 Cordless Telephone Systems

---

Cordless telephone systems are full-duplex communication systems that use radio to connect a handset to a dedicated base station. The base unit is connected to a dedicated telephone line with a specific telephone number. In first-generation cordless telephone systems (manufactured in the 1980s), the portable unit communicates only to the dedicated base unit, and only over distances of a few tens of meters. Early cordless telephones operate as extension telephones to a **transceiver** connected to a subscriber line on the PSTN and were developed primarily for in-home use. Second-generation cordless telephones have recently been introduced that allow subscribers to use their handsets at many outdoor locations within urban centers, such as London or Hong Kong. Modern cordless telephones are sometimes combined with paging receivers so that a subscriber may first be paged and then respond to the page using the cordless telephone. Cordless telephone systems provide the user with limited range and limited mobility, as it is usually not possible to maintain a call if the user travels outside the range of the base station. Typical second-generation base stations provide coverage ranges of a few hundred meters.

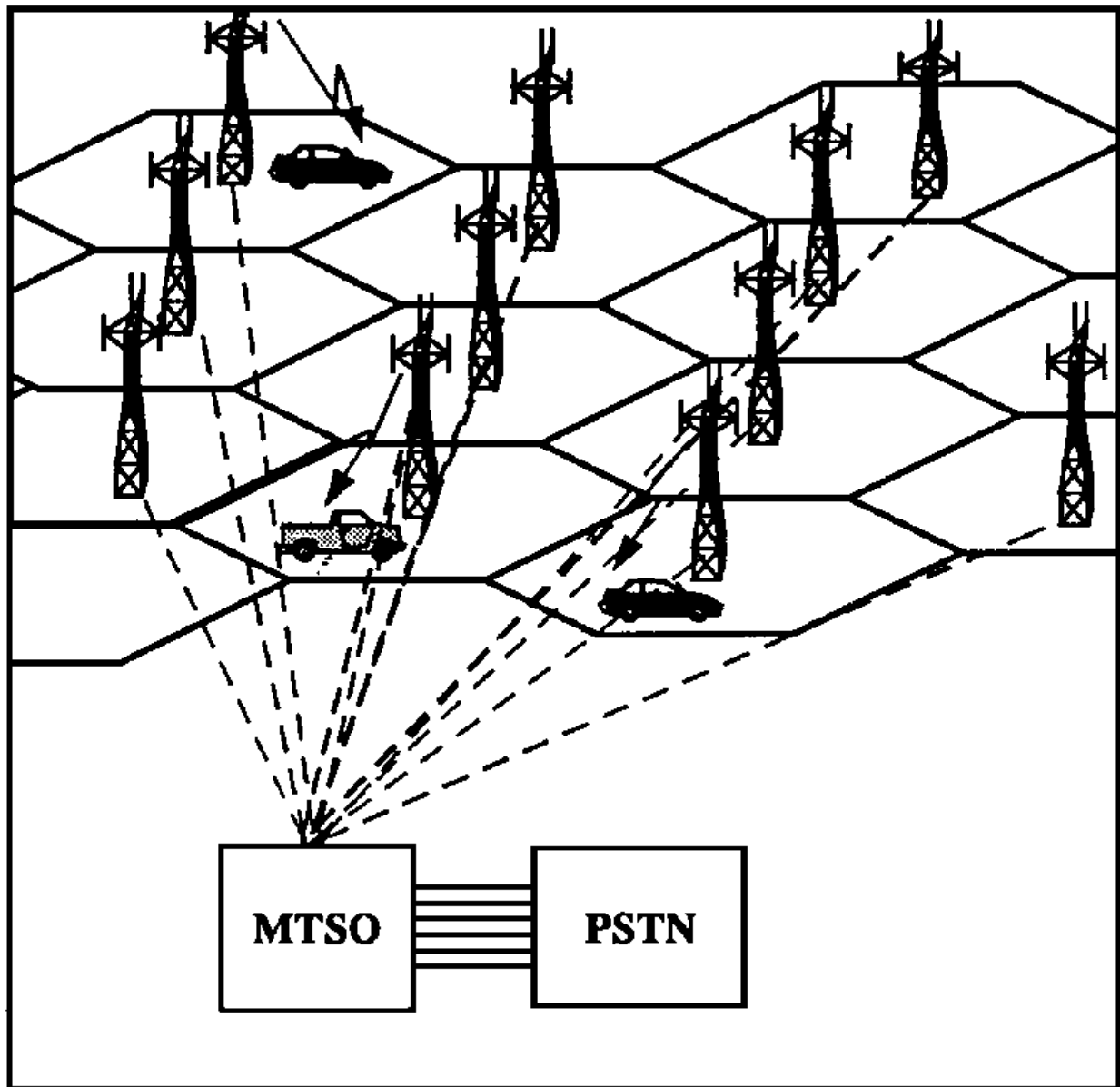
## 130.3 Cellular Telephone Systems

---

A cellular telephone system provides wireless access to the PSTN for any user located within the radio range of the system. Cellular systems accommodate a large number of users over a large geographic area, within a limited frequency spectrum. Cellular radio systems provide high-quality service, often comparable to that of landline telephone systems. High capacity is achieved by limiting the coverage of each transmitter to a small geographic area so that the same radio channels can be reused by another transmitter located a small distance away. A sophisticated switching technique called a **handoff** enables a call to proceed uninterrupted when the user moves from one area to another.

[Figure 130.3](#) shows a basic cellular system that consists of **mobile stations**, **base stations**, and a **mobile telephone switching office (MTSO)**. Each user communicates via radio with one of the base stations and may be handed off to any number of base stations throughout the duration of a call. The mobile station contains a transceiver, an antenna, and a control unit and can be mounted in a vehicle or handheld package. The base stations consist of several transmitters and receivers and generally have towers that support several transmitting and receiving antennas. The base station serves as a bridge between all mobile users in a geographic area and is connected via telephone lines or microwave links to the MTSO. The MTSO coordinates the activities of all the base stations and connects the entire cellular system to the PSTN.

**Figure 130.3** An illustration of a cellular system. The towers represent base stations that provide radio access between mobile users and the mobile telephone switching office (MTSO).



Communication between the base station and the mobile takes place over four distinct channels. The channel used for voice transmission from base station to mobile is called the *forward voice channel (FVC)*, and the channel used for voice transmission from mobile to base station is called the *reverse voice channel (RVC)*. The other two channels are the forward and reverse *control channels*. Control channels transmit and receive data messages that carry call initiation and service requests. Control channels are always monitored by mobiles that do not have an active call in progress.

Most cellular systems provide a special service called *roaming*, which allows subscribers to move into service areas other than the one from which service is subscribed. Once a mobile enters

a city or geographic area that is different from its home service area, it is registered as a roamer in the new service area. Roaming mobiles are allowed to receive and place calls from wherever they happen to be.

## 130.4 Personal Communications System (PCS)

---

In the mid-1990s the worldwide demand for cellular telephone service led to the development of the personal communication system (PCS). PCS offers personal wireless communications and advanced data networking using the cellular radio concept. PCS has been allocated the radio spectrum in the 1.8–2.0 GHz band in many countries throughout the world. Under the auspices of the International Telecommunications Union (ITU), an international consortium is developing a universal standard for PCS, so that the same equipment may be used throughout the world. This consortium, called the Future Public Land Mobile Telecommunications System (FPLMTS), and recently renamed IMT-2000, is considering local, regional, national, and international networking using cellular and satellite radio communication.

## 130.5 The Cellular Concept and System Fundamentals

---

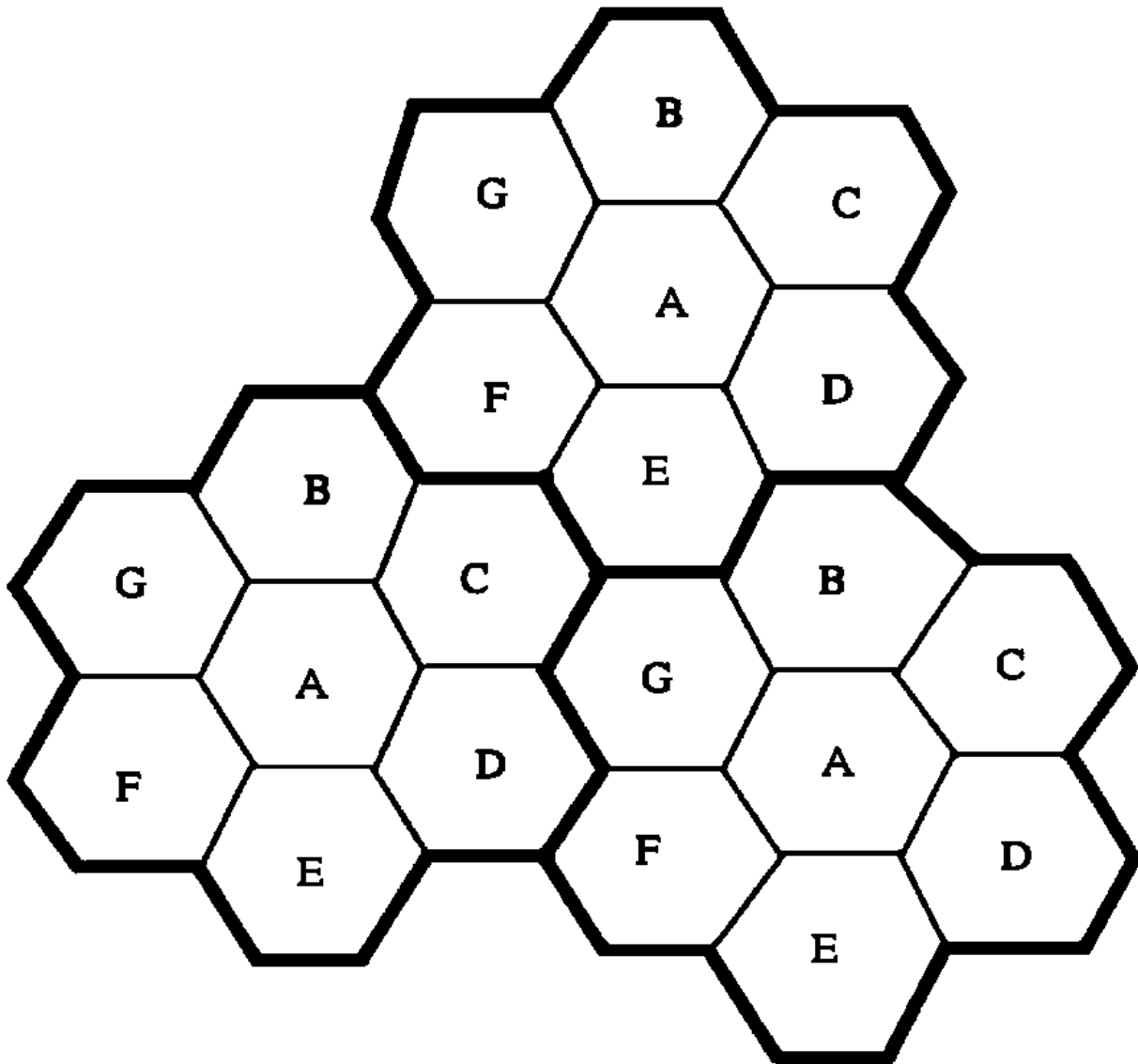
The need to grow a mobile subscriber population within a limited radio spectrum led to the development of systems based on the cellular concept [MacDonald, 1979]. The key to providing capacity in a cellular system is a technique called **frequency reuse**.

### Frequency Reuse

Cellular radio systems rely on an intelligent allocation and reuse of channels throughout a coverage region. A subset of channels is assigned to a small geographic area called a *cell*. Each cell is served by a base station that uses the assigned channel group. The power radiated by a base station is deliberately kept low, and antennas are located so as to achieve coverage within the particular cell. By limiting the coverage area within a cell, the same group of channels can be used to cover various cells that are separated from one another by distances large enough to keep the cochannel interference level within tolerable limits.

Figure 130.4 shows a cellular layout where cells labeled with the same letter use the same group of channels. Due to random propagation effects, actual cell coverage areas are amorphous in nature. However, for system design purposes it is useful to visualize cells as hexagons.

**Figure 130.4** Illustration of the cellular concept. Cells labeled with the same letter use the same set of frequencies. A cell cluster is outlined in bold and is replicated over the coverage area. In this example the cluster size,  $N$ , is equal to 7, and each cell contains  $1/7$  of the total number of available channels.



To understand the frequency reuse concept, consider a cellular system that has a total of  $S$  duplex channels available for use. If each cell is allocated a group of  $k$  channels ( $k < S$ ), and if the  $S$  channels are divided among  $N$  cells into unique and disjoint channel groups with the same number of channels, the total number of available radio channels can be expressed as

$$S = kN \quad (130.1)$$

The factor  $N$  is called the *cluster size* and is typically equal to 7 or 4. The  $N$  cells that use the complete set of frequencies are collectively called a *cluster*. If a cluster is replicated  $M$  times within the system, the total number of duplex channels,  $C$ , available to the system is given by

$$C = MkN = MS \quad (130.2)$$

As seen from Eq. (130.2), the capacity of a cellular system is directly proportional to the number of times a cluster is replicated in a given service area. If the cluster size  $N$  is reduced, more clusters are used to cover a given area and hence more capacity (larger value of  $C$ ) is achieved. The choice of  $N$  depends on the cochannel interference level that can be tolerated, as discussed later.

## Channel Assignment and Handoff Strategies

For efficient utilization of the radio spectrum, a frequency reuse scheme that is consistent with the objectives of increasing capacity and minimizing interference is required. Channel assignment strategies can be classified as either fixed or dynamic. The particular type of channel assignment employed affects the performance of the system, particularly in how calls are managed when a mobile user travels from one cell to another [Tekinay and Jabbari, 1991].

In a fixed channel assignment strategy, each cell is allocated a predetermined set of channels. Any call attempt within the cell can be served only by the unoccupied channels in that particular cell. If all the channels in that cell are occupied, the call is blocked and the subscriber does not receive service. In dynamic channel assignment, channels are not allocated to various cells permanently. Instead, each time a call is attempted, the cell base station requests a channel from the MTSO, which allocates a channel based on an algorithm that minimizes the cost of channel allocation.

Since neighboring cells use separate channels, when a mobile passes into a separate cell while a conversation is in progress, it is required to transfer the connection to the new cell base station automatically. This handoff operation not only involves identifying a new base station, but also requires that the voice and control signals be allocated to channels associated with the new base station.

Processing handoffs is an important task in any cellular mobile system. Many system designs prioritize handoff requests over call initiation requests. It is required that every handoff be performed successfully and that they happen as infrequently and imperceptibly as possible. In cellular systems the signal strength on either the forward or reverse channel link is continuously monitored and, when the mobile signal begins to decrease (e.g., when the reverse channel signal



### THE ENVOY

The Envoy wireless communicator is creating a whole new industry by enabling mobile professionals to exchange Internet messages, send faxes, check flight schedules, and manage appointments, addresses, and other personal information from wherever they happen to be—even an airport lounge. This handheld device uses Motorola two-way wireless communications and thus requires no phone lines or external connection to access information. Wireless products, like the Envoy, will enable millions of travelling workers to have the same tools as those workers who are at their desks. (Courtesy of Motorola.)

strength drops to below between  $-90$  dBm and  $-100$  dBm at the base station), a handoff occurs. In first-generation analog cellular systems, the MTSO monitor(s) the signals of all active mobiles at frequent intervals to determine a rough estimate of their location and decide if a handoff is necessary. In second-generation systems that use digital TDMA technology, handoff decisions are mobile assisted. In a **mobile-assisted handoff (MAHO)** the mobile stations make measurements of the received power from several surrounding base stations and continually report the results of these measurements to the base station in use, which initiates the handoff. The MAHO method enables faster handoff than in first-generation analog cellular systems since the MTSO is not burdened with additional computation.

## 130.6 System Capacity and Performance of Cellular Systems

---

Interference is the major limiting factor in the capacity and performance of cellular radio systems. The source of interference may be from an adjacent channel in the same cell, a signal from other base stations operating in the same frequency band, or signals from cochannel mobiles. Interference caused by signals from adjacent channels is called *adjacent channel interference*, and the interference between signals from different cells using the same frequency is called *cochannel interference*. Adjacent channel interference occurs due to imperfect receiver filtering, especially when an undesired transmitter is significantly closer to a receiver than the desired source. Adjacent channel interference is reduced by maximizing the frequency separation between channels in each cell through careful frequency planning in the cellular system.

### Radio Interference and System Capacity

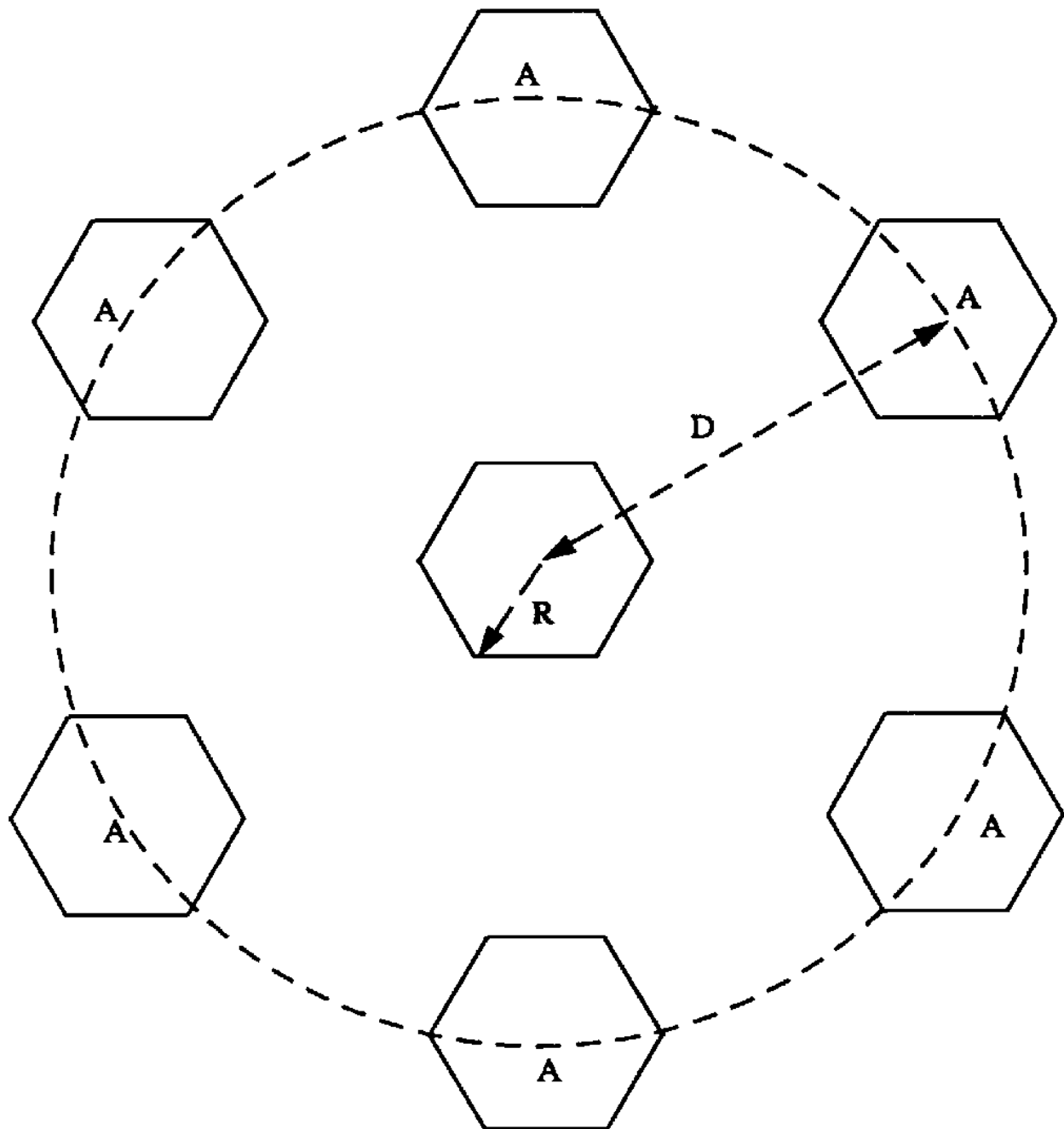
Cochannel interference is the major bottleneck for increasing capacity. Unlike noise, cochannel interference cannot be combated by simply increasing the carrier power. This is because an increase in carrier transmit power increases the interference as well. To reduce cochannel interference, cells using the same set of frequencies (cochannel cells) must be separated to provide sufficient isolation.

When the size of each cell in a cellular system is roughly the same, cochannel interference is independent of the transmitted power and is an increasing function of a parameter  $Q$ , called the **cochannel reuse ratio**. The value of  $Q$  is related to the cluster size  $N$  and is defined for a hexagonal geometry as

$$Q = \frac{D}{R} = \sqrt{3N} \quad (130.3)$$

where  $R$  is the major radius of the cells and  $D$  is the separation between cochannel cells as shown in Fig. 130.5. A small value of  $Q$  provides larger capacity for a particular geographic coverage region, whereas a large value of  $Q$  improves the transmission quality, due to a smaller level of cochannel interference. A trade-off is made between these two objectives in actual cellular design.

**Figure 130.5** Illustration of the first tier of cochannel cells for a cluster size of  $N = 7$ .



If  $i_0$  is the number of cochannel interfering cells, the signal to interference ratio ( $S/I$ ) at a receiver on the forward link can be expressed as



$$\frac{S}{I} = \frac{S}{\sum_{i=1}^{i_0} I_i} \quad (130.4)$$

where  $S$  is the desired signal from the desired base station and  $I_i$  is the interference caused by the  $i$ th outlying cochannel cell base station. If the signal levels of cochannel cells are known, then the  $S/I$  ratio for the reverse link can be found using Eq. (130.4).

Propagation measurements in a mobile radio channel show that the average received power at any point decays exponentially with respect to distance of separation between the transmitter and the receiver. The average received power  $P_r$  at a distance  $d$  from the transmitting antenna is approximated by either of the following:

$$P_r = P_0 \left( \frac{d_0}{d} \right)^n \quad (130.5)$$

$$P_r(\text{dBm}) = P_0(\text{dBm}) - 10n \log_{10} \frac{d}{d_0} \quad (130.6)$$

where  $P_0$  is the power received at a close-in reference point in the far-field region of the antenna at a distance  $d_0$  from the transmitting antenna, and  $n$  is the path loss exponent. Therefore, if  $D_i$  is the distance of the  $i$ th interferer from the mobile, the received power at a given mobile due to the  $i$ th interfering cell will be proportional to  $(D_i)^{-n}$ . The path loss exponent typically ranges between 2 and 4 in urban cellular systems [Rappaport and Milstein, 1992].

Assuming that the transmit power of each base station is equal and the path loss exponent is the same throughout the coverage area, the  $S/I$  ratio can be approximated as

$$\frac{S}{I} = \frac{R^{-n}}{\sum_{i=1}^{i_0} (D_i)^{-n}} \quad (130.7)$$

Considering only the first layer of interfering cells, if all the interfering base stations are equidistant from the desired base station and if this distance is equal to the distance  $D$  between cell centers, then Eq. (130.7) simplifies to

$$\frac{S}{I} = \frac{(D/R)^n}{i_0} = \frac{(\sqrt{3N})^n}{i_0} \quad (130.8)$$

Equation (130.8) relates  $S/I$  to the cluster size  $N$ , which in turn determines the overall capacity of the system. Hence, it is clear that cochannel interference determines the capacity of cellular systems.

## Grade of Service

Cellular systems rely on **trunking** to allow a large population of users to share a finite number of

radio channels. The quality of service in any trunked system is often measured using a benchmark called the **grade of service** (GOS). The grade of service is a measure of the ability of a particular user to access a trunked system during the busiest hour. GOS is typically given as the likelihood that a call is blocked or the likelihood of a call experiencing a delay greater than a certain queuing time.

The GOS for a trunked system that provides no queuing for blocked calls is given by the Erlang B formula,

$$\Pr[\text{blocking}] = \frac{A^C / C!}{\sum_{k=0}^C (A^k / k!)} \quad (130.9)$$

where  $C$  is the number of channels offered by the cell and  $A$  is the total traffic offered. The total offered traffic  $A$  is measured in erlangs, where one erlang represents the load over a channel that is completely occupied at all times. For a system containing  $U$  users, the total offered traffic can be expressed as

$$A = U\mu H \quad (130.10)$$

where  $\mu$  is the average number of call requests per unit time and  $H$  is the average duration of a typical call.

For trunked systems in which a queue is provided to hold calls that are blocked, the likelihood of a call not having immediate access to a channel is determined by the Erlang C formula:

$$\Pr[\text{delay} > 0] = \frac{A^C}{A^C + C![1 - (A/C)] \sum_{k=0}^{C-1} (A^k / k!)} \quad (130.11)$$

The GOS for a queued system is measured as the probability that a call is delayed greater than  $t$  seconds and is given by the probability that a call is delayed by a nonzero duration of time, multiplied by the conditional probability that the delay is greater than  $t$  seconds, as shown in Eq. (130.12):

$$\begin{aligned} \Pr[\text{delay} > t] &= \Pr[\text{delay} > 0] \Pr[\text{delay} > t \mid \text{delay} > 0] \\ &= \Pr[\text{delay} > 0] \exp[-(C - A)t/H] \end{aligned} \quad (130.12)$$

## 130.7 Mobile Radio Systems Around the World

Numerous mobile radio systems and services are in use around the world. There is a repertoire of standards that have been developed for the operation of these mobile radio systems, and many more are likely to emerge. [Tables 130.1, 130.2, and 130.3](#) provide listings of the most common paging, cordless, and cellular telephone standards in North America, Europe, and Japan.

**Table 130.1** Mobile Radio Standards in North America

Standard	Type	Year of Introduction	Multiple Access	Frequency Band	Modulation	Channel Bandwidth
AMPS	Cellular	1983	FDMA(FDD)	824–894 MHz	FM	30 kHz
NAMPS	Cellular	1992	FDMA	824–894 MHz	FM	10 kHz
USDC	Cellular	1991	TDMA(FDD)	824–894 MHz	$\pi/4$ -DQPSK	30 kHz
IS-95	Cellular	1993	CDMA	824–894 MHz	O-QPSK	1.25 MHz
GSC	Paging	1970s	Simplex FDM	Several	FSK	12.5 kHz
POCSAG	Paging	1970s	Simplex FDM	Several	FSK	12.5 kHz
FLEX	Paging	1993	Simplex FM	Several	4-FSK	15 kHz
PACS	Cordless/PCS	1993	TDMA/FDM A	1.8–2.2 GHz	$\pi/4$ -QPSK	300 kHz

**Table 130.2** Mobile Radio Standards in Europe

Standard	Type	Year of Introduction	Multiple Access	Frequency Band	Modulation	Channel Bandwidth
E-TACS	Cellular	1985	FDMA	900 MHz	FM	25 kHz
NMT-450	Cellular	1981	FDMA	450–470 MHz	FM	25 kHz
NMT-900	Cellular	1986	FDMA	890–960 MHz	FM	12.5 kHz
GSM	Cellular	1990	TDMA	890–960 MHz	GMSK	200 kHz
C-450	Cellular	1985	FDMA	450–465 MHz	FM	20 kHz/10kHz
ERMES	Paging	1993	FDMA	Several	4-FSK	25 kHz
CT-2	Cordless	1989	FDMA/(TDD)	864–868 MHz	GFSK	100 kHz
DECT	Cordless	1993	TDMA/(TDD)	1880–1900 MHz	GFSK	1.728 MHz
DCS-1800	Cordless	1993	TDMA	1710–1880 MHz	GMSK	200 kHz

**Table 130.3** Mobile Radio Standards in Japan

Standard	Type	Year of Introduction	Multiple Access	Frequency Band	Modulation	Channel Bandwidth
JTACS	Cellular	1988	FDMA	860–925 MHz	FM	25 kHz
JDC	Cellular	1992	TDMA	810–1513 MHz	$\pi/4$ -DQPSK	25 kHz
NTT	Cellular	1979	FDMA	400–800 MHz	FM	25 kHz
NTACS	Cellular	1993	FDMA	843–925 MHz	FM	12.5 kHz
NTT	Paging	1979	FDMA	280 MHz	FSK	12.5 kHz
NEC	Paging	1979	FDMA	Several	FSK	10 kHz
PHS	Cordless	1993	TDMA	1895–1907 MHz	$\pi/4$ -DQPSK	300 kHz

The two most common paging standards are the POCSAG (Post Office Code Standard Advisory Group), and GSC (Golay Sequential Code) paging standards. POCSAG was developed by British Post Office in the late 1970s and supports binary FSK signaling at 256 bps, 512 bps, 1200 bps, and 2400 bps. GSC is a Motorola paging standard that uses 300 bps for the pager address and 600 bps binary FSK for message transmission. New paging systems, such as FLEX and ERMES, will provide up to 6400 bps transmissions by using 4-level modulation.

The CT-2 and DECT standards developed in Europe are the two most popular cordless telephone standards throughout Europe and Asia. The CT-2 system makes use of microcells that cover small distances, usually less than 100 m, using base stations with antennas mounted on street lights or at low heights. The CT-2 system uses battery efficient frequency-shift keying along with a 32 kbps ADPCM speech coder for high-quality voice transmission. Handoffs are not supported in CT-2, as it is intended to provide short-range access to the PSTN. The DECT system accommodates data and voice transmissions for office and business users. In the U.S. the PACS standard, developed by Motorola and Bellcore, is likely to be used inside office buildings as a wireless voice and data telephone system.

The world's first cellular system was implemented by NTT in Japan. The system was deployed in 1979 and uses 600 FM duplex channels (25 kHz per one-way channel) in the 800 MHz band. In Europe the Nordic Mobile Telephone system (NMT450) was developed in 1981 for the 450 MHz band and uses 25 kHz channels. The Extended European Total Access Cellular System (ETACS) was deployed in 1985 and enjoys about 15% of the market share in Europe. In Germany a cellular standard called C-450 was introduced in 1985. The first-generation European cellular systems are generally incompatible with one another because of the different frequencies and communication protocols used. These systems are now being replaced by the Pan-European digital cellular standard GSM (Global System Mobile), which was first deployed in 1990. The GSM standard is gaining worldwide acceptance as the first digital cellular standard.

Unlike the incompatible first-generation cellular systems in Europe, the Advanced Mobile Phone System (AMPS) was introduced in the U.S. in 1983 as a nationwide standard ensuring that all cellular telephones are compatible with any cellular radio base station within the country. AMPS, like other first-generation cellular systems, uses analog FM modulation and frequency-division multiple access. As the demand for services continues to increase, cellular radio systems using more efficient digital transmission techniques are being employed. A U.S. TDMA-based digital standard called U.S. Digital Cellular (USDC or IS-54) has been in operation since 1991. In order to be compatible with the existing analog system, the IS-54 standard requires that the mobiles be capable of operating in both analog AMPS and digital voice channels. TDMA systems are able to provide capacity improvements of the order of 5–10 times that of analog FM without adding any new cell sites [Raith and Uddenfeldt, 1991]. Using bandwidth efficient  $\pi/4$ -DQPSK modulation, the IS-54 standard offers a capacity of about 50 erlangs/km<sup>2</sup>.

A CDMA-based cellular system has been developed by Qualcomm, Inc., and standardized by the Telecommunications Industry Association (TIA) as an interim standard (IS-95). This system supports a variable number of users in 1.25 MHz wide channels. Whereas the analog AMPS system requires that the signal be at least 18 dB above the interference to provide acceptable call quality, CDMA systems can operate with much larger interference levels because of their inherent

interference-resistant properties. This fact allows CDMA systems to use the same set of frequencies in every cell, which provides a large improvement in capacity. Unlike other digital cellular systems, the Qualcomm system uses a variable rate vocoder with voice activity detection, which considerably reduces the effective data rate and also the battery drain.

## Defining Terms

**Base station:** A station in the cellular radio service used for radio communication with mobile stations. They are located either in the center or edges of every cell and consist of transmitting and receiving antennas mounted on towers.

**Cochannel reuse ratio:** The ratio of the radius of a cell to the distance between the centers of two nearest cochannel cells.

**Frequency reuse:** The use of radio channels on the same carrier frequency to cover various areas that are separated from one another so that cochannel interference is not objectionable.

**Grade of service:** Likelihood that a call is blocked or delayed in a trunked system.

**Handoff:** The process of transferring a mobile station from one channel to another.

**Mobile-assisted handoff (MAHO):** A process in which a mobile, under directions from a base station, measures signal quality of specified RF channels. These measurements are forwarded to the base station upon request to assist in the handoff process.

**Mobile station:** A station in the cellular radio service intended to be used while in motion at unspecified locations. They could be either handheld personal units or units installed in vehicles.

**Mobile telephone switching office (MTSO):** Switching center that coordinates the routing of cellular calls in a service area. The MTSO connects the cellular base stations and mobiles to the PSTN.

**Transceiver:** A device capable of both transmitting and receiving radio signals.

**Trunking:** Method of accommodating a large number of users using a small number of radio channels by allocating them on a demand basis.

## References

- MacDonald, V. H. 1979. The cellular concept. *The Bell Systems Tech. J.* 58(1):15–43.
- Raith, K. and Uddenfeldt, J. 1991. Capacity of digital cellular TDMA systems. *IEEE Trans. Vehic. Technol.* 40(2):323–331.
- Rappaport, T. S. and Milstein, L. B. 1992. Effects of radio propagation path loss on DS-CDMA cellular frequency reuse efficiency for the reverse channel. *IEEE Trans. Vehic. Technol.* 41(3):231–241.
- Tekinay, S. and Jabbari, B. 1991. Handover and channel assignment in mobile cellular networks. *IEEE Comm. Mag.* November, p. 42–46.

## Further Information

A detailed treatment of cellular system design is presented in *Wireless Communications*, by T. S.

Rappaport, Prentice-Hall, 1996.

A special issue on mobile radio systems published in the *IEEE Transactions on Vehicular Technology*, May 1991, contains papers on emerging mobile radio systems and technologies. The *IEEE Communications Magazine* and the *IEEE Personal Communications Magazine* are good sources of information for the latest developments in the field.

Palais, J. C. "Optical Communications"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Optical Communications

## 131.1 Optical Communications Systems Topologies

### 131.2 Fibers

### 131.3 Other Components

### 131.4 Signal Quality

## Joseph C. Palais

Arizona State University

Electronic communications over conducting wires or by atmospheric radio transmission began in the latter part of the 19th century and was highly developed by the middle part of the 20th century. Widespread communication via beams of light traveling over thin glass fibers is a relative newcomer, beginning in the 1970s, reaching acceptance as a viable technology in the early 1980s, and continuing to evolve since then [Chaffee, 1988]. Fibers now form a major part of the infrastructure for a national telecommunications information highway in the U.S. and elsewhere.

The fundamentals of optical communications are covered in many textbooks [e.g., Palais, 1992; Keiser, 1991], which elaborate on the information presented in this chapter.]

*Optical communications* refers to the transmission of messages over carrier waves that oscillate at optical frequencies. The frequency spectrum of electromagnetic waves can range from DC to beyond  $10^{21}$  Hz. (A tabulation of units and prefixes used in this chapter appears in Tables 131.1 and 131.2.) Optical waves oscillate much faster than radio waves or even microwaves. Their characteristically high frequencies (on the order of  $3 \cdot 10^{14}$  Hz) allow vast amounts of information to be carried. An optical channel utilizing a bandwidth of just one percent of this center frequency would have an enormous bandwidth of  $3 \cdot 10^{12}$  Hz. Numerous schemes exist for taking advantage of the vast bandwidths available. These include **wavelength-division multiplexing (WDM)** and **optical frequency-division multiplexing (OFDM)**, which allocate various information channels to bands within the optical spectrum.

**Table 131.1** Units Used in Optical Communications

1 Unit	Symbol	Measure
1 Meter	m	Length
1 Second	s	Time
1 Hertz	Hz	Frequency
1 Bits per second	b/s	Data rate
1 Watt	W	Power



**Table 131.2** Commonly Used Prefixes

1 Prefix	Symbol	Multiplication Factor
1 Giga	G	$10^9$
1 Mega	M	$10^6$
1 Kilo	k	$10^3$
1 Milli	m	$10^{-3}$
1 Micro	$\mu$	$10^{-6}$
1 Nano	n	$10^{-9}$

For historical reasons optical waves are usually described by their wavelengths rather than their frequencies. The two characteristics are related by

$$\lambda = c/f \quad (131.1)$$

where  $f$  is the frequency in hertz,  $\lambda$  is the wavelength, and  $c$  is the velocity of light in empty space ( $3 \cdot 10^8$  m/s). A frequency of  $3 \cdot 10^{14}$  Hz corresponds to a wavelength of  $10^{-6}$  meters (or  $1 \mu\text{m}$ , often called a *micron*). Wavelengths of interest for optical communications are on the order of a micron.

The Gaussian beam profile is common in optical systems. Its intensity pattern in a plane transverse to the direction of wave travel is described by

$$I = I_0 e^{-2(r/w)^2} \quad (131.2)$$

where  $r$  is the polar radial coordinate and the factor  $w$  is called the *spot size*. This is the light pattern emitted by many lasers and is (approximately) the intensity distribution in a **single-mode fiber**.

Optical transmission through the atmosphere is possible, but it suffers serious liabilities. A fundamental limit is imposed by diffraction of the light beam as it propagates away from the transmitting station. The full divergence angle of a Gaussian beam is given by

$$\theta = 2\lambda/\pi w \quad (131.3)$$

Because of the continual beam enlargement with propagation distance, the amount of light captured by a finite receiving aperture diminishes with increasing path length. The resultant low received power limits the distances over which atmospheric optical systems are feasible.

The need for an unobstructed line-of-sight connection between transmitter and receiver and a clear atmosphere also limits the practicality of atmospheric optical links. Although atmospheric applications exist, the vast majority of optical communications is conducted over glass fiber.

A key development leading to fiber communications was the demonstration of the first **laser** in 1960. This discovery was quickly followed by plans for numerous laser applications. Progress on empty space optical systems in the 1960s laid the groundwork for fiber communications in the 1970s. The first low-loss optical waveguide, the glass fiber, was produced in 1970. Soon after, multimode fiber transmission systems were being designed, tested, and installed. The mass-production commercialization of the single-mode fiber in 1983 brought about the fiber revolution in the communications industry. Fibers are practical for a range of path lengths, from under a meter to as long as required on the earth's surface and beneath its oceans (e.g., almost 10000 kilometers for transpacific links).

Fiber systems are limited in length by the bandwidth of their components (a fiber's bandwidth decreases with length) and by component losses (a fiber's loss increases with length). Loss is usually expressed in the decibel scale, which compares two power levels and is defined by

$$\text{dB} = 10 \log P_2/P_1 \quad (131.4)$$

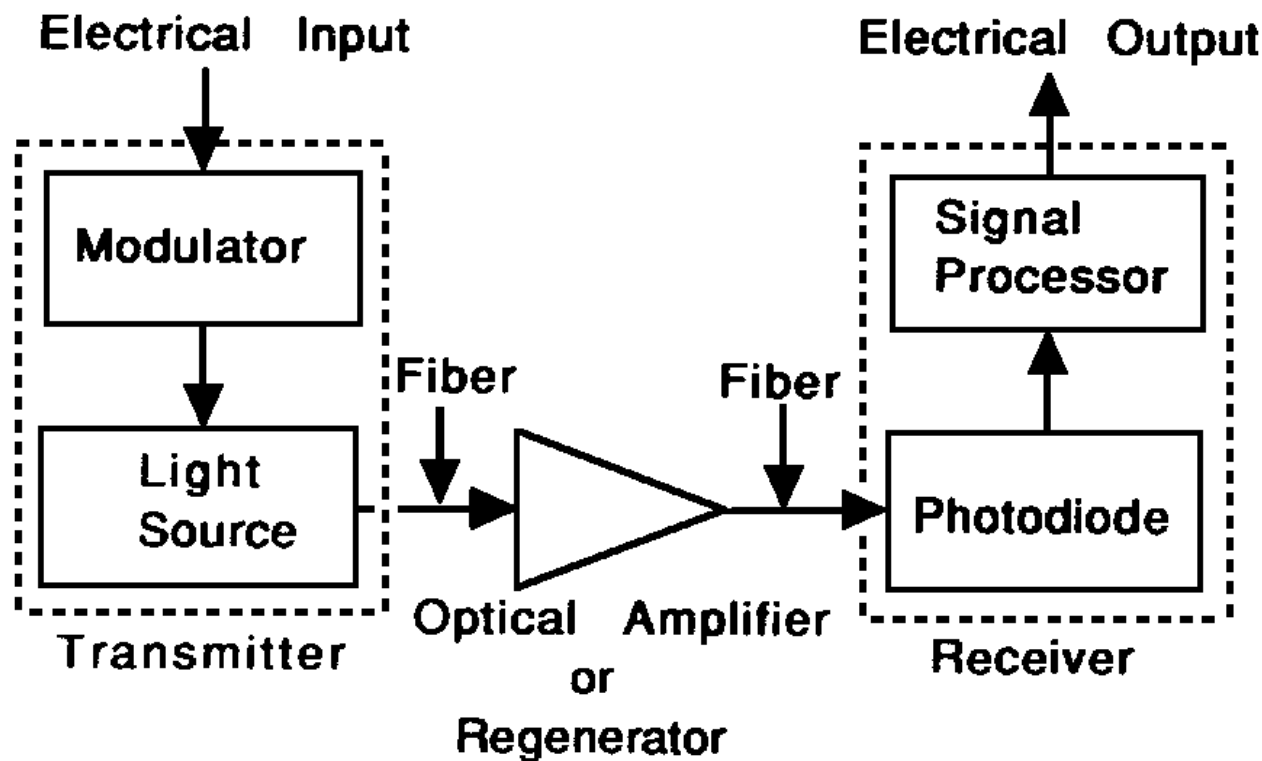
## 131.1 Optical Communications Systems Topologies

---

Fibers find their greatest use in telephone networks, local-area networks, and cable television networks. They are also useful for short data links, closed-circuit video links, and elsewhere.

A block diagram of a point-to-point fiber optical communications system appears in [Fig. 131.1](#). This diagram represents the architecture typical of the telephone network. The fiber telephone network is digital, operating at levels indicated in [Table 131.3](#). Fiber systems often transmit at multiples of the rates listed in the table. For example, rates in the range of 2.5 Gb/s are commonplace. At this rate, several thousand digitized voice channels (each operating at 64 kb/s) can be transmitted along a single fiber using time-division multiplexing (TDM). Higher-rate fiber systems (tens of gigabits per second) will be developed as technology evolves.

**Figure 131.1** Point-to-point fiber transmission system.



**Table 131.3** Digital Transmission Rates

Designation	Data Rate (Mb/s)
DS-1	1.5444
DS-3	44.736
DS-4	274.175
OC-1	51.84
OC-3	155.52
OC-12	622.08
OC-24	1244.16
OC-48	2488.32

The DS levels refer to U.S. telephone signaling rates, whereas the OC levels are those of the SONET standard.

Because fiber cables may contain more than one fiber (in fact, some cables contain hundreds of fibers), a single cable may carry hundreds of thousands of voice channels.

Telephone applications may be broken down into several distinct areas: transmission between telephone exchanges, long-distance links, undersea links, and distribution in the local loop (i.e., to subscribers). Although similarities exist between these systems, the requirements are somewhat different. Between telephone exchanges, large numbers of calls must be transferred over moderate distances. Because of the moderate path lengths, optical amplifiers or regenerators are not required.

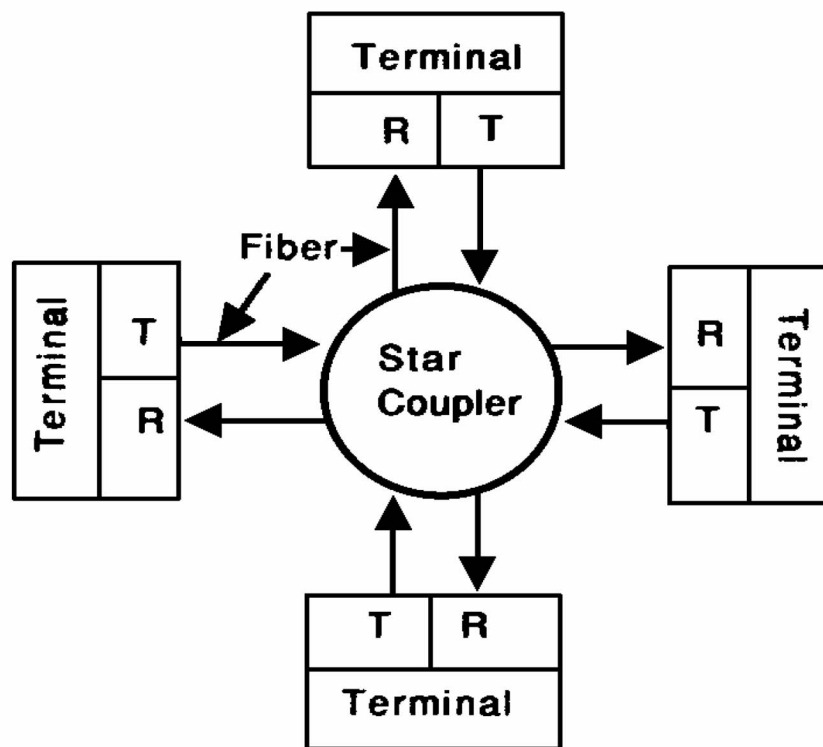
Long-distance links, such as between major cities, require signal boosting of some sort (either regenerators or optical amplifiers). Undersea links (such as transatlantic or transpacific) require multiple boosts in the signal because of the long path lengths involved [Thiennot *et al.*, 1993].

Fiber-to-the-home (for broadband services, such as cable television distribution) does not involve long path lengths but does include division of the optical power in order to share fiber transmission paths over all but the last few tens of meters into the subscriber's premises. In many subscriber networks, the fibers terminate at optical-to-electrical conversion units located close to the subscriber. From that point, copper wires transmit the signals over the remaining short distance to the subscriber. Because of the power division, optical amplifiers are needed to keep the signal levels high enough for proper reception.

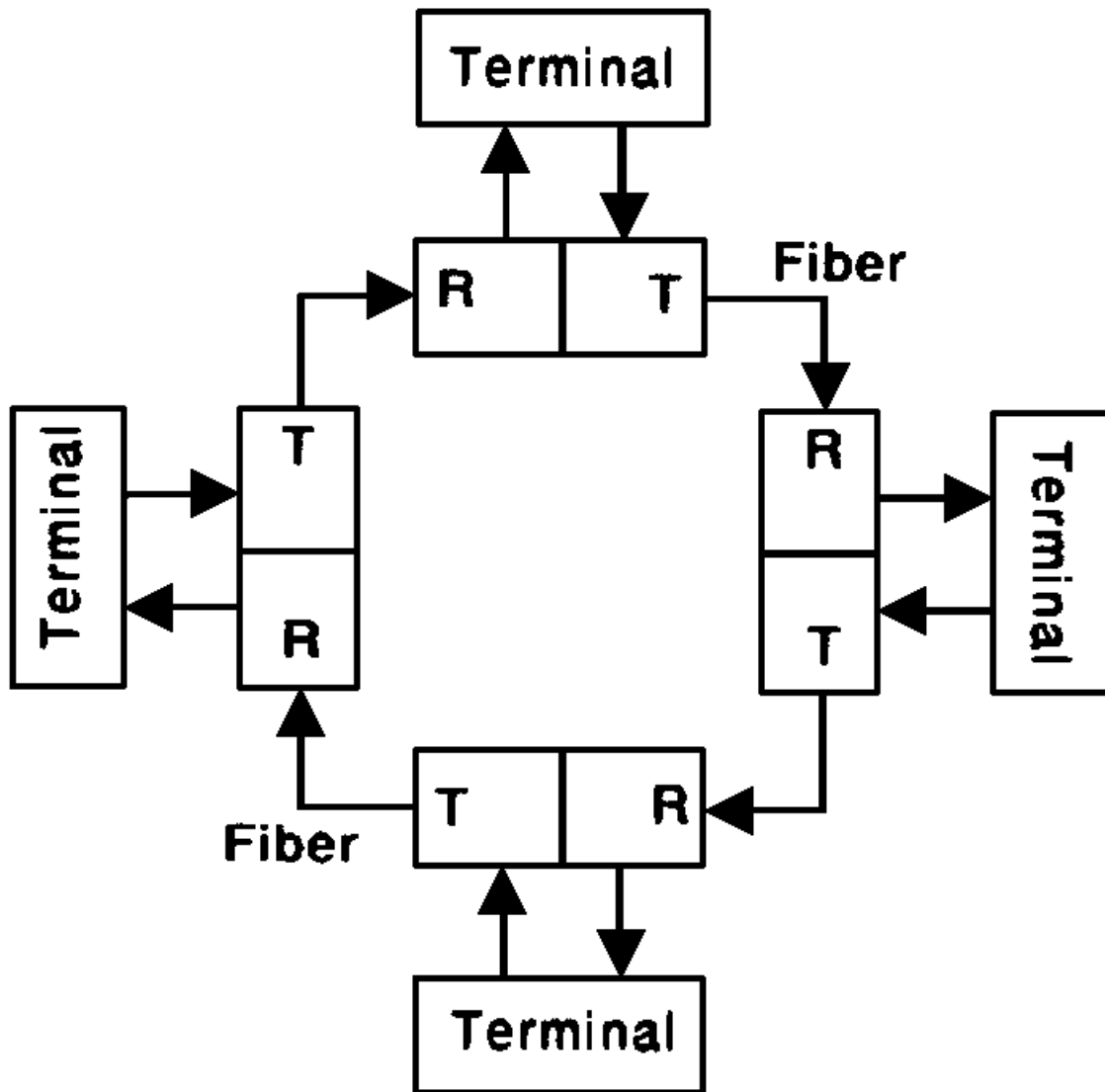
Cable television distribution remained mostly coaxial cable based for many years, which was partly due to the distortion produced by optical analog transmitters. Production of highly linear laser diodes [such as the distributed feedback (DFB) laser diode] permitted the design of practical television fiber distribution links.

Some applications, such as local-area networks (LANs), require distribution of the signals over shared transmission fiber. Topologies include the passive star, the active star, and the ring network [Hoss, 1990]. The passive star and the active ring are illustrated in Figs. 131.2 and 131.3. LANs are *broadcast systems* that allow all terminals to receive messages sent from any one transmitting terminal.

**Figure 131.2** Star topology. Blocks T and R are, respectively, the optical transmitter and optical receiver.



**Figure 131.3** Ring topology. Blocks T and R are, respectively, the optical transmitter and optical receiver.



The star coupler distributes optical signals from any input port to all output ports. Its loss is simply

$$L = 10 \log (1/N) \quad (131.5)$$

where  $N$  is the number of output ports on the star coupler.

In the ring each node acts as a regenerator. The ring is simply a series of connected point-to-point networks.

The major components found in optical communications systems are modulators, light sources, fibers, photodetectors, connectors, splices, star couplers, regenerators, and optical amplifiers.

## 131.2 Fibers

Silica glass fibers make up the majority of optical transmission lines. The wavelength regions around 0.85, 1.3, and 1.55  $\mu\text{m}$  have been heavily utilized because of their low losses. These regions are called, respectively, the *first*, *second*, and *third windows*. Properties of fibers in these windows are tabulated in Table 131.4.

**Table 131.4** Typical Fiber Properties

Window	Wavelength (nm)	Loss (dB/km)	Dispersion (ps/nm · km)
First	800–900	3	120
Second	1290–1330	0.4	Nearly zero
Third	1520–1570	0.25	15

**Dispersion** causes pulse spreading, leading to intersymbol interference. This interference limits the fiber's allowable data rate. The amount of pulse spreading is given by

$$\Delta\tau = ML\Delta\lambda \quad (131.6)$$

where  $M$  is the dispersion (values are given in Table 131.4),  $L$  is the fiber length, and  $\Delta\lambda$  is the spectral width of the light source emission. Dispersion is created by the material and the waveguide structure, both of which have a wavelength-dependent pulse velocity.

Multimode fibers allow many modes to simultaneously traverse the fiber. This produces *multimode distortion* (resulting in additional pulse spreading) because the various modes travel at different speeds. For this reason multimode fibers can only be used for applications in which the product of data rate (or modulation frequency) and path length is not high.

Because single-mode fibers eliminate multimode spreading, they have larger bandwidths and are used for all long-distance, high-rate links.

Multimode fibers have relatively high loss and large dispersion in the 850 nm first window region. Applications are, therefore, restricted to moderately short path lengths (typically less than a kilometer). Components in this window tend to be cheaper than those operating in the 1300 nm and 1550 nm windows.

The 1300 nm second window exhibits low losses and nearly zero dispersion. Single-mode, nonrepeated paths up to 70 km or so are attainable in this window. Here, multimode fiber is feasible for modest lengths required by local-area networks and campus-based networks and single-mode fiber for longer point-to-point links.

Fiber systems operating in the 1550 nm third window cover the highest rates and longest unamplified, unrepeated distances. Unamplified lengths over 100 km are possible. Typically, only single-mode fibers are used in this region.

## 131.3 Other Components

---

Most systems utilize semiconductor **laser diodes** (LD) or **light emitting diodes** (LED) for the light source. These sources are typically modulated by controlling their driving currents. The conversion from current,  $i$ , to optical power,  $P$ , is given by

$$\begin{aligned} P &= a_1(i - I_{th}), & i > I_{th} \\ P &= 0, & i < I_{th} \end{aligned} \quad (131.7)$$

where  $a_1$  is a constant and  $I_{th}$  is the turn-on threshold current for the diode. The threshold current for LEDs is zero and in the range of a few milliamperes to a few tens of milliamperes for LDs. For example,  $a_1$  may be 0.1 mW/mA. Thus, the optical power waveform is a replica of the modulation current if the light source is operated above its threshold current.

Laser diodes are more coherent (i.e., they have smaller spectral widths) than LEDs and thus produce less dispersion. In addition, LDs can be modulated at higher rates (above 10 GHz), whereas LEDs are limited to rates of just a few hundred MHz. LEDs have the advantage of lower cost and simpler circuitry requirements.

The photodetector converts the light beam back into an electrical signal. Semiconductor **PIN photodiodes** and **avalanche photodiodes** are normally used. The conversion for the PIN diode is given by the linear equation

$$i = \rho P \quad (131.8)$$

where  $i$  is the detected current,  $P$  is the incident optical power, and  $\rho$  is the photodetector's *responsivity*. Typical values of the responsivity are on the order of 0.5 A/W. The avalanche photodiode response follows this same equation but includes an amplification factor that can be as high as several hundred.

## 131.4 Signal Quality

---

Signal quality is measured by the **signal-to-noise ratio** (S/N) in analog systems and by the **bit-error rate** (BER) in digital links. High-quality analog video links may require signal-to-noise ratios on the order of a hundred thousand (50 dB) or more. Good digital systems operate at error rates of  $10^{-9}$  or better.

In a thermal noise-limited system the probability of error,  $P_e$ , (which is the same as the bit-error rate) is

$$P_e = 0.5 - 0.5 \operatorname{erf} (0.354\sqrt{S/N}) \quad (131.9)$$

where erf is the error function, tabulated in many references. An error rate of  $10^{-9}$  requires a signal-to-noise ratio of nearly 22 dB ( $S/N = 158.5$ ).

## Defining Terms

**Avalanche photodiode:** Semiconductor photodetector having internal gain.

**Bit-error rate:** Probability of error.

**Dispersion:** Wavelength-dependent pulse group velocity caused by the material and the fiber structure. Results in pulse spreading due to the nonzero spectral emission widths of the light source.

**Laser:** Source producing highly coherent light.

**Laser diode:** Semiconductor laser. Spectral emission widths typically in the range of 1 to 5 nm. Special devices available with spectral widths of 0.1 nm or less.

**Light emitting diode:** Semiconductor emitter having spectral emission widths much larger than those of the laser diode. Typical emission widths are in the range of 20 to 100 nm.

**Optical frequency-division multiplexing (OFDM):** Transmission of closely spaced carrier wavelengths (a few tenths of a nm, or less) along a single fiber. Demultiplexing usually requires an optical coherent receiver.

**PIN photodiodes:** Semiconductor photodetector.

**Signal-to-noise ratio:** Ratio of signal power to noise power.

**Single-mode fiber:** Fiber that allows only one mode to propagate; has larger bandwidth, allowing greater data rate, than do multimode fibers.

**Wavelength-division multiplexing (WDM):** Transmission of multiple-carrier wavelengths simultaneously to increase the capacity of a single fiber. Typically, the individual carriers are widely spaced (from a few tens of nm to several hundred nm), permitting just a few independent channels on the fiber.

## References

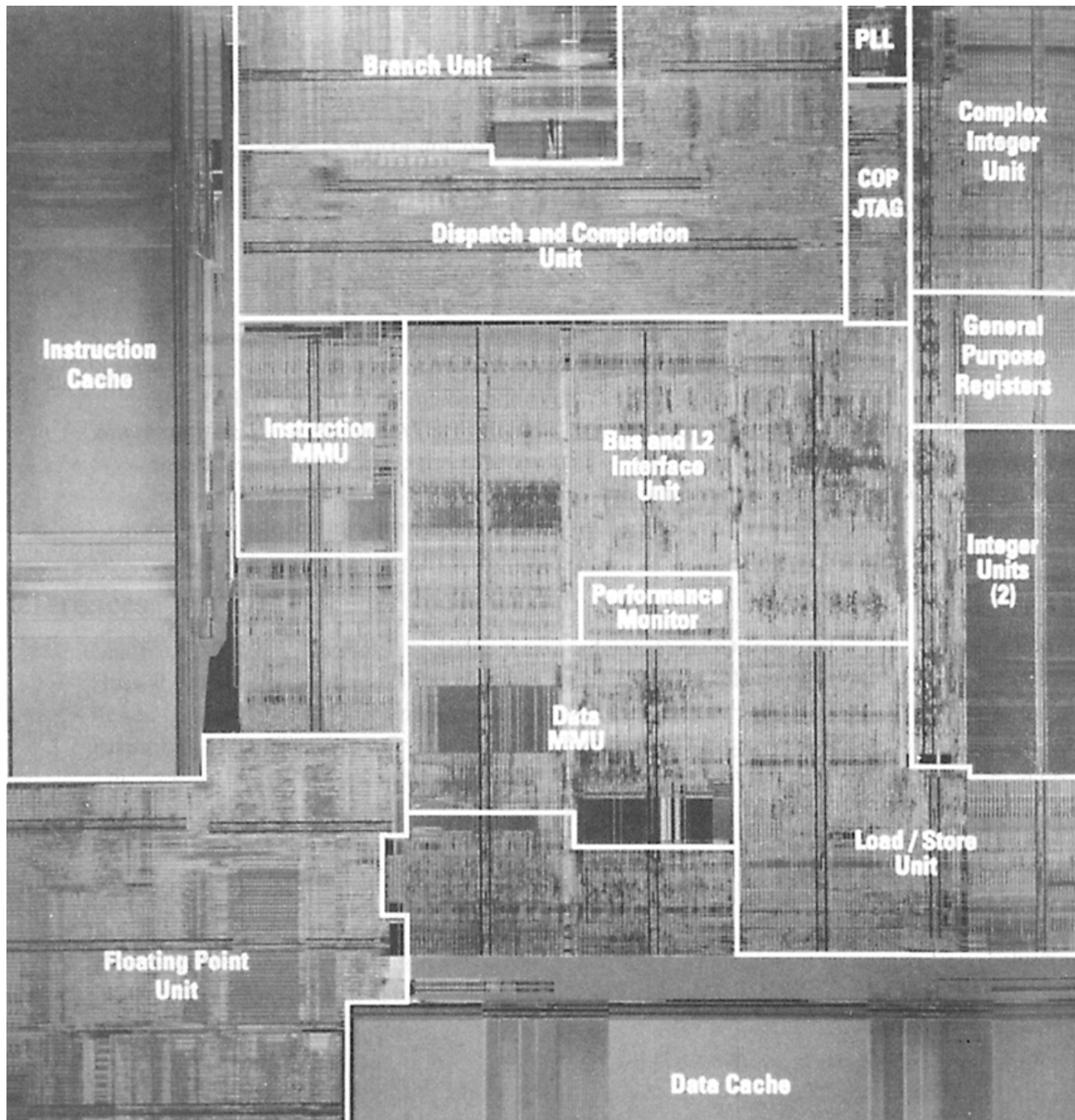
- Chaffee, C. D. 1988. *The Rewiring of America*. Academic Press, Orlando, FL.
- Hoss, R. J. 1990. *Fiber Optic Communications*. Prentice Hall, Englewood Cliffs, NJ.
- Keiser, G. 1991. *Optical Fiber Communications*. McGraw-Hill, New York.
- Palais, J. C. 1992. *Fiber Optic Communications*. Prentice Hall, Englewood Cliffs, NJ.
- Thiennot, J., Pirio, F., and Thomine, J.-B. 1993. Optical undersea cable systems trends. *Proc. of the IEEE*. 81(11):1610–1611.

## Further Information

Information on optical communications is included in several professional society journals. These include *IEEE Journal of Lightwave Technology* and *IEEE Photonics Technology Letters*. Valuable information is also contained in several trade magazines such as *Lightwave* and *Laser Focus World*.



Oklobdzija, V. G. "Computers"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



Microphotograph of IBM PowerPC™ 620, a 64-bit super-scalar processor containing 7 million transistors and capable of issuing four instructions in a single cycle built in 0.5 $\mu$  CMOS technology. The processor has 32 Kbytes of instruction and data cache memory, and employs register renaming and branch prediction mechanisms. It runs at 100 MHz and achieves 250 SPEC integer and 300 SPEC floating point operations. (Photo by Tom Way and courtesy of IBM.)

# XXI

## Computers

---

**Vojin G. Oklobdzija**  
*University of California, Davis*

**132 Computer Organization: Architecture** *V. G. Oklobdzija*

Instruction Set • RISC Architecture

**133 Operating Systems** *L.-F. Cabrera*

Typical Services • General Flow of Control • Structure • Communication • Advanced Data Management Services

**134 Programming Languages** *D. M. Volpano*

Principles of Programming Languages • Program Verification • Programming Language Paradigms

**135 Input/Output Devices** *R. Freitas*

Input/Output Subsystem • I/O Devices

**136 Memory and Mass Storage Systems** *P. M. Chen*

Aspects of Storage Devices • Types of Storage Devices • Storage Organization

A COMPUTER IS A SYSTEM capable of solving various scientific and nonscientific problems; storing, retrieving, and manipulating data; communicating and sharing data; controlling processes; and interacting with the surrounding environment.

Today a typical computer consists of a complex electronic system containing a massive amount of components which are very highly integrated. Even the simplest computer today is far more complex than the computers from the not-so-distant past and may contain millions or hundreds of millions of transistors.

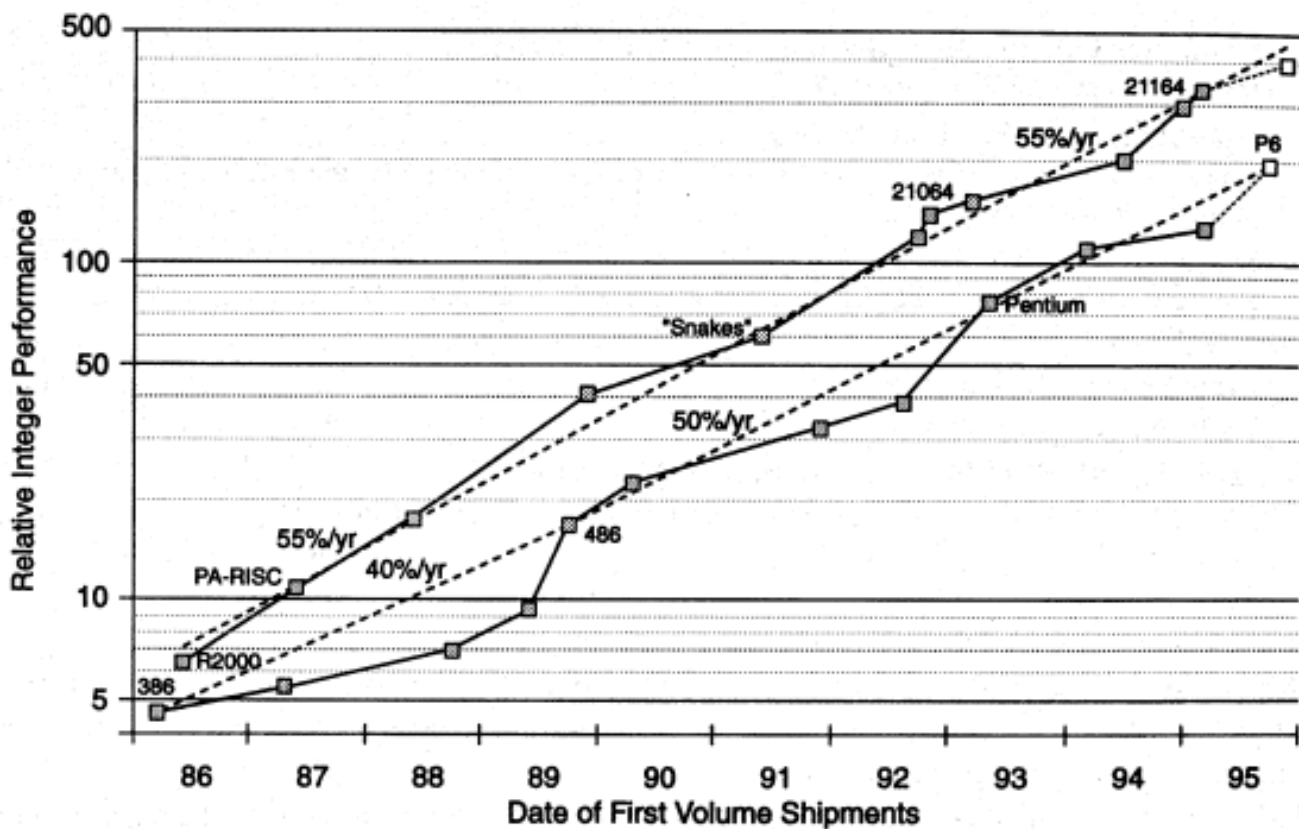
Computers emerged from complex digital systems and controllers in areas where the behavioral specification for such a system could be satisfied with some sort of general-purpose digital system. One of the first minicomputers (18-bit PDP-4) was built as a generalization of an atomic plant controller. The first microprocessor, Intel 4004 (S. Shima), was originally commissioned as a generalized calculator chip.

Because of the computers versatility and general-purpose orientation, there is hardly any place today that does not contain a computer in some form. There are two reasons for this:

1. The computer's structure and organization is general, and therefore it can be easily customized in many different forms.
2. Because of their general structure, computers can be mass-produced, which keeps their cost down.

The 4- and 8-bit microcontrollers, which represent over 90% of the microprocessors sold, generally sell for well under a dollar, which is due to the fact that they can be produced in very large quantities.

Based on their power, computers are traditionally classified in four major categories: personal computers, midrange or workstations, mainframes, and supercomputers; however, the boundaries between these categories are blurred. The reason is that the same technology is being used for all four categories, although mainframes and supercomputers are resorting to bipolar and gallium arsenide. Therefore, the performance increases are being achieved mainly through the improvements in technology. The performance is roughly doubling every two years, as shown in Fig. 1. This is not the result of changes in the architecture because the architecture issue has been settled around RISC, which has clearly demonstrated its advantage over CISC (RISC stands for Reduced Instruction Set Computer, while CISC stands for Complex Instruction Set Computer). RISC is characterized by simple instructions in the instruction set, which is constructed to fit the machine pipeline in such a way that one instruction can be issued in every cycle. CISC is characterized by complex instructions which have grown mainly out of the microprogramming design style of the computer.



The performance of leading RISC processors has increased steadily at a rate of about 55% per year since 1986. The gap between the trend lines for RISC and CISC performance is now slightly more than 2× and is increasing very slowly. (Reprinted with permission from the 1/23/95 issue of *Microprocessor Report*. Copyright 1995 MicroDesign Resources, Sebastopol, CA.)

The so-called supercomputers are the ones that have been driving all of the advanced concepts in architecture as well as driving the technology to its limits. Though several computers were designated as "supercomputers" in the past, such as IBM-Stretch, IBM System 360 Model 91, and Control Data's CDC 6600, the real era of supercomputers started with CRAY-1, engineered and designed by Seymour Cray. Probably the best description of a supercomputer is a design where performance is the prime objective and the cost is ignored. They are manufactured in small numbers for very special customers requiring very high performance who are willing to pay a

premium cost for that performance. CRAY-1 was introduced in 1976 and had a clock cycle of 12.5 n. The latest CRAY is the CRAY-4, built in gallium arsenide technology with a cycle time of 1 n and capable of achieving 256 gigaflops in a 128-processor configuration. CRAY-4 is truly state of the art in almost all aspects of engineering.

Today, typical high-performance computer systems employ multiple processors in various arrangements. There has been an ongoing effort to parallelize the execution of the programs and use a number of the relatively inexpensive processors in order to achieve a high processing rate. These efforts have achieved limited success. A number of parallel machines have been introduced with varying degrees of success. Although they can be divided in several categories, most of the machines introduced fall into one of two structures: SIMD and MIMD. SIMD, which stands for Single Instruction Multiple Data, is characterized by the execution of one instruction at a time, operating on an array of data elements in parallel. A typical example of SIMD architecture is the so-called Connection Machine, CM-1, introduced by Connection Machines of Cambridge, MA, in the first half of 1984. This machine is characterized by an array of up to 64K processors divided in four quadrants containing 16K processors each. CM-1 has been superseded by CM-2, 3, and CM-5. The operations of the processors are controlled by the same instruction issued from the central instruction unit. Another example of parallel SIMD architecture is the IBM GF-11 machine, capable of a peak execution rate of 11 billion floating-point operations per second.

The current trend is toward distributed computing on a large, even global, scale. This involves a network of workstations connected via high-bandwidth, low-latency networks acting as a computing platform. The goal is to take advantage of a large resource pool of workstations comprising hundreds of gigabytes of memory, terabytes of disk space, and hundreds of gigaflops of processing power that is often idle. This new paradigm in computing is expected to impact the fundamental design techniques for large systems and their ability to solve large problems, serve a large number of users, and provide a computing infrastructure.

Oklobdzija, V. G. "Computer Organization: Architecture"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Computer Organization: Architecture

---

## 132.1 Instruction Set

Addressing Modes • Data Representation Formats

## 132.2 RISC Architecture

**Vojin G. Oklobdzija**

*University of California, Davis*

On 7 April 1964 the term **computer architecture** was first defined by Amdahl, Blaauw, and Brooks of IBM Corporation in the paper announcing the IBM System/360 computer family [1964]. On that day IBM Corporation introduced, in the words of an IBM spokesman, "the most important product announcement that this corporation has made in its history." There were six models introduced originally, ranging in performance from 25 to 1. Six years later this performance range was increased to about 200 to 1. This was the key feature that prompted IBM's effort to design an architecture for a new line of computers that are to be code-compatible with each other. The recognition that architecture and **computer implementation** could be separated and that one need not imply the other led to establishment of a common System/360 machine architecture implemented in the range of models.

In their milestone paper Amdahl, Blaauw, and Brooks identified three interfaces: architecture, implementation, and realization. They defined computer *architecture* as the attributes of a computer seen by the machine language programmer, as described in the **principles of operation**. IBM referred to the principles of operation as a definition of the machine that enables the machine language programmer to write functionally correct, time-independent programs that run across a number of implementations of that particular architecture. Therefore, the architecture specification covers all functions of the machine that are observable by the program [Siewiorek *et al.*, 1982]. On the other hand, the principles of operation are used to define the functions that the implementation should provide. In order to be functionally correct the implementation must conform to the principles of operation. Accordingly, for the first time in the history of computer development, IBM has separated *machine definition* from *machine implementation*, thus enabling them to bring several machine implementations in a wide price and performance range that has reached—22 years after the introduction of System/360—2000 to 1.

The principles of operation define computer architecture, which includes:

- Instruction set
- Instruction format
- Operation codes

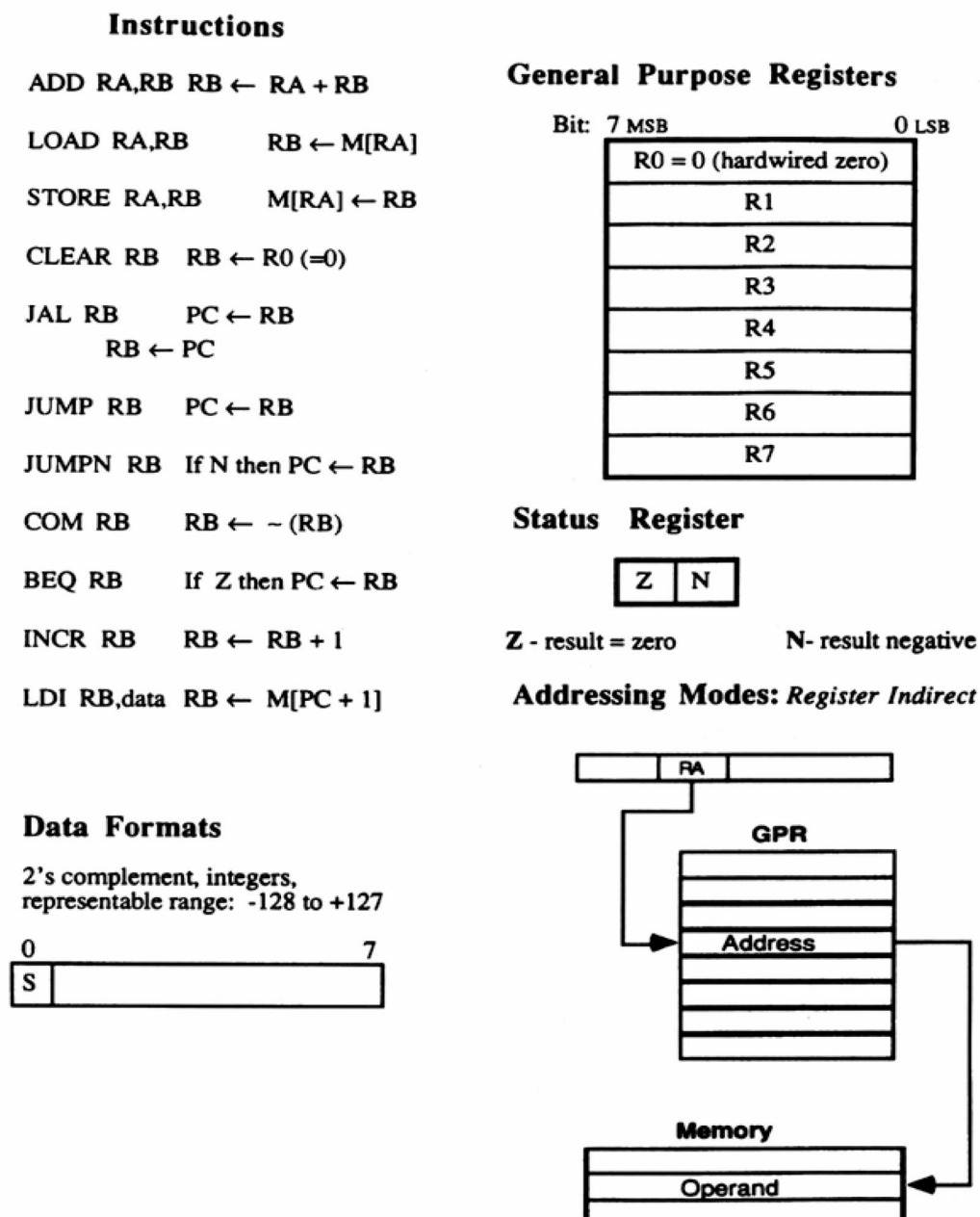


- Addressing modes
- All registers and memory locations that may be directly manipulated or tested by a machine language program
- Formats for data representation

**Machine implementation** was defined as the actual system organization and hardware structure encompassing the major functional units, data paths, and control. **Machine realization** includes issues such as logic technology, packaging, and interconnections.

An example of simple architecture of an 8-bit processor that uses 2s complement representation to represent integers and contains 11 instructions is shown in Fig. 132.1. The figure contains all of the necessary information for the architecture to be defined.

**Figure 132.1** Example of a minimal architecture: PRISC.





Separation of the machine architecture from implementation enables several embodiments of the same architecture to be built. Operational evidence proved that architecture and implementation could be separated and that one need not imply the other. This separation made it possible to transfer programs routinely from one model to another and expect them to produce the same result, which defined the notion of **architectural compatibility**. Implementation of the whole line of computers according to a common architecture requires unusual attention to details and some new procedures, which are described in the architecture control procedure. The design and control of system architecture is an ongoing process whose objective is to remove ambiguities in the definition of the architecture and, in some cases, to adjust the functions provided.

Definition of an architecture facilitated future development and introduction of not only new models but new **upwardly compatible architectures**. The architecture is upwardly compatible if the user's programs written for the old architecture run efficiently on the new models without modifications to the program. The limitations to upward compatibility are (1) that new systems have the same or equivalent facilities, and (2) that the programs have no time dependence, use only model-independent functions defined in the principles of operation, and do not use unassigned formats and operation codes [Case and Padegs, 1978]. An example of upward compatibility is IBM System/360, introduced in June 1970.

## 132.1 Instruction Set

---

Instruction set defines a basic set of operations, as specified by the architecture, that a particular implementation of that architecture is required to perform. An *instruction* of the instruction set defines an atomic operation that may alter data or the machine state or may perform an I/O operation. In terms of the operation performed, instructions of the instruction set are broadly classified in one of the four general categories:

1. Instructions performing transformation of data
2. Instructions altering the program flow
3. Instructions performing data movement
4. System instructions

The first category includes instructions performing arithmetic and logical operations. The operations can be arithmetic, string, logical, or floating point. They are performed in the appropriate functional units of the particular implementation of the architecture.

Instructions affecting the flow of the program and/or machine state are branches, calls, and returns as well as loop control instructions.

The third category of instructions performs data movement across various functional units of the machine. Examples of such instructions are the load instruction, which loads a content of a memory location to a particular register in the general purpose register file (GPR), and the store instruction, which does the opposite. The move instruction moves a block of data from one memory location to another, or to and from the Stack or GPR.

The system instructions change the system's mode and are not generally visible by the programmer that programs in the *problem state*. Problem state is the domain of the machine visible to a programmer executing a general purpose program—as opposed to the *system state*, which is visible to the operating system.

An example of the instruction set specified in the IBM System/360 architecture is given in [Fig. 132.2](#).

**Figure 132.2** IBM System/360 instruction set. (Source: Blaauw, G. A. and Brooks, F. P. 1964. The structure of system/360. *IBM Syst. J.* 3(2):119–135. Copyright 1964 by International Business Machines Corporation. With permission.)

**RR Format**

Branching and status switching 0000xxxx		Fixed-point fullword and logical 0001xxxx		Floating-point long 0010xxxx		Floating-point short 0011xxxx	
0000		LPR	LOAD POSITIVE	LPDR	LOAD POSITIVE	LPER	LOAD POSITIVE
0001		LNR	LOAD NEGATIVE	LNR	LOAD NEGATIVE	LNR	LOAD NEGATIVE
0010		LTR	LOAD AND TEST	LDR	LOAD AND TEST	LTR	LOAD AND TEST
0011		LCR	LOAD COMPLEMENT	LDR	LOAD COMPLEMENT	LTR	LOAD COMPLEMENT
0100	SPM SET PROGRAM MASK	NR	AND	HDR	HALVE	HER	HALVE
0101	BALR BRANCH AND LINK	CLR	COMPARE LOGICAL				
0110	BCTR BRANCH ON COUNT	OR	OR				
0111	BCR BRANCH/CONDITION	XR	EXCLUSIVE OR				
1000	SSK SET KEY	LR	LOAD	LDR	LOAD	LER	LOAD
1001	ISK INSERT KEY	CR	COMPARE	CDR	COMPARE	CER	COMPARE
1010	SVC SUPERVISOR CALL	AR	ADD	ADR	ADD N	ALR	ADD N
1011		SR	SUBTRACT	SDR	SUBTRACT N	SER	SUBTRACT N
1100		MR	MULTIPLY	MDR	MULTIPLY	MER	MULTIPLY
1101		DR	DIVIDE	DDR	DIVIDE	DER	DIVIDE
1110		ALR	ADD LOGICAL	AWR	ADD U	AUR	ADD U
1111		SLR	SUBTRACT LOGICAL	SWR	SUBTRACT U	SUR	SUBTRACT U

**RX Format**

Fixed-point halfword and branching 0100xxxx		Fixed-point fullword and logical 0101xxxx		Floating-point long 0110xxxx		Floating-point short 0111xxxx	
0000	STH STORE ADDRESS	ST	STORE	STD	STORE	STE	STORE
0001	LA LOAD ADDRESS						
0010	STC STORE CHARACTER						
0011	IC INSERT CHARACTER						
0100	EX EXECUTE	N	AND				
0101	BAL BRANCH AND LINK	CL	COMPARE LOGICAL				
0110	BCT BRANCH ON COUNT	O	OR				
0111	BC BRANCH/CONDITION	X	EXCLUSIVE OR				
1000	LH LOAD	L	LOAD	LD	LOAD	LE	LOAD
1001	CH COMPARE	C	COMPARE	CD	COMPARE	CE	COMPARE
1010	AH ADD	A	ADD	AD	ADD N	AE	ADD N
1011	SH SUBTRACT	S	SUBTRACT	SD	SUBTRACT N	SE	SUBTRACT N
1100	MH MULTIPLY	M	MULTIPLY	MD	MULTIPLY	ME	MULTIPLY
1101		D	DIVIDE	DD	DIVIDE	DE	DIVIDE
1110	CYD CONVERT-DECIMAL	AL	ADD LOGICAL	AW	ADD U	AU	ADD U
1111	CVB CONVERT-BINARY	SL	SUBTRACT LOGICAL	SW	SUBTRACT U	SU	SUBTRACT U

**RS, SI Format**

Branching status switching and shifting 1000xxxx		Fixed-point logical and input/output 1001xxxx		1010xxxx		1011xxxx	
0000	SSM SET SYSTEM MASK	STM	STORE MULTIPLE				
0001		TM	TEST UNDER MASK				
0010	LPSW LOAD PSW	MVI	MOVE				
0011	DIAGNOSE	TS	TEST AND SET				
0100	WRD WRITE DIRECT	NI	AND				
0101	RDD READ DIRECT	CLI	COMPARE LOGICAL				
0110	BXH BRANCH/HIGH	OI	OR				
0111	BXLE BRANCH/LOW-EQUAL	XI	EXCLUSIVE OR				
1000	SRL SHIFT RIGHT SL	LM	LOAD MULTIPLE				
1001	SLL SHIFT LEFT SL						
1010	SRA SHIFT RIGHT S						
1011	SLA SHIFT LEFT S						
1100	SRDL SHIFT RIGHT DL	SIO	START I/O				
1101	SLDL SHIFT LEFT DL	TIO	TEST I/O				
1110	SRDA SHIFT RIGHT D	HIO	HALT I/O				
1111	SLDA SHIFT LEFT D	TCH	TEST CHANNEL				

**SS Format**

1100xxxx		Logical 1101xxxx		1110xxxx		Decimal 1111xxxx	
0000		MVN	MOVE NUMERIC			MVO	MOVE WITH OFFSET
0001		MVC	MOVE			PACK	PACK
0010		MVZ	MOVE ZONE			UNPK	UNPACK
0011		NC	AND				
0100		CLC	COMPARE LOGICAL				
0101		OC	OR				
0110		XC	EXCLUSIVE OR				
0111							
1000						ZAP	ZERO AND ADD
1001						CP	COMPARE
1010						AP	ADD
1011						SP	SUBTRACT
1100		TR	TRANSLATE			MP	MULTIPLY
1101		TRT	TRANSLATE AND TEST			DP	DIVIDE
1110		ED	EDIT				
1111		EDMK	EDIT AND MARK				

NOTE: N = NORMALIZED DL = DOUBLE LOGICAL S = SINGLE  
SL = SINGLE LOGICAL U = UNNORMALIZED D = DOUBLE

We can further classify instructions in terms of the number of *explicit operands*, *operand locations*, and *type* and *size* of the operands.

Instruction architecture that specifies no explicit operands is better known as *stack architecture*. In stack architecture all operations are performed on the data that are on the top of the stack. Examples of stack architecture are HP 3000/70 by Hewlett-Packard and B5500 by Burroughs. In the **accumulator architecture** all of the operations are performed between the operand specified in the instruction and the *accumulator*, which is a special register. An example of accumulator architecture is the PRISC processor shown in [Fig. 132.1](#). One of the well-known accumulator-based architectures is PDP-8, by Digital Equipment Corporation. Almost all of the modern machines have a repertoire of available general purpose registers whose numbers range from 16 to 32 and in some cases even more than 32 (SPARC). The number of operands explicitly specified in the instructions of a modern architecture today can be two or three. In the case of three operands an instruction explicitly specifies the location of both operands and the location where the result is to be stored. In some architectures (IBM System/360) only two operands are *explicitly specified* in order to save the bits in the instruction. As a consequence, one of the operands is always replaced by the result and its content is destroyed. This type of instruction is sometimes referred to as *diadic instruction*.

In terms of the operand locations, instructions can be classified as:

Register to register (or R-R) instructions

Memory to register (R-M) instructions

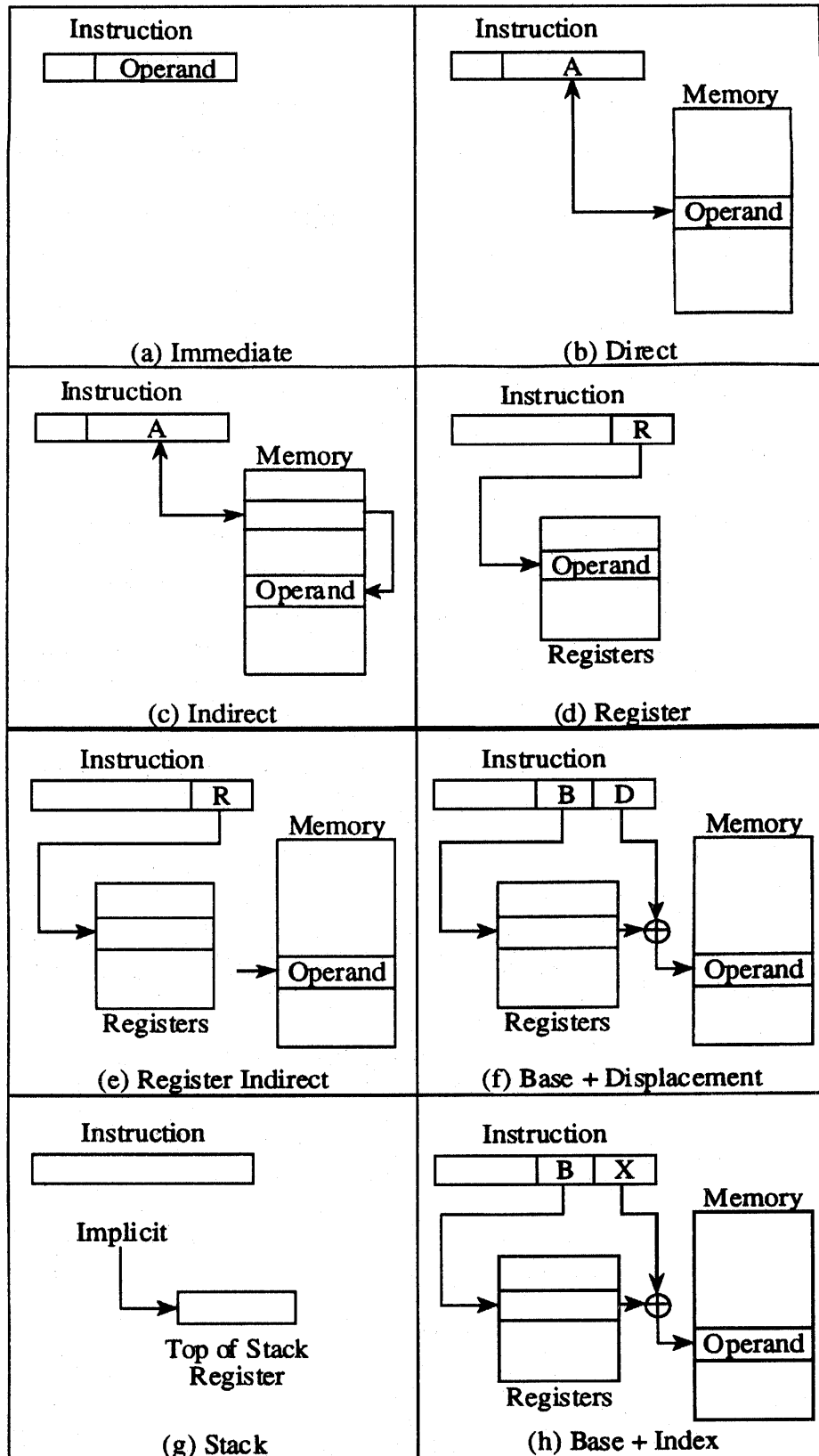
Memory to memory (M-M) instructions

The addresses of the operands are specified within the instruction. From the information contained in the particular operand field of the instruction, the address of the particular operand can be formed in various ways, described in the section that follows.

## Addressing Modes

The way in which the address of the operand is formed depends on the location of the operand as well as choices given in the instruction architecture. It is obvious that, in the case of stack or accumulator architecture, the address of the operand is implied and there is no need to specify the address of the operand. If the operand is in one of the GPRs the operand field in the instruction contains the number (address) of that particular register. This addressing mode is known as *register direct addressing* and is one of the simplest ways of pointing to the location of the operand. The addressing of an operand can be even simpler, in the case where the operand is contained within the instruction. This mode is called the *immediate addressing mode*. The location pointed to by the address formed from the information contained in the operand field of an instruction can contain the operand itself or an address of the operand. The latter case is referred to as *indirect addressing*. Examples of several ways of forming an address of the operand are given in [Fig. 132.3](#)

**Figure 132.3** Example of addressing modes. (Adapted from Stallings, W. 1993. *Computer Organization and Architecture*. Macmillan, New York.)



## Data Representation Formats

Another important issue in computer architecture is the determination of data formats. Data formats, along with instruction formats, were formerly of much influence in determining **word size**. Today it is commonly assumed that most of the machines use a 32-bit word size (which is gradually shifting toward 64-bit). This standard was not common in the past, and there was not a common word size used by the majority of the machines. A 36-bit word size was quite common (IBM's early machines: 701, 704), and word sizes of 12, 18, and 60 bits were represented as well (PDP-8, CDC 6600). In the early days of computer development, interaction with the operator was done mainly via the teletype machine (TTY), which used 6 bits to represent each character. Therefore the word sizes of the machines of that period were determined with the objective of being able to pack several characters in the machine word. The size of I/O interfaces was commonly 12 bits (two characters). Anticipation of the new standard for the representation of digits (USASCII-8) prompted IBM to introduce an 8-bit character (EBCDIC) in their introduction of System/360 architecture, which was also its reason for switching from 36-bit to a new 32-bit word size. Since then (and until today) 32-bit word size and the multiples of the 8-bit quantity (**byte**) have been the most common data formats among various computer architectures. The new standard for representation of digits, USASCII-8, however, did not materialize. Instead, a 7-bit standard for data representation, ASCII, has been commonly used almost everywhere, except by IBM, which could not diverge from the 8-bit character representation defined in its architecture.

Every architecture must specify its representation of

Characters

Integers

Floating-point numbers

Logical operands

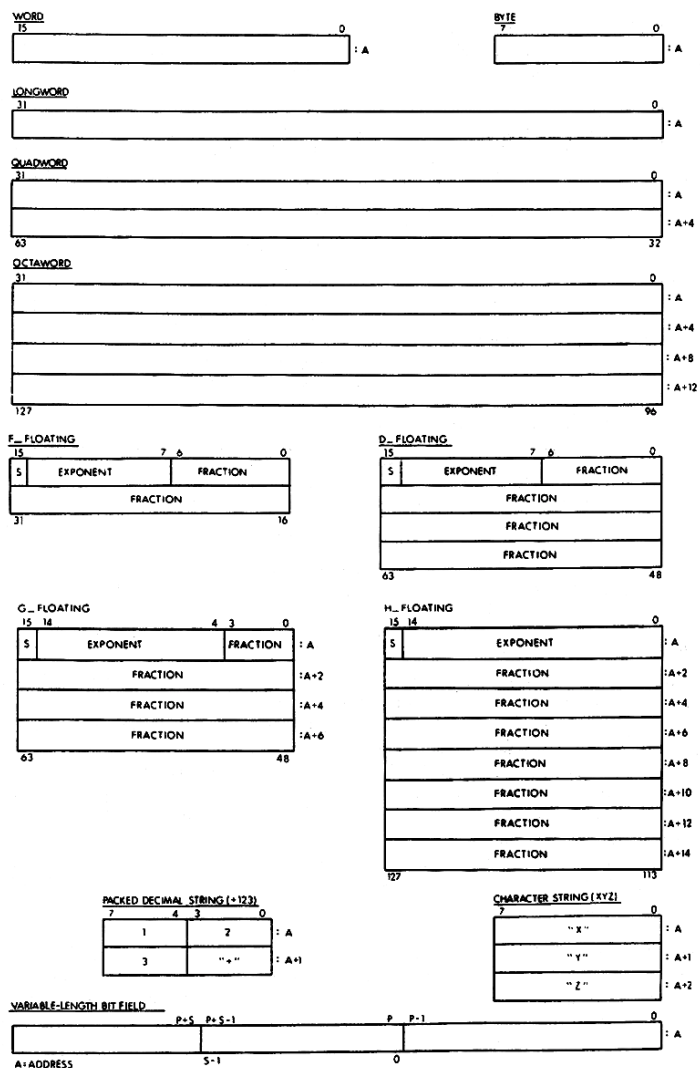
This representation must specify the number of bits used for every particular field, their order in the computer word, meaning of the special bits, interpretation, and the total length of data. Data types and data formats as defined in Digital Equipment Corporation VAX 11/780 architecture are shown in [Fig. 132.4](#).

**Figure 132.4** (a) Data types and (b) data formats as defined in Digital Equipment Corporation VAX 11/780 architecture. (*Source*: Digital Equipment Corporation. 1981. *VAX Architecture Handbook*. Digital Equipment Corporation, Maynard, MA. With permission.)

Figure 132.4

DATA TYPE	SIZE	RANGE (decimal)	
Integer		Signed	Unsigned
Byte	8 bits	-128 to + 127	0 to 255
Word	16 bits	-32768 to + 32767	0 to 65535
Longword	32 bits	$-2^{31}$ to $+ 2^{31} - 1$	0 to $2^{32} - 1$
Quadword	64 bits	$-2^{63}$ to $+ 2^{63} - 1$	0 to $2^{64} - 1$
Octaword	128 bits	$-2^{127}$ to $+ 2^{127} - 1$	0 to $+ 2^{128} - 1$
Floating Point			
F floating	32 bits	approximately seven decimal digits precision	
D floating	64 bits	approximately sixteen decimal digits precision	
G floating	64 bits	approximately fifteen decimal digits precision	
H floating	128 bits	approximately thirty-three decimal digits precision	
Packed Decimal String	0 to 16 bytes (31 digits)	numeric, two digits per byte sign in low half of last byte	
Character String	0 to 85535 bytes	one character per byte	
Variable-length Bit Field	0 to 32 bits	dependent on interpretation	
Numeric String	0 to 31 bytes (DIGITS)	$-10^{31}-1$ to $+ 10^{31}-1$	
Queue	> 2 longwords/queue entry	0 through 2 billion entries	

(a)



(b)

## Fixed-Point Data Formats

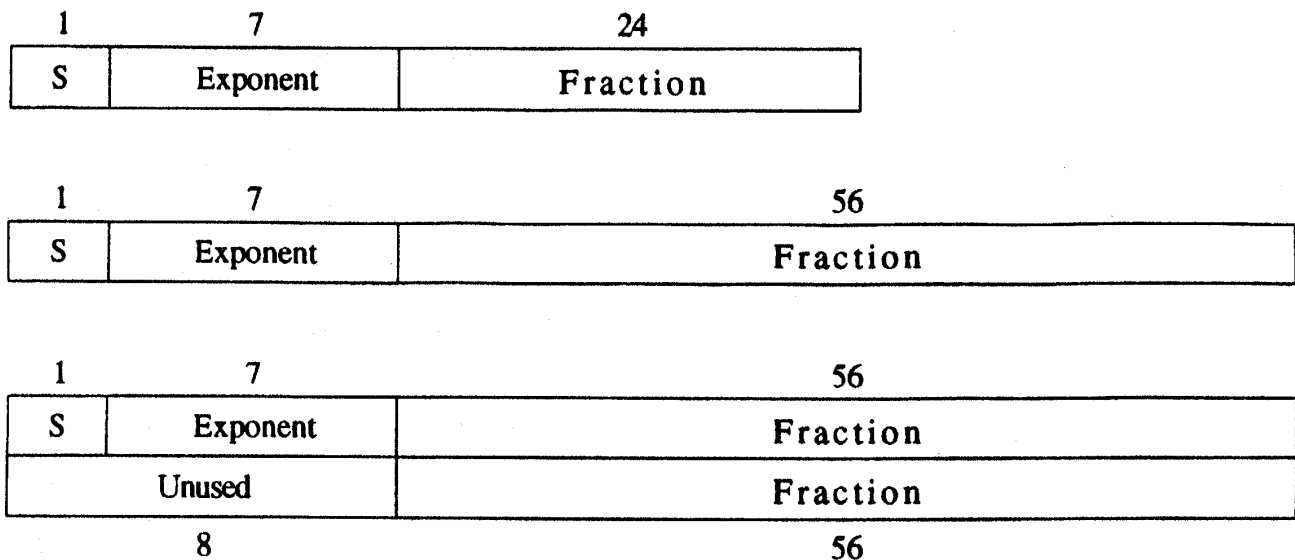
**Fixed-point** data forms are used to represent integers. Full-word (32 bit), half-word (16-bit), or double-word (64-bit) quantities are used for representation of integers. They can be signed or unsigned positive integers. In the case of signed integers one bit is used for representation of the sign, in order to represent a range of positive and negative integers. The most common representation of integers is 2s complement format. Another, not so common representation of integers is binary coded decimal representation (BCD), used to represent integers as decimal numbers. Each digit position is represented with 4 bits. The coding is straightforward for the numbers from 0 to 9, and the unused bit combinations are used to represent the sign.

For the *logical operand* a word is treated as a collection of individual bits, where each bit is assigned a Boolean value. A *variable bit field* can also be defined in cases where the field can be treated as a signed or unsigned field of bits.

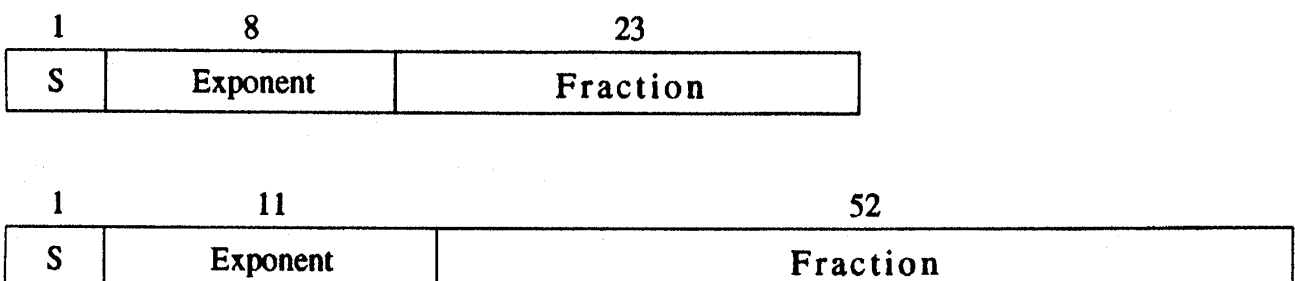
## Floating-Point Data Formats

For scientific computation, the dynamic range achievable using integers is not sufficient, and **floating-point** data representation is therefore defined. Each number is represented with the *exponent* and *fraction* (or mantissa). For the representation of a single number, one or more words could be used if required by the desired precision. Floating-point data formats specified in IBM System/360 architecture are shown in Fig. 132.5(a).

**Figure 132.5** Floating-point data representation formats.



(a) IBM System/370 Formats



(b) IEEE 754 Formats



A floating-point standard known as IEEE 754 has recently been introduced. The standard specifies the way data are to be represented, as well as the way computation should be performed. The purpose of this standard is to ensure that floating-point computation always produces exactly the same results—regardless of the machine or machine architecture being used. This result can be achieved only if the architecture complies to the IEEE 754 standard for floating-point computation. Data formats prescribed by IEEE 754 are shown in [Fig. 132.5\(b\)](#).

## 132.2 RISC Architecture

---

A special place in computer architecture has been given to RISC. RISC architecture was developed as a result of the 801 project, which started in 1975 at the IBM T. J. Watson Research Center and was completed by the early 1980s [[Radin, 1982](#)]. This project was not widely known to the world—outside of IBM and two other projects with similar objectives started in the early 1980s at the University of California, Berkeley, and Stanford University [[Patterson and Sequin, 1982](#); [Hennessy, 1984](#)]. The term **RISC** (reduced instruction set architecture), used for the Berkeley research project, is the term under which this architecture became widely known and recognized today.

Development of RISC architecture started as a "fresh look at existing ideas" [[Hopkins, 1987](#)] after first revealing evidence that surfaced as a result of examination of how the instructions are actually used in the real programs. This evidence came from the analysis of the *trace tapes*, a collection of millions of the instructions that were executed in the machine running a collection of representative programs. This evidence showed that, for 90% of the time, only about 10 instructions from the instruction repertoire were actually used. Then the obvious question was asked: "Why not favor implementation of those selected instructions so that they execute in a short cycle, and emulate the rest of instructions." The following reasoning was used: "If the presence of a more complex set adds just one logic level to a 10 level basic machine cycle, the CPU has been slowed down by 10%. The frequency and performance improvement of the complex functions must first overcome this 10% degradation, and then justify the additional cost" [[Radin, 1982](#)]. Therefore, RISC architecture starts with a small set of the most frequently used instructions, which determines the pipeline structure of the machine, enabling fast execution of those instructions in one cycle. One cycle per instruction is achieved by exploitation of parallelism through the use of **pipelining**. It turns out that *parallelism through pipelining* is the single most important characteristic of RISC architecture—from which all the rest of the RISC features could be derived. We can characterize RISC basically as a performance-oriented architecture based on exploitation of parallelism through pipelining. A list of the remaining features of RISC architecture is given in [Table 132.1](#).

RISC architecture has proven itself; several *mainstream architectures* today are of the RISC type. These include SPARC (used by Sun Microsystems workstations, an outgrowth of Berkeley RISC), MIPS (an outgrowth of the Stanford MIPS project, used by Silicon Graphics), and a **super-scalar** implementation of RISC architecture, IBM RS/6000 (also known as PowerPC architecture).



**Table 132.1** Features of RISC Architecture

Feature	Characteristic
Load/store architecture	All of the operations are register to register. In this way, operation is decoupled from the access to memory.
Carefully selected subset of instructions	Control is implemented in hardware. There is no microcoding in RISC. Also, this set of instructions is not necessarily small.*
Simple addressing modes	Only the most frequently used addressing modes are used. It is also important that they can fit into the existing pipeline.
Fixed size and fixed fields instructions	This is necessary to be able to decode instruction and access operands in one cycle (though there are architectures using two sizes for the instruction format, IBM PC-RT).
Delayed branch instruction (known also as branch and execute)	The most important performance improvement through instruction architecture.
One instruction per cycle execution rate, CPI=1.0	Possible only through the use of pipelining.
Optimizing compiler	Close coupling between the architecture and the compiler. Compiler "knows" about the pipeline.
Harvard architecture	Separation of instruction and data cache, resulting in increased memory bandwidth.

\*IBM PC-RT instruction architecture contains 118 instructions, whereas IBM RS/6000 (PowerPC) contains 184 instructions. This should be contrasted to the IBM System/360, containing 143 instructions, and IBM System/370, containing 208. The first two are representatives of RISC architecture, whereas the latter two are not.

## Defining Terms

**Accumulator:** A special register always containing one operand and possibly also receiving the result.

**Architectural compatibility:** Ability to run programs on separate machines and expect them to produce the same results.

**Byte:** An 8-bit quantity being treated as a unit.

**Computer architecture:** The attributes of a computer, as seen by the machine language programmer that enable this programmer to write functionally correct, time-independent programs.

**Computer implementation:** System organization and hardware structure.

**Computer organization:** Hardware structure encompassing the major functional units, data paths, and control.

**Fixed point:** Positive or negative integer.

**Floating point:** A number format, containing a fraction and an exponent, used for representation of numbers covering a wide range of values. Used for scientific computation where the range is important.

**Pipelining:** The technique used to initiate one operation in every cycle without waiting for the final result to be produced, or completion of previously initiated operations.

**Principles of operation:** A definition of the machine. Term used for computer architecture in IBM.

**RISC:** Reduced instruction set computer.

**Super scalar:** Implementation of an architecture capable of executing more than one instruction in the same cycle.

**Upwardly compatible architectures:** Ability to efficiently run user programs written for the old architecture on the new models without modifications to the program, lacking however, the capacity to do the reverse.

**Word size:** A quantity defined as the number of bits being operated upon as a unit.

## References

- Amdahl, G. M., Blaauw, G. A., and Brooks, F. P. 1964. Architecture of the IBM System/360. *IBM J. Res. Dev.* 8(2):87–101.
- Blaauw, G. A., and Brooks, F. P. 1964. The structure of System/360. *IBM Syst. J.* 3(2):119–135.
- Case, R. P. and Padegs, A. 1978. Architecture of the IBM System/370. *Commun. ACM.* 21(1):73–96.
- Digital Equipment Corporation. 1981. *VAX Architecture Handbook*. Digital Equipment Corporation,
- Hennessy, J. L. 1984. VLSI processor architecture. *IEEE Trans. Comput.* C-33 (12): pp–pp.
- Hopkins, M. E. 1987. A perspective on the 801/Reduced Instruction Set Computer. *IBM Syst. J.* 26(1): pp–pp.
- Patterson, D. A. and Sequin, C.H. 1982. A VLSI RISC. *IEEE Comput. Mag.*
- Radin, G. 1982. The 801 minicomputer. *SIGARCH Comput. Architecture News.* 10(2):39–47.
- Siewiorek, D. P., Bell, C. G., and Newell, A. 1982. *Computer Structures: Principles and Examples*. McGraw-Hill, New York.
- Stallings, W. 1993. *Computer Organization and Architecture*. Macmillan, New York.

## Further Information

A good introductory text for computer architecture is a book by William Stallings, *Computer Organization and Architecture*, Macmillan Publishing Company, 1993.

For the advanced reader, more information on computer hardware, design, and performance analysis can be found in a book by David A. Patterson and John L. Hennessy, *Computer Organization and Design: The Hardware/Software Interface*, Morgan Kaufmann Publishers, 1994. For quantitative analysis of instruction usage and various factors affecting performance, as well as insight into RISC architecture, *Computer Architecture: A Quantitative Approach*, by the same authors and publisher, is highly recommended.

An important historical insight into the development of computer architecture is an interview with Richard Case and Andris Padegs, "Case Study: IBM's System/360-370 Architecture," conducted by editors David Gifford and Alfred Spector in *Communications of ACM*, volume 30, number 4, April 1987, as well as the paper "The Architecture of IBM's Early Computers," published in the *IBM Journal of Research and Development*, volume 25, number 5, September 1981. The first chapter of the book by David J. Kuck, *The Structure of Computers and Computation* (John Wiley & Sons, 1978), contains an excellent overview of the history of computer development.

Various useful articles on computer architecture, performance, and systems can be found in *Computer Magazine*, published by the Computer Society of IEEE. More advanced articles on the subject of computer architecture, performance, and design could be found in the *IEEE Transactions on Computers*, published by IEEE. For subscription information regarding IEEE publications, contact: IEEE Service Center, 445 Hoes Lane, P. O. Box 1331, Piscataway, NJ 08855-1331, or phone (800)678-IEEE.

Cabrera, L. F. "Operating Systems"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

- 133.1 Typical Services
- 133.2 General Flow of Control
- 133.3 Structure
- 133.4 Communication
- 133.5 Advanced Data Management Services

**Luis-Felipe Cabrera**

*IBM Almaden Research Center*

The operating system of a computer is the software program responsible for controlling and administering all the hardware resources present in the system. The operating system is to be understood as the **control program** of a computer. It is through the basic set of services provided by the operating system that hardware devices can be brought into a computer system, used by the system, and exploited by other system services or by application software. Thus, the role of an operating system is to interface and communicate with all the hardware devices present in the system and to define, provide, and administer all the basic system facilities, or system services, that are present in a computer system. These system facilities are the building blocks for all other system services and for application software.

The system facilities offered by an operating system are normally exported to the rest of the system through a set of software functions named **system calls**. There are different manners in which system calls are implemented, yet their use is rather standard: Applications issue the calls as if they were part of the application program itself. To applications, system calls appear as functions provided by the environment.

Operating systems have evolved from being one-of-a-kind programs useful for only one type of hardware configuration to being portable programs that can be made to operate in a homogeneous family of hardware configurations and even in heterogeneous hardware platforms. The first operating system that operated on several different kinds of hardware configurations supporting many different kinds of devices was called OS/360. OS/360 was introduced by IBM in the early 1960s as a computing platform for commercial applications. OS/360 enabled, for the first time, the support of different generations of hardware devices in a continuous manner, without having to modify the software that ran in the systems and avoiding the creation of complete new operating systems to accommodate new devices.

The first operating system to operate in a wide variety of hardware platforms was Unix, introduced by AT&T Bell Laboratories in the mid-1970s as a computing platform for research and development of engineering applications. Unix had the characteristic that its size was small, and

thus it could operate in inexpensive computers used for small groups of people. The first operating system to be used massively was MS-DOS, introduced by Microsoft in the early 1980s as a computing platform for personal, or individual, use.

The services provided by operating systems vary depending on the intended use of the computer. In MS-DOS, for example, the computer can be doing only one **task** at the time, whereas in Unix the computer may be doing several tasks concurrently. Unix is called a multitasking operating system for this reason. The advantage of multitasking systems is that a user may have several different kinds of activities happening at once and thus users can make better use of time and resources while getting work done.

## 133.1 Typical Services

---

The services provided by an operating system can be organized in categories. Four possible categories are task control, file manipulation, device manipulation, and information maintenance. Task control has all the services needed to administer tasks, such as initiation and termination of a task, as well as services that may control the coordination of activities between tasks. File manipulation services are used to organize and access the data of users in files. File manipulation in advanced systems can be quite complex as there may be different kinds of files to optimize data availability or access time performance. The experimental system Swift/RAID is an example of a client/server file system in which files are stored in a manner that enhances the speed at which their data can be stored and retrieved and also enhances the availability of the data.

Device manipulation services provide the ability to communicate with and between devices in a system. The software in an operating system that directly accesses a hardware device is called a **device driver**. Information maintenance is an important category of services that tracks information on various aspects of a system and helps administer the resources present in the system.

Table 133.1 presents a base list of system services. The complexity of the operating system increases when the list of services is larger. For example, in computer systems composed of many processors and many devices where data are stored, the access to this data requires services that are substantially more complex than the one present in a personal computer. These advanced services are not represented in Table 133.1

**Table 133.1** Types of System Services

---

### **Task Control**

End abort

Load, execute

Create task, terminate task

Get task attributes, set task attributes

Wait for time

Wait event, signal event

### **File Manipulation**

Create file, delete file

Open, close

Read, write, reposition  
Get file attributes, set file attributes  
**Device Manipulation**  
Request device, release device  
Read, write, reposition  
Get device attributes, set device attributes  
**Information Maintenance**  
Get time or date, set time or date  
Get system data, set system data  
Get task, file, or device attributes; set task, file, or device attributes

---

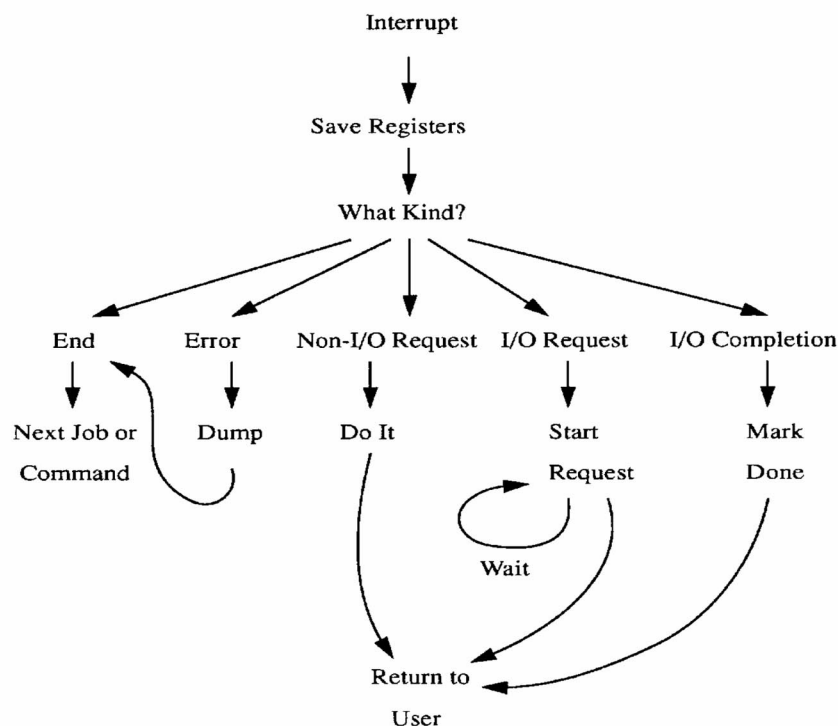
## 133.2 General Flow of Control

All devices—those that interact directly with the user, such as a keyboard, and those that are internal to the system, such as a hard disk—generate event **interrupts** when they complete a request made to them. (The interaction between a user and a computer through its input/output devices is explained in **Chapter 135**.) The operating system is programmed to first save the relevant information of an interrupt in locations called **registers** and then to service the interrupt.

With the information saved in registers the system determines what kind of event caused the interrupt. Then the operating system does the activities required by this interrupt and eventually returns control to the user application.

Figure 133.1 depicts a simplified general flow of an operating system. An important aspect missing in Figure 133.1 is the representation of all the synchronization activities that happen in the computer while concurrent tasks are active in one system.

**Figure 133.1** Simplified flow of an operating system.

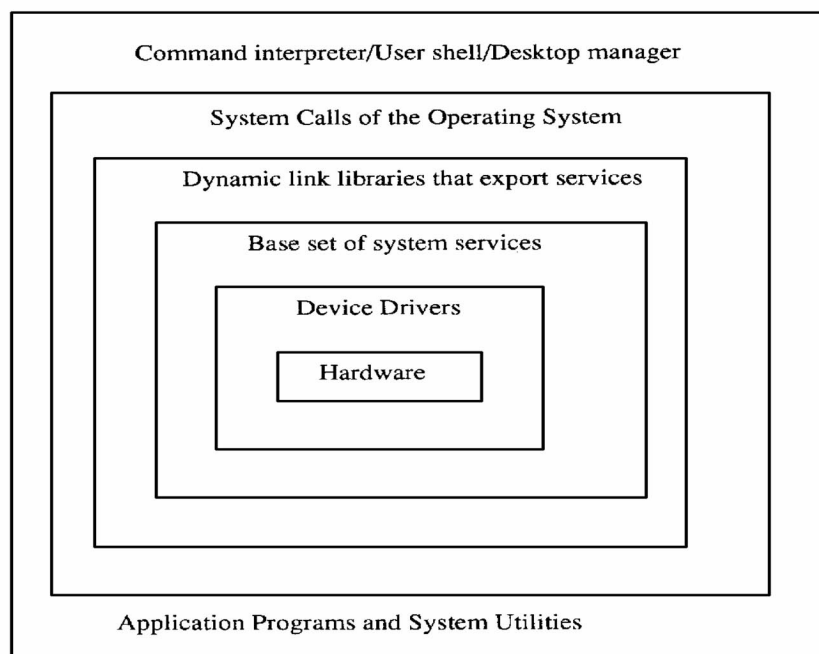


## 133.3 Structure

The organization of the structure of operating systems has also evolved over the years from being a monolithic set of system services whose boundaries were difficult to establish to being a structured set of system services with clear boundaries between them. Today, for example, some of these services, like the file system in charge of administering the file data stored by users, have boundaries, or software interfaces, that are standards regulated by organizations that are not related to any computer vendor. This modularization of operating systems into well-understood components provides the possibility to "mix and match" components from different vendors in one system. This trend toward pluggable system parts is accelerating with the formation of various consortia whose charters are to specify common, appropriate, and widely accepted interfaces for various system services.

Current operating systems are all based on the idea of building higher-level hardware abstraction from lower-level hardware-oriented function. In other words, current operating systems build higher-level abstractions from lower-level functions that are closer to the raw hardware. Thus, all kinds of hard disks, for example, are made to look and operate in the same manner by their low-level device drivers. Then, in turn, the operating system presents, with all other services in the system (such as the file system), a uniform, common view of a hard disk. This process of successive layers of abstraction is also followed within other services. In file systems, for example, a layer providing the abstraction of continuous storage that does not have device boundaries is built from the abstraction provided by the individual storage in several disks. [Figure 133.2](#) shows a layered depiction of an operating system.

**Figure 133.2** Layered depiction of an operating system.

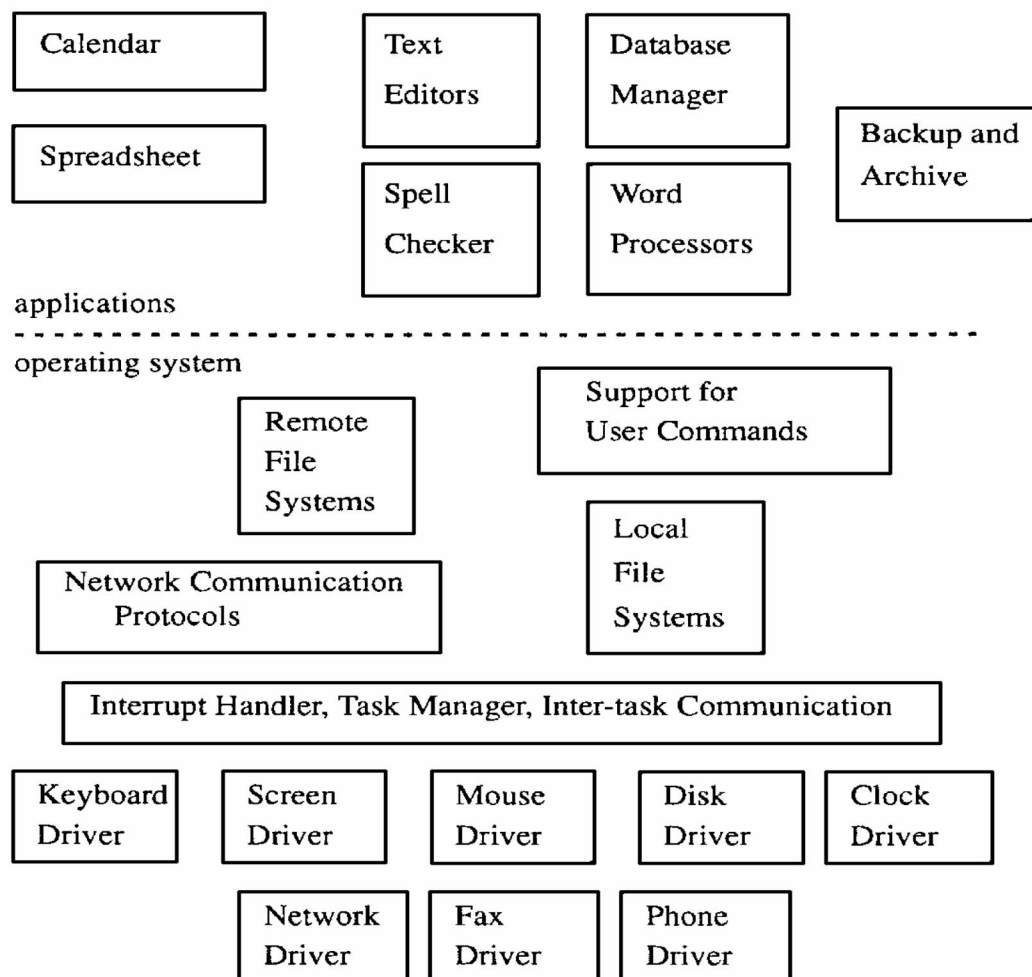




All computer systems have a base set of hardware components that need to be administered by the operating system. These components are the building set for all computing that can be done in the machine. Most current personal computers, for example, have a keyboard, a pointing device or mouse, a high-resolution screen, an internal clock, a phone modem, a diskette reader, and a hard disk.

The **kernel** of an operating system is the most basic set of service needed to build the complete system. The kernel always contains the interrupt handler, the task manager, and the interprocess communication manager. Sometimes it may also contain the **virtual memory** manager and the network subsystem manager. With these services the system can operate all the hardware present in the system and also coordinate the activities between tasks. Using these services, subsystems such as the file system and the network system can build their services. Applications can then provide their function by using these subsystems. A remote file service, for example, will use the network subsystem to provide the illusion of local files by connecting to a file system that can be accessed over the network. [Figure 133.3](#) shows a set of base operating system components and applications that may be found in a computer system.

**Figure 133.3** A set of base operating system components and applications that may be found in a computer system.



## 133.4 Communication

---

A fundamental characteristic that may vary from system to system is the manner of communication between tasks. The two manners in which this is done is via messages sent between tasks or via the sharing of memory where the communicating tasks can both access the data. Operating systems can support either. In fact, both manners can coexist in a system. In message-passing systems, the sender task builds a message in an area that it owns and then contacts the operating system to send the message to the recipient. There must be a location mechanism in the system so that the sender can identify the receiver. The operating system is then put in charge of delivering the message to the recipient. To minimize the overhead of the message delivery process, some systems try to avoid copying the message from the sender to the kernel and then to the receiver and provide means by which the receiver can read the message directly from where the sender wrote it. This mechanism requires that the operating system intervene if the sender wants to modify the contents of a message before the recipient has gone through its content.

In memory-sharing systems the sender and receiver use a common area of memory to place the data that is to be exchanged. To guarantee appropriate concurrent manipulation of these shared areas, the operating system has to provide synchronization services for mutual exclusion. A common synchronization primitive is the **semaphore**, which provides mutual exclusion for two tasks using a common area of memory. In a shared memory system the virtual memory subsystem must also collaborate to provide the shared areas of work.

There are several examples of commercial and experimental operating systems that share memory and send messages. The office-oriented computing systems commercialized by Xerox are examples of memory-sharing systems. In fact, in those systems that ran a version of the operating system called Pilot, all the tasks would share a substantial amount of memory, each having only a small area private to itself. A message-passing system that is increasing its commercial acceptance is Mach, which was first developed at Carnegie Mellon University.

There are systems in which the amount of data that can be shared or sent between tasks is minimal. In the original Unix, for example, different tasks would not have shared memory, and only a very limited form of message passing was provided. The messages were constrained to be event notification. In this environment, processes that desired to share data had to write in files and then share the files. To support file sharing, some systems provide a **locking** service that is used to synchronize the accesses to files by different tasks.

## 133.5 Advanced Data Management Services

---

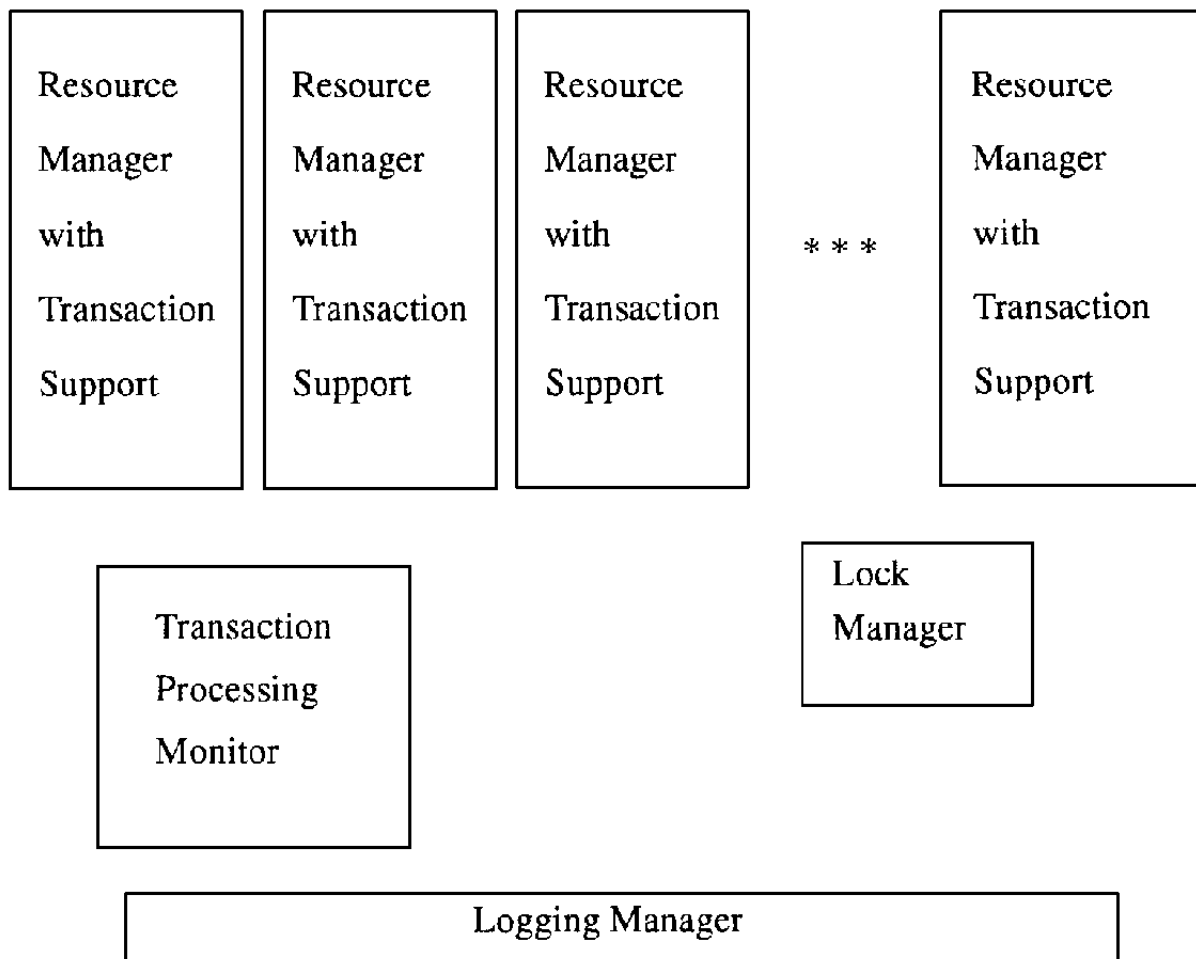
There are some operating systems in which data integrity is given a special attention so as to provide special function to support the transactional manipulation of data. Data are manipulated transactionally when sets of updates are treated as complete units of work, in that either all of them are made to happen at once or none of them happen. This atomicity property is useful to guard mission-critical data against partial updates caused by unexpected software or hardware failures. Commercial applications, such as banking, have used these techniques for decades.

Transactional manipulation of data acquires new benefits when the operations desired are to be done by a set of interconnected computers. To minimize the need for human intervention in the

administration of these installations, the transactional service is also useful as, upon restart from a failure, the system automatically rebuilds itself in a consistent state without the need of human intervention.

Figure 133.4 shows the typical services present in a system that provides transactional data administration. The transaction-processing monitor is in charge of the overall coordination of activities that allow to implement the atomicity and serialization properties. The resource managers with transaction support are the repositories of the data that are being manipulated transactionally. The lock and log managers are system services necessary to provide transactions.

**Figure 133.4** Typical components of a transaction-processing system.



## Defining Terms

**Control program:** A software program that administers the hardware resources present in a computer system. Through this software resources are accessed and the activities between different resources are coordinated and prioritized.

**Device driver:** Software that directly accesses a hardware device controlling its operation. Device drivers are specific, or one-of-a-kind, for each device in each operating system. As like devices have similar functions, there are similarities in their device drivers, particularly in the interfaces they provide the operating system.

**Interrupt:** An action initiated by the hardware of the computer that transfers control from any running activity to the operating system. The operating system relinquishes its control after having determined the type of interrupt and having performed the activities needed of its behalf.

**Kernel:** Essential set of services needed to build the complete operating system. Kernel services are the most basic services provided by a given system.

**Locking:** System service that enables callers of the service to explicitly synchronize their activities by placing locks on entities that require access synchronization and coordination.

**Register:** Locations in the processing unit of the computer used to store data for operations, the result of operations and the status of hardware devices.

**Semaphore:** Synchronization primitive used between tasks that share a common region of memory to achieve isolation of activity.

**System calls:** The set of software functions callable from different programming languages used by all applications to perform the functions provided by the operating system.

**Task:** A logical unit of work being done by a computer at a given time. This can be anything from a system activity, such as reading some data from a hardware disk or printing some region of a screen, to an activity done by an application, such as editing a file with an editor, using a spreadsheet program for some financial calculations, keeping track of a schedule of events with a calendar program, or displaying the time of day with a clock program.

**Virtual memory:** A system service that presents to applications the abstraction that there is more memory available for them in the system than the real amount of memory installed in the system.

## References

- Cabrera, L.-F., McPherson, J., Schwarz, P. M., and Wyllie, J. C. 1993. Implementing atomicity in two systems: Techniques, tradeoffs, and experience. *IEEE Trans. Software Eng.* 19(10): 950–961.
- Deitel, H. M. and Kogan, M. S. 1992. *The Design of OS/2*. Addison-Wesley, Reading, MA.
- Gray, J. and Reuter, A. 1993. *Transaction Processing: Concepts and Techniques*. Morgan Kaufmann, San Mateo, CA.
- Haskin, R., Malachi, Y., Sawdon, W., and Chan, G. 1988. Recovery management in QuickSilver. *ACM TOCS*. 6(1):82–108.

Long, D. D. E., Montague, B. R., and Cabrera, L.-F. 1994. Swift/RAID: A distributed RAID system. *Computing Syst.* 7(3):333–359.

Peterson, J. L. and Silberschatz, A. 1985. *Operating System Concepts*. Addison-Wesley, Reading, MA.

## **Further Information**

In the U.S. the three most relevant professional organizations that sponsor publications and conferences in the area of operating systems are the Computer Society of the Institute of Electrical and Electronics Engineers (IEEE CS), the Usenix Association, and the Association for Computing Machinery (ACM). In particular, the *IEEE Transactions on Computers*, the *IEEE Transactions on Software Engineering*, the proceedings of the IEEE conferences on distributed computing systems, the journal *Computing Systems*, the proceedings of the winter and summer Usenix conferences, the *ACM Transactions of Computer Systems*, and the proceedings of the ACM symposia on operating systems principles are excellent sources of up-to-date technical information on operating systems.

Dennis M. Volpano. "Programming Languages"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Programming Languages

---

## 134.1 Principles of Programming Languages

Data Types • Syntax and Semantics • Encapsulation

## 134.2 Program Verification

## 134.3 Programming Language Paradigms

Concurrent and Distributed Programming • Real-Time Programming

**Dennis M. Volpano**

*Naval Postgraduate School*

The evolution of programming languages is an ongoing process. It is driven by new machine architectures, different models of computation, and the need for new techniques to express, at a more abstract level, algorithms for certain applications. Our understanding of programming languages and our ability to implement them has improved dramatically since the early days of Fortran. Programming paradigms, such as functional, logic, concurrent, and distributed programming, continue to be developed. Programming languages are typically distinguished according to these paradigms. However, for a given paradigm, a more fundamental distinction can be made based upon a set of principles. These principles are presented, as are other facets of modern programming paradigms.

## 134.1 Principles of Programming Languages

---

### Data Types

A programming language provides a collection of data types that determines the values that can be manipulated in the language. Languages often can be distinguished based on their data types. The primitive data types of a language are formed from two sets—the empty set (denoted **0**) and the singleton set {unit} (denoted **1**)—and from type-forming operations discriminated union (+), Cartesian product ( $\times$ ), function types ( $\rightarrow$ ), and recursive types (**fix**).

#### Discriminated Union

If  $S$  and  $T$  are types then their discriminated union is given by

$$S + T = \{\text{left } x \mid x \in S\} \cup \{\text{right } y \mid y \in T\}$$

Elements of  $S + T$  are tagged by applying constructors *left* and *right* to elements of  $S$  and  $T$ ,

respectively. For example, an enumeration type of truth values can be constructed by

$$\text{bool} = 1 + 1 = \{\text{left unit}, \text{right unit}\}$$

where *left unit* may denote true or false. Discriminated union is not the same as set union for  $1 \cup 1 = \{\text{unit}\}$ . Languages manifest the discriminated union type as the type of a variant record.

### Cartesian Product

If  $S$  and  $T$  are types then their Cartesian product is given by

$$S \times T = \{(x, y) \mid x \in S \wedge y \in T\}$$

For example, the following Cartesian product type is isomorphic to `bool`:

$$1 \times \text{bool} = \{(\text{unit}, \text{left unit}), (\text{unit}, \text{right unit})\}$$

Languages manifest the Cartesian product type as the type of a record or tuple.

### Function Types

If  $S$  and  $T$  are types then  $S \rightarrow T$  consists of all functions from  $S$  to  $T$ . For example, the function type `bool`  $\rightarrow$  `bool` consists of the constant true and false functions, the identity function, and the complement function. Languages manifest the function type as the type of a function declaration or function abstraction and, in the case of finite functions, also as the type of an array.

### Recursive Types

A poset  $(S, \leq)$  is a domain if there is a unique least element  $x \in S$ , meaning  $x \leq y$ , or  $x$  approximates  $y$ , for all  $y \in S$  and for every infinite increasing sequence  $x_0 \leq x_1 \leq x_2 \leq \dots$  there is a unique element  $z$ , called its *limit*, that is approximated by every  $x_k$ , and that approximates any other element that is approximated by every  $x_k$ . A function  $F$  between domains is *continuous* if it is monotonic and preserves limits—that is,  $F(x) \leq F(y)$  if  $x \leq y$ —and if the increasing sequence  $x_0 \leq x_1 \leq x_2 \leq \dots$  has limit  $z$  then increasing sequence  $F(x_0) \leq F(x_1) \leq F(x_2) \leq \dots$  has limit  $F(z)$ . Then if  $F$  is a continuous function between elements of a domain then **fix**  $F$  is the least fixed point of  $F$ .

For example, let  $F = \mathbf{fn} \ z \Rightarrow 1 + (H \times z)$ , where  $H$  stands for characters. Then **fix**  $F$  is a recursive type containing all finite strings. Successive approximations of **fix**  $F$  are given by  $Q_{k+1} = F(Q_k)$ , with  $Q_0 = 0$  and set inclusion as the partial order. Intuitively, the approximation

$$Q_1 = 1 + (H \times Q_0) = \{\text{left unit}\}$$



contains only the empty string, whereas

$$Q_2 = 1 + (H \times Q_1) = Q_1 \cup \{\text{right}(h, \text{left unit}) \mid h \in H\}$$

contains the empty string and all strings of length one. In general,  $Q_k$  contains all strings of length less than  $k$ . The limit of sequence

$$Q_0 \subseteq Q_1 = F(Q_0) \subseteq Q_2 = F(Q_1) \subseteq \dots$$

is given by  $Q = Q_0 \cup Q_1 \cup Q_2 \cup \dots$ . But  $Q$  is also the limit of sequence

$$F(Q_0) \subseteq F(Q_1) \subseteq F(Q_2) \subseteq \dots$$

which has as its limit  $F(Q)$ , since  $F$  is continuous. Therefore,  $Q = F(Q)$ , or  $Q$  is a fixed point of  $F$ . For any other fixed point  $Q'$  of  $F$ ,  $Q_0 \subseteq Q'$ , and if  $Q_k \subseteq Q'$  then  $Q_{k+1} = F(Q_k) \subseteq F(Q') = Q'$ . Since  $Q$  is the limit and  $Q'$  is approximated by every  $Q_k$ ,  $Q \subseteq Q'$ . This implies  $Q = \mathbf{fix} F$ , or  $Q$  is the least fixed point of  $F$ . Languages manifest the recursive type as the type of an element belonging to an *infinite* set. Some languages permit recursive types to be defined, and their run-time support typically includes **garbage collection**. Other languages allow only definition of self-referential structures using pointers. Here, memory allocation and deallocation are typically a programmer's responsibility.

## Syntax and Semantics

A programming language consists of three syntactic classes of phrases: definitions (**Def**), expressions (**Exp**), and commands (**Com**). These classes arise out of semantic differences between phrases. Expressions are evaluated for their values, commands are executed for their effect on a store, and definitions are evaluated for their effect on an environment. Let **L** denote locations, **B** basic values, and  $\mathbf{R} = \mathbf{L} + \mathbf{B}$  storable values. A *store* is an element of domain

$$\mathbf{S} = \mathbf{L} \rightarrow (\mathbf{R} + \{\text{error}\})$$

and maps a location to the value stored at that location. An *environment* is an element of domain

$$\mathbf{Env} = \mathbf{Id} \rightarrow (\mathbf{R} + \{\text{error}\})$$

and maps an identifier to a storable value. If an environment maps an identifier to a location, then the identifier stands for a *variable*; otherwise, it denotes a *constant*. A formal description of a phrase's semantics (meaning) may be given by semantic functions. There is one for each class:

$$\begin{aligned}
\mathcal{E}: \text{Exp} &\rightarrow \text{Env} \rightarrow \mathbf{S} \rightarrow (\mathbf{R} + \{\text{error}\}) \\
\mathcal{C}: \text{Com} &\rightarrow \text{Env} \rightarrow \mathbf{S} \rightarrow (\mathbf{S} + \{\text{error}\}) \\
\mathcal{D}: \text{Def} &\rightarrow \text{Env} \rightarrow \mathbf{S} \rightarrow (\text{Env} \times \mathbf{S})
\end{aligned}$$

where  $\mathcal{E}$ ,  $\mathcal{C}$ , and  $\mathcal{D}$  give meaning to expressions, commands, and statements, respectively. This kind of semantic description is called a *denotational semantics*. Each semantic function maps an element of a syntactic class to a function over semantic domains **Env**, **S**, and **R**.

## Expressions

An expression is evaluated for its value, although in some languages like C it may be evaluated for its effect as well. This kind of effect is called a *side effect*; it makes reasoning about programs more difficult since it compromises *referential transparency*, the ability to replace all occurrences of an expression in a program by a single value without changing the program's meaning. Examples of expressions are literals, identifiers, compound expressions (e.g., arithmetic expressions), block expressions, and procedure expressions.

Languages may differ in their interpretation of compound expressions in that subexpression evaluation order may vary. But if expressions are side effect-free, then evaluation order is irrelevant. Even without side effects, languages may still interpret expressions differently. For example, the conventional interpretation of an identifier  $I$  is given by

$$\mathcal{E}[I] \ e \ s = s(e \ I) \quad \text{if} \quad (e \ I) \in \mathbf{L} \quad \text{otherwise} \quad e \ I$$

which specifies that addresses of variables are implicitly dereferenced. Therefore, in languages like Pascal, C, and Ada, different occurrences of the same variable within a program may have different meanings! This is true in the assignment command  $x := x + I$ , where on the left side, the occurrence of variable  $x$  means  $(e \ I)$  and on the right it means  $s(e \ I)$ . Standard ML eliminates this inconsistency by adopting

$$\mathcal{E}[I] \ e \ s = e \ I$$

as its interpretation of identifiers. Now the store is applied explicitly via a dereference  $!x$  so that the assignment becomes  $x := !x + 1$ . Both occurrences of variable  $x$  have the same meaning  $(e \ I)$ .

Another kind of expression that may be interpreted differently by various languages is the procedure expression. Such an expression is used to create procedure abstractions, which in some languages may remain anonymous [Watt, 1990]. One common interpretation is given by

$$\mathcal{E}[\text{procedure } C] \ e \ s = \mathcal{C}[C] \ e$$

which conveys that procedure body  $C$  must be closed with respect to environment  $e$  but not with respect to store  $s$ , since  $e$  appears on the right side but  $s$  does not. Consequently, the procedure will be executed in the environment of its *definition* rather than the environment of its *invocation* and in the store of its invocation, not the store of its definition. This interpretation is called a **static**

**binding** semantics because any free identifiers in  $C$  will get their bindings from the environment at the point of definition. Programming languages Pascal, Modula, C, Ada, Scheme, and Common Lisp interpret procedure expressions in this way. In contrast, a **dynamic binding** semantics is given by

$$\mathcal{E}[\text{procedure } C] \text{ } e \text{ } s = \mathcal{C} [C]$$

where the environment is now supplied when the procedure is invoked. Free identifiers then get their bindings from the environment at the point of invocation. Early dialects of Lisp adopt this semantics.

The temporary rebinding of predefined identifiers for the purpose of debugging is one advantage of dynamic binding. There are many disadvantages, however. A free identifier may not have a unique binding, or, in other words, scope rules cannot be used to determine meaning. Local variables can be modified by procedures that fall outside the scopes of these local variables. Further, bound variables can no longer be consistently renamed without possibly changing a program's meaning. Achieving, in a language with dynamic binding, the static binding of free identifiers in functions passed as arguments has come to be known as the *funarg problem*.

## Commands

A command is executed for its effect on a store. Examples of commands are assignment, command composition, block commands (as in block-structured languages), conditionals, and procedure calls. A procedure call may transmit actual parameters that are referenced in the called procedure via its formal parameters. Parameter transmission is typically one of three kinds: *call-by-value*, *call-by-value-result*, or *call-by-reference*. Of the three, only call-by-value-result affects the epilogue of a procedure call, since it has a copy-out phase where values of the formal parameters are copied back to the corresponding actual parameters. In languages where identifiers may share a location, i.e., where they are *aliases* for each other, these transmission techniques can produce different results. A procedure whose correctness can be established independently of the type of one or more of its formal parameters is called a *polymorphic procedure*.

Call-by-value is also referred to as *eager evaluation*. Sometimes a called procedure may not need the values of all its formal parameters. In this case, evaluation of the corresponding actuals can be avoided. Evaluating an actual parameter if and only if its value is needed in the called procedure is called *call-by-name*. Under call-by-name, an actual parameter's evaluation is delayed until its corresponding formal is referenced, but each reference calls for another evaluation of it. If it is evaluated at most once, however, then parameter transmission is *call-by-need* [Friedman *et al.*, 1992]. Call-by-name is also referred to as *lazy evaluation*.

## Definitions

A definition is evaluated for its effect on an environment. Every definition has a *scope* as prescribed by a scope rule. The scope is a region of text. An applied occurrence of identifier  $I$  is *free* in  $R$  if it does not fall within the scope of a definition for  $I$ ; otherwise, it is bound in  $R$ . If a definition  $I = E$  has scope  $R$ , then every free occurrence of  $I$  in  $R$  denotes  $E$ .

There are two kinds of definitions in traditional imperative languages—*constant* and *variable*

definitions. Their meanings are given by

$$\mathcal{D}[\text{const } I = E] \ e \ s = (\text{fix fn } e' \Rightarrow e[I \mapsto \llbracket \mathcal{E} \rrbracket e' \ s], s)$$

$$\mathcal{D}[\text{var } I = E] \ e \ s = (e[I \mapsto l], s[l \mapsto \mathcal{E}[E] \ e \ s])$$

where  $l$  is a new location and  $e[I \mapsto l]$  is environment  $e$  updated so that  $I$  is mapped to  $l$ . A constant definition, **const**  $I = E$ , unlike a variable definition, never modifies a store. It also requires the scope of identifier  $I$  to include  $E$  in order to express *recursion*. One disadvantage of this requirement is that references to predefined identifiers can no longer be intercepted, say during debugging, unless some form of scope resolution is provided.

Suppose, for example, that *eof* is a predefined end-of-file predicate of one argument. Then

$$\text{const } eof = \text{fn } f \Rightarrow \dots eof \ f \dots;$$

is a recursive definition. However, our intention may be to create a user-defined version of *eof* from which some additional operations are performed before calling the predefined version of *eof*. Standard ML [Paulson, 1992] and Scheme [Friedman *et al.*, 1992] permit the desired definition to be expressed because they have forms of definition with different scope rules. For example, in Scheme,

$$(\text{let } ((eof \ (\text{lambda } (f)(\dots eof \ \dots)))) \dots)$$

gives the desired definition, as does

$$\text{val } eof = \text{fn } f \Rightarrow \dots eof \ f \dots;$$

in Standard ML. In each case the scope of *eof* does not include the definition. To include it, **letrec** is used in Scheme and **fun** or **val rec** in Standard ML:

$$\text{val rec } eof = \text{fn } f \Rightarrow \dots eof \ f \dots;$$

Though this definition appears circular, it is actually shorthand for the noncircular definition

$$\text{val } eof = \text{fix fn } eof \Rightarrow \text{fn } f \Rightarrow \dots eof \ f \dots;$$

which conveys that *eof* is defined as the least fixed point of functional  $\text{fn } eof \ \text{fn } f \Rightarrow \dots eof \ f \dots$ , assuming that the functional is continuous over some domain.

## Encapsulation

Large programs without a discipline for accessing global variables are hard to maintain and reuse because procedures that affect a global variable cannot be developed independently of other procedures that also access the variable. Modifications needed to accommodate a change in the global variable's declaration are not confined to a specific region of the program. To confine them, a global variable is declared within a *module* together with those procedures that affect it, so that its scope is limited to the module. Thus the variable is effectively hidden or inaccessible to all procedures declared outside the module. This is called *encapsulation* or *information hiding*.

### Modules and Objects

There are two basic kinds of modules: *objects* (modules with constant or variable definitions) and *structures* (modules with constant definitions only). An object also has an interface indicating the definitions whose scopes extend beyond the object, that is, are exported. An object declaration affects the environment and store. A package body in Ada with a variable declaration that is not specified in the package specification is an example of an object with a hidden variable [Barnes, 1992]. An *object class* declaration, unlike an object, affects only the environment, not the store; but its instantiation, which creates an object, affects both. A *class* in C++ is an example of an object class. An object class may be derived from multiple object classes in the sense that it *inherits* operations from these classes. Languages that permit derived object class declarations of this kind, such as C++ and Eiffel [Meyer 1989], are called **object-oriented languages**.

A structure is a tuple of constant definitions. As such, its declaration affects only the environment. A structure is used to group logically related definitions or to implement an abstract data type. An *abstract data type* is a heterogeneous algebra that may have multiple representations. Each representation has a *coupling invariant* that relates abstract and concrete values [Gries and Volpano, 1990]. A language supports abstract data types if it prohibits *impersonation*, applying abstract operations to representation values, and *unauthorized access*, applying representation operations to abstract values. A language that does not prevent unauthorized access cannot guarantee that programs will always be independent of abstract data type representations. Some languages, such as Standard ML, permit type abstractions and their representations to be separately specified and allow mappings between structures, called *functors*, to be defined, which serve as useful system-building operations.

## 134.2 Program Verification

---

Program verification is concerned with mathematically proving the correctness of a program. It requires a set of axioms and inference rules for programming language constructs, called a *programming logic*. A formal proof of a statement about a program is a sequence of facts ending with the statement, such that every fact is an instance of an axiom or follows from preceding facts by an inference rule.

An example of a programming logic is Floyd-Hoare logic. In this logic,  $\{P\}C\{Q\}$  is a *partial correctness* specification, where  $P$  is a precondition,  $Q$  is a postcondition, and  $C$  a command. It is true if whenever  $C$  is executed in a state that satisfies  $P$  and execution of  $C$  terminates, then it does

so in a state that satisfies  $Q$ . The specification may be true even though  $C$  does not always terminate. A stronger kind of specification is a **total correctness** specification, which also requires that  $C$  terminate whenever it is executed in a state satisfying  $P$  [Gries, 1981]. Let  $P[E/V]$  denote  $P$  with all free occurrences of  $V$  replaced by  $E$ . Then, an example of a Floyd-Hoare axiom is the assignment axiom,

$$\vdash \{P[E/V]\} \quad V := E \quad \{P\}$$

where  $V$  is any variable,  $E$  any side effect-free expression, and  $P$  a postcondition. An instance is

$$\vdash \{a[j] = 0\} \quad a[i] := 0 \quad \{a[i] = a[j]\}$$

Notice that the precondition is not implied by  $i = j$ , even though  $\{i = j\} \quad a[i] := 0 \quad \{a[i] = a[j]\}$  is true. A more suitable precondition is the *weakest precondition*, as is given in the instance

$$\vdash \{a[j] = 0 \vee i = j\} \quad a[i] := 0 \quad \{a[i] = a[j]\}$$

A programming logic is *sound* with respect to a semantics if deductions in the logic are consistent with the semantics. For example, the above assignment axiom is unsound in the language C since

$$\vdash \{j = 0\} \quad i = j++; \quad \{j = 0\}$$

can be deduced but is false. A logic is *complete* with respect to a semantics if every specification that is true for the semantics can be proved in the logic [Gordon, 1988].

## 134.3 Programming Language Paradigms

---

The kinds of phrases that a programming language allows is primarily what distinguishes it from other languages. A language with commands is called an **imperative language**. Programs written in such a language are formally characterized as store (state) transitions. Examples include block-structured languages, such as Algol-60 and its descendants—Pascal, Modula, and Ada. A programming language without commands is a declarative language and includes **functional languages** [Hudak, 1989] and **logic languages** [Shapiro, 1989]. Distributed programming languages support computation by independent processes that communicate by message passing. Processes may be distributed across distant nodes of a network. Parallel or concurrent languages, on the other hand, are aimed at supporting concurrent execution of sequential processes that communicate by shared variables.

# Concurrent and Distributed Programming

## Concurrent Programming

A *concurrent program* is a set of sequential programs that execute simultaneously on a computer in which processes communicate via shared variables and are able to call a centralized operating system to receive a service. This is typically called *large-grained concurrency*, compared to the small-grained variety, such as data and control parallelism provided by parallel vector processors and parallel function units.

Reasoning about the behavior of a concurrent program is facilitated by regarding its execution as an interleaved execution sequence of atomic instructions of sequential processes. There may be more than one sequence, so a property of a concurrent program is a property that holds for all interleavings. Sequential processes may share a resource for which mutually exclusive access is necessary. That portion of a process in which the resource is needed is called its *critical section*. A concurrent program has the *mutual exclusion property* if instructions from the critical sections of two or more processes are never interleaved. Dekker's algorithm ensures this property for two processes, whereas Peterson's algorithm [Peterson, 1983] ensures it for any number of processes.

Other important properties of a concurrent program include freedom from deadlock and starvation. *Deadlock* arises when all sequential processes are prevented from making further progress due to mutually unsatisfiable demands for additional resources. A concurrent program may detect, prevent, or avoid deadlock. *Starvation* exists if one process is prevented indefinitely from entering its critical section.

Higher-level support for mutual exclusion is provided by semaphores and monitors. A *semaphore* is an integer-valued variable  $S$  ranging over natural numbers with two atomic operations—**wait**( $S$ ), which may cause a process to suspend on semaphore  $S$ , and **signal**( $S$ ), which may resume execution of a suspended process. There are various kinds of semaphores, including blocked-set, blocked-queue, busy-wait, and binary semaphores. A semaphore is considered unstructured because its correct use depends on independent calls to primitives **wait** and **signal**. Structured support for mutual exclusion is provided by a *monitor*, which is an object with the property that at most one process can execute its procedures at a time. Monitors are used extensively in concurrent programming, and dialects of Pascal have been developed with monitor-based programming primitives [Bustard et al., 1988].

## Distributed Programming

A *distributed program* is like a concurrent program, except that sequential programs may execute on different computers and communicate by sending messages to each other. In a physically distributed system, processes communicate through the exchange of messages using a *protocol*, which prescribes a format for messages and detects and corrects errors in their transmission. With *synchronous communication*, message exchange requires both the sender and receiver to participate. Each must reach a designated point in its execution, called a *rendezvous*, before a message can be exchanged between them. If one reaches this point before the other, then it must wait. This form of communication is found in Ada. If a sender is allowed to send a message and continue without blocking, then communication is *asynchronous*. In this case the sender may send multiple messages before the receiver responds, so message buffering is needed. Linda is a concurrent language with asynchronous communication [Gelernter, 1985].



## Real-Time Programming

A *real-time program* is a program whose execution is subject to real-time constraints on response time. Such programs are found in embedded computer systems, such as flight-control systems. A real-time program may be implemented by a synchronous clock-driven scheduler. Processor time is divided into frames and tasks into segments so that every segment can complete in one frame. A *scheduling table* assigns segments to frames so that all segments in a frame complete their execution by the end of the frame. When the clock signals the start of a frame, the scheduler calls the segments for that frame as prescribed by the table. An alternative to partitioning tasks is to assign them priorities instead, as in Ada [Barnes, 1992]. Tasks execute asynchronously but are scheduled by a preemptive scheduler, which interrupts a running process if a higher-priority task becomes ready to execute.

## Defining Terms

**Dynamic binding:** A free variable's binding in the environment at the point of invocation.

**Functional language:** A language whose programs are mappings between values expressed using only expressions and function declarations.

**Garbage collection:** The systematic recycling of inaccessible memory cells during program execution.

**Imperative language:** A language whose programs are mappings between stores expressed using variables, commands, and side effects.

**Logic language:** A language whose programs are relations expressed using only expressions and clauses.

**Object-oriented language:** A language whose programs may be written with derived object classes.

**Static binding:** A free variable's binding in the environment at the point of definition.

**Total correctness:** A correctness criterion requiring partial correctness and termination.

## References

- Barnes, J. G. P. 1992. *Programming in Ada*. Addison-Wesley, Reading, MA.
- Bustard, D., Elder, J., and Welsh, J. 1988. *Concurrent Programming Structures*. Prentice Hall, Hemel Hempstead, UK.
- Friedman, D., and Wand, M., and Haynes, C. 1992. *The Essentials of Programming Languages*. McGraw-Hill, New York.
- Gelernter, D. 1985. Generative communication in Linda. *ACM Trans. Programming Languages and Syst.* 7(1):80–112.
- Gordon, M. 1988. *The Theory and Implementation of Programming Languages*. Prentice Hall, Hemel Hempstead, UK.



- Gries, D. 1981. *The Science of Programming*. Springer Verlag, New York.
- Gries, D. and Volpano, D. 1990. The transform— A new language construct. *Structured Programming*. 11(1):1–10.
- Hudak, P. 1989. Conception, evolution, and application of functional programming languages. *ACM Comput. Surv.* 21(3):359–411.
- Meyer, B. 1989. *Object-Oriented Software Construction*. Prentice Hall, Hemel Hempstead, UK.
- Paulson, L. 1992. *ML for the Working Programmer*. Cambridge University Press, Cambridge, UK.
- Peterson, G. L. 1983. A new solution to Lamport's concurrent programming problem using small shared variables. *ACM Trans. Programming Languages and Syst.* 5(1):56–65.
- Shapiro, E. 1989. The family of concurrent logic programming languages. *ACM Comput. Surv.* 21(3):359–411.
- Watt, D. A. 1990. *Programming Language Concepts and Paradigms*. Prentice Hall, Hemel Hempstead, UK.

## Further Information

A good introduction to the theory and practice of concurrent and distributed programming can be found in Ben-Ari's textbook *Principles of Concurrent and Distributed Programming*, Prentice Hall, 1990. It puts a wide range of solutions to well-known problems in perspective. Another good book on the same topic is by Greg Andrews, *Concurrent Programming: Principles and Practice*, Benjamin/Cummings, 1991. David Watt's book *Programming Language Processors*, Prentice Hall, 1993, provides a nice overview of the implementation of programming languages. New implementation techniques can be found in the proceedings of the annual ACM Conference on Programming Language Design and Implementation. Theoretical and practical aspects of programming in a wide variety of paradigms can be found in the quarterly journal *ACM Transactions on Programming Languages and Systems* and the proceedings of the biennial IEEE International Conference on Computer Languages. Foundations of programming languages are the subject of the annual ACM Symposium on Principles of Programming Languages. For information on practical issues in distributed and parallel programming, see the quarterly journal *IEEE Parallel and Distributed Technology*.

Freitas, R. "Input/Output Devices"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

### 135.1 Input/Output Subsystem

#### 135.2 I/O Devices

Human-Oriented I/O • Communications-Oriented I/O • Storage-Oriented I/O

#### Rich Freitas

*IBM Almaden Research Center*

The computer senses and affects the environment around it through input/output (I/O). In some sense its CPU characteristics are secondary; the computer is really "defined" by the quantity, variety, and attachment strategy of its I/O devices. Therefore, a full and complete understanding of computers and how to use them must be founded on a good understanding of the characteristics and capabilities of the computer's I/O subsystem and the I/O devices attached to it. This chapter presents brief descriptions of typical components of I/O subsystems and several I/O devices.

## 135.1 Input/Output Subsystem

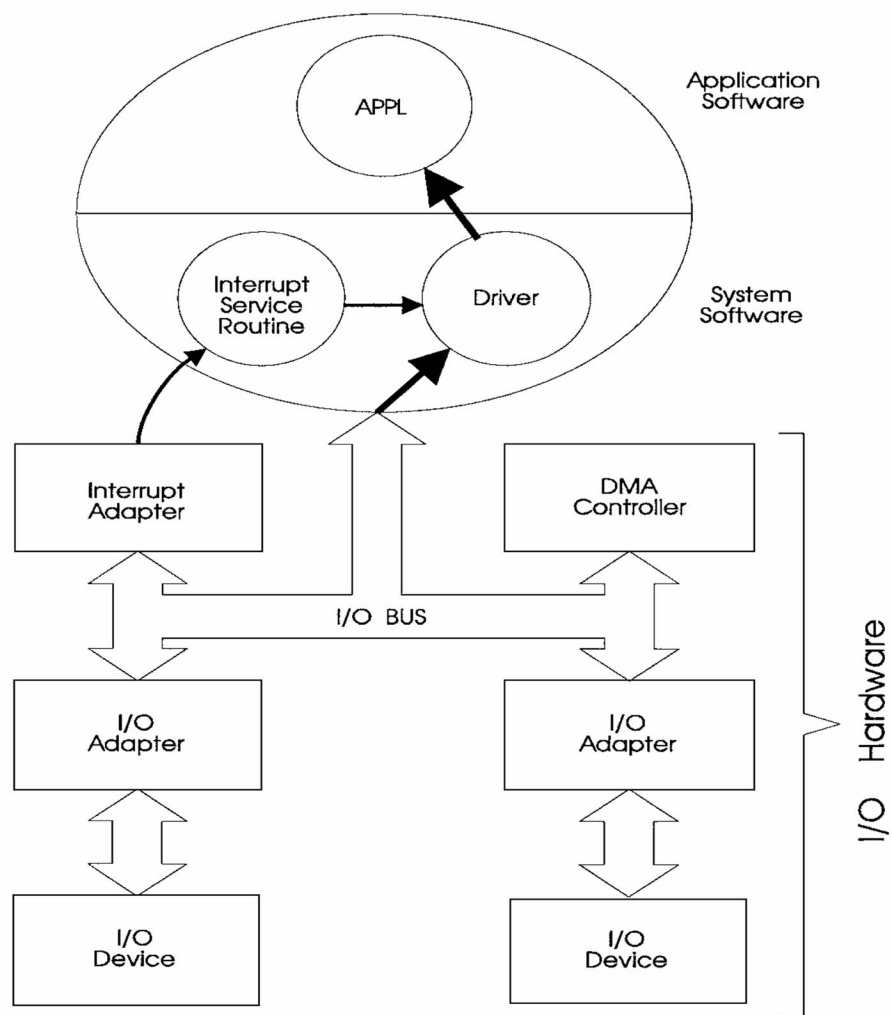
---

[Figure 135.1](#) depicts an I/O subsystem. The circle represents software and the rest is the I/O hardware. The upper half of the circle is application software, which generates or consumes the I/O data. The application software selects or names some I/O data or an I/O device. It may also specify the type of transfer and a target location in memory.

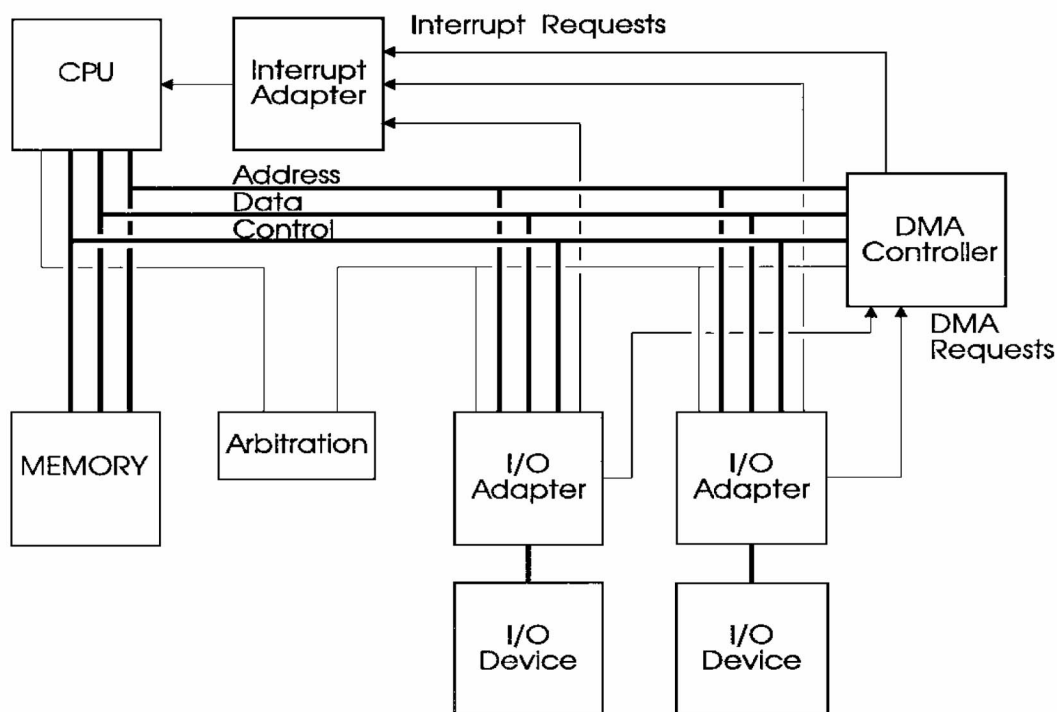
**Device drivers** are part of the system software, which includes the file system and operating system. The role of the device driver is to map the high-level names used by the application software into the addresses recognized by adapters on the I/O bus, to manage the sequence of events, and to perform error recovery. Device drivers deal with the device-specific aspects of the I/O software, and thus may have to queue the operation, select a device, address it, assess its status, check the legality of the operation, arm and enable **interrupts**, initialize the **direct memory access (DMA)** controller, start the **I/O adapter**, and deal with the resulting interrupts, status conditions, and errors. In personal computers this function is performed by a combination of the operating system (DOS, OS/2, or Windows) and the ROM resident BIOS.

The I/O hardware consists of the I/O bus, the interrupt controller, the DMA controller, the I/O adapters, and the I/O devices. The I/O bus connects the I/O hardware with the memory and the CPU on which the software runs. [Figure 135.2](#) depicts a typical I/O bus. It consists of lines for data, address, control, arbitration, and interrupts. The functions of the address, data, and control lines is straightforward; these lines provide the means and control signals to move data to and from the CPU, memory, and the adapters. The arbitration lines govern who has control of the bus. Many such buses exist in the industry, and a few are described in [Table 135.1](#).

**Figure 135.1** Input/output subsystem.



**Figure 135.2** Input/output busing.



**Table 135.1** Characteristics of Typical I/O Buses

Bus	Data Width	Single Transfer Wait State	Burst Transfer Wait State
ISA	8, 16	8.33 MB/s	8.33 MB/s
EISA	32	8.33	33
MCA	32, 64	20	80
PCI	32, 64	33/44 (R/W)	132
VMEbus	16, 32	25	27.9
FutureBus	32,...,256	37	95

ISA—industry standard architecture

MCA—micro channel architecture

PCI—peripheral component interconnect

EISA—extended ISA

VMEbus—VERSAbus-E

Interrupts are signals from the I/O adapters to the interrupt controller indicating that a hardware event has occurred, to which the software is expected to respond. The interrupt controller notifies the CPU, which suspends execution, saves its state, and executes an interrupt service routine. When the routine is finished, the CPU restores the saved state and resumes executing the interrupted program. For complicated situations, such as two or more interrupts occurring "simultaneously" or preventing interrupts in critical sections of the software, the interrupt controllers prioritize, disable, or mask interrupts. A signalled interrupt with high priority blocks all interrupts at its priority and any of lower priority. A masked interrupt is blocked because its associated mask or arm bit is off. The whole interrupt system can be masked; this is called *disabling* the interrupt system. There are interrupts that cannot be disabled or disarmed. Serious error indications (e.g., uncorrected memory error) are often reported as nonmaskable interrupts, and many kinds of programming errors (e.g., divide by zero) are reported by a lower priority nonmaskable interrupt called a *trap*.

The I/O adapter is the hardware entity directly addressed by the device driver. Its role is to "adapt" I/O devices to the shared I/O environment. It responds to commands from the device driver and then manages the sequencing and timing of operations between the I/O bus and the I/O device; does error detection and error correction; collects, maintains, and reports status information; buffers data and/or acts as a data cache; and converts or translates data from one physical or logical environment to another.

There are three I/O modes: programmed I/O, interrupt-driven I/O, and direct memory access. In programmed I/O the software has direct access and control of all I/O data, control, and status information. It writes and reads hardware registers either by using special I/O commands or, in memory-mapped I/O systems, by accessing the address in memory address spaces assigned to the hardware register. If the software needs to test for the occurrence of a hardware event, it loops reading the status register and checking for the indication of the event; this process is called *polling*. Polling I/O events is time consuming and inefficient. If the hardware event is signaled as an interrupt, then the software need not poll. This is called *interrupt driven I/O*. Programmed I/O and interrupt-driven I/O are effective for transferring control and status information and small amounts of data. However, if large block transfers of data—especially at high speed—are required, the software overhead is very high and can dramatically decrease the overall system performance.

Direct memory access, the third I/O mode, was devised to free the software from dealing with long contiguous data transfers. The DMA controller performs all the necessary tasks once the

software has selected a DMA channel (i.e., selected an adapter); stored the memory address, byte count, and transfer direction into DMA control registers; and started the selected I/O adapter. At the transfer of each byte (word, etc.) the count is decremented and the address is incremented. When the count goes to zero, the hardware prepares to stop. In some DMA controllers, additional sets of address and byte count are stored in DMA registers so that the DMA controller can continue to transfer data. Such hardware automatically continues the data transfer using this new information; this process is called chaining. If chaining is not performed, the hardware interrupts the software to indicate that the count equals zero. The software then completes the termination of the transfer. If the DMA controller manages the address and count registers for the adapter, a third-party DMA transfer has occurred. If the adapter manages the registers, a first-party DMA transfer has occurred.

## 135.2 I/O Devices

---

There are three classes of I/O devices: those devices that interface with humans, those that interface with machines, and those that interface with storage media. We will briefly discuss representative devices in each of these categories.

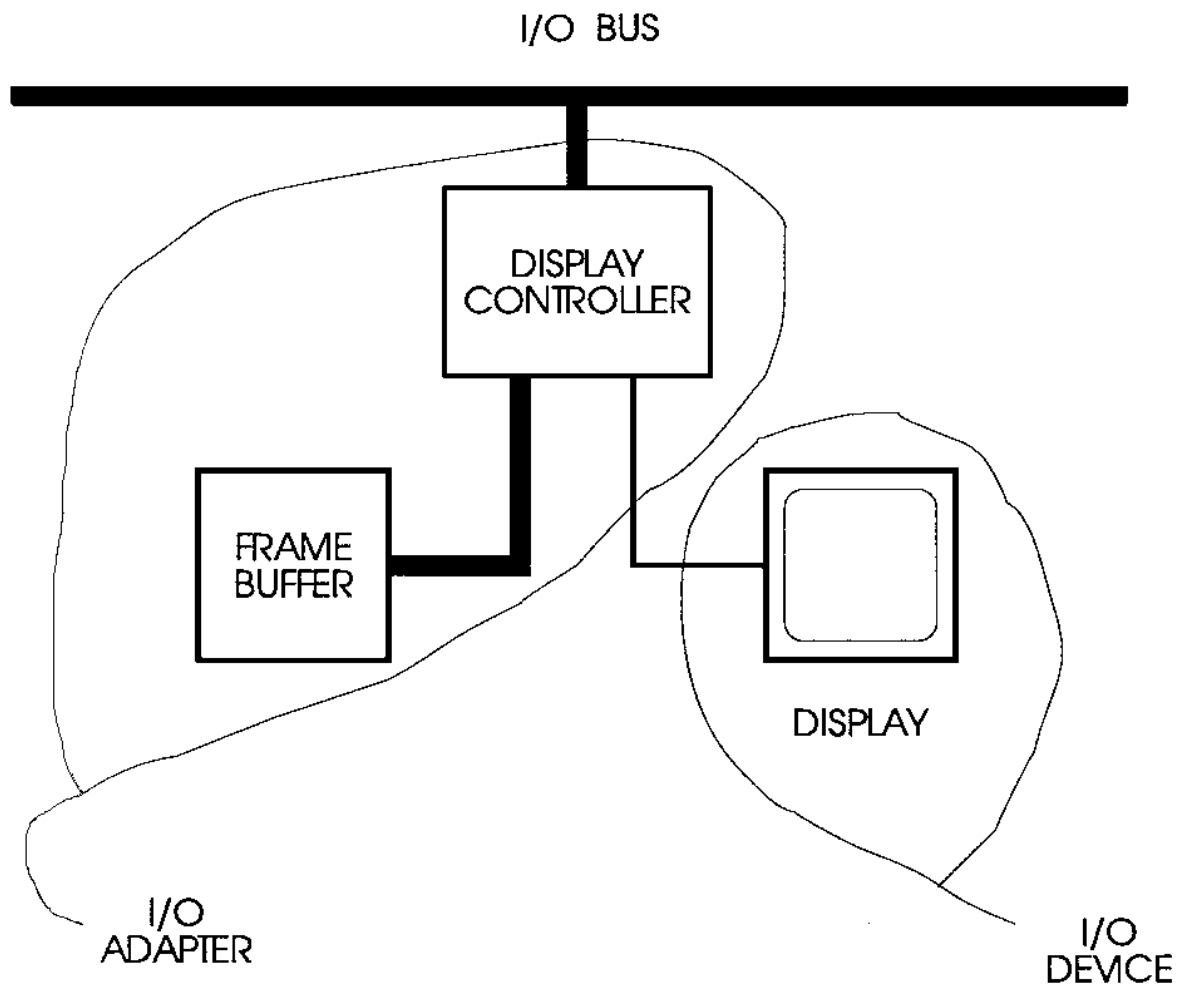
### Human-Oriented I/O

The video display; printer; keyboard; and pointing devices such as the mouse, TrackPoint II™, and tablet are typical human-oriented I/O devices—that is, a transducer and associated electronics that interface a computer with a human sense. By far the most common output device is the video display. The block diagram of a typical video display system is shown in [Fig. 135.3](#). It consists of a **frame buffer**, a **display controller**, and a display. The frame buffer contains information about the intensity and color of each picture element (pixel) on the screen. The display controller uses the information in the frame buffer to paint a picture on the display screen. If the software wishes to change that picture, it must change the contents of the frame buffer. Such changes are made through and managed by the display controller.

For desktop computers the most commonly used video display system is based on the use of the color cathode ray tube (CRT). A CRT consists of an evacuated glass vessel, electron gun(s), a beam-focusing yoke, and a screen coated with a fast phosphor. To draw a picture on the CRT's screen, the beam must irradiate every pixel on the screen that information stored in the frame buffer says should be on. The information must be fetched in order and in synchrony with the movement of the beam. In present day CRTs the beam is scanned across the screen and then down the screen. One complete scan of the screen is called a *frame*. Since the phosphor decays rapidly, the frame must be scanned many times per second to provide a stable, flicker-free image, typically 30 to 80 times per second.

CRTs fail to meet the requirements of one of the fastest-growing segments of the personal computer market: mobile or portable computing. CRTs are heavy, power-hungry, and large in volume. The requirements of mobile computing are presently being met by another display, the flat panel display. The most common flat panel display today is the **liquid crystal display (LCD)**, which consists of two sheets of glass encasing a liquid crystal. On the surface of the glass

**Figure 135.3** Display adapter.



are parallel conductors. On one sheet all the conductors are horizontal and on the other they are all vertical. At each intersection of the two perpendicular lines is a pixel. When the lines are energized the electric field formed between them causes the optical properties of the liquid crystal at the intersection to change. See [Table 135.2](#) for typical characteristics of CRT and LCD displays.

**Table 135.2** Characteristics of Typical Displays

Characteristic	CRT	LCD Flat Panel
Diagonal	12"–19"	6"–10"
Weight	Heavy	Light
Volume	Large	Small
Bandwidth	High	Medium
Refresh rate	30–80 Hz	.01–50 Hz
Brightness	20–60 fL	3–10 fL
Contrast ratio	~30 : 1	3.5–15 : 1

Another output device is the printer. It is the counterpart to the video display in that it produces a hard copy. Printers may be classified according to their technology: impact versus nonimpact (i.e., does the print head strike the paper to print the character?). They may also be classified according to their print mode (character, line, and page) or according to their mechanism (drum printers, band printers, dot matrix printers, ink-jet printers, and electrographic or laser printers). Regardless of the classification scheme chosen, printers almost always interface with the computer using one of two interface standards. One is serial; it is the RS232 protocol, which is used to asynchronously ship data at moderate speed to the printer over potentially long lines. The other is the byte-parallel, 25-pin centronics interface; it is used to send data to the printer at higher speed but much shorter distances.

Typical input devices include the keyboard and pointing devices. The keyboard is an electrical device for translating pressure on one or more of the keys into a 7- or 8-bit number. That transformation is effected by a matrix of switches monitored by a small microcomputer housed in the keyboard. Once the microcomputer has determined a key code, it formats the code into a serial bit stream and sends it over the keyboard cable to an adapter in the computer. The computer may also send messages to the keyboard to drive the buzzer and turn on indicator lights.

Pointing devices such as the mouse, TrackPoint II™, and tablet are used to select a point on the computer display. This point is indicated by an icon called the *cursor*. The software picks a (0,0) point and then derives the cursor's position by keeping track of the  $\delta x$  and  $\delta y$  registers maintained by the pointing device hardware. Associated with both the mouse and the TrackPoint II™ are two or three buttons whose state is returned by the hardware.

The mouse consists of a ball and two shaft encoders mounted at right angles to each other. Attached to each shaft encoder is a wheel. As the mouse is moved over the surface, the ball moves and the wheels, which touch the ball, turn proportionally with the  $x$  and  $y$  displacement; the shaft encoders generate pulses accordingly. These pulses increment and/or decrement the  $\delta x$  and  $\delta y$  registers. These registers are typically sampled at 5 ms intervals, and the changes are reported to the computer. The mouse is separate from the computer and keyboard and tethered to the computer



by a cable over which power and the sampled data are sent.

The TrackPoint II™, on the other hand, is built into the keyboard. It is a small shaft that sticks up between the G, H, and B keys of the keyboard. At its base are a set of strain gages. When the user pushes on the shaft, the pressure is measured and used to determine the position of the cursor. Basically, the vector sum of the strain gages indicates the direction in which the cursor should move, and its speed is a nonlinear function of the magnitude of the force applied. For "small" forces the speed is zero. For "medium" forces the speed is linearly proportional to the force applied, and for "high" forces the speed is a constant.

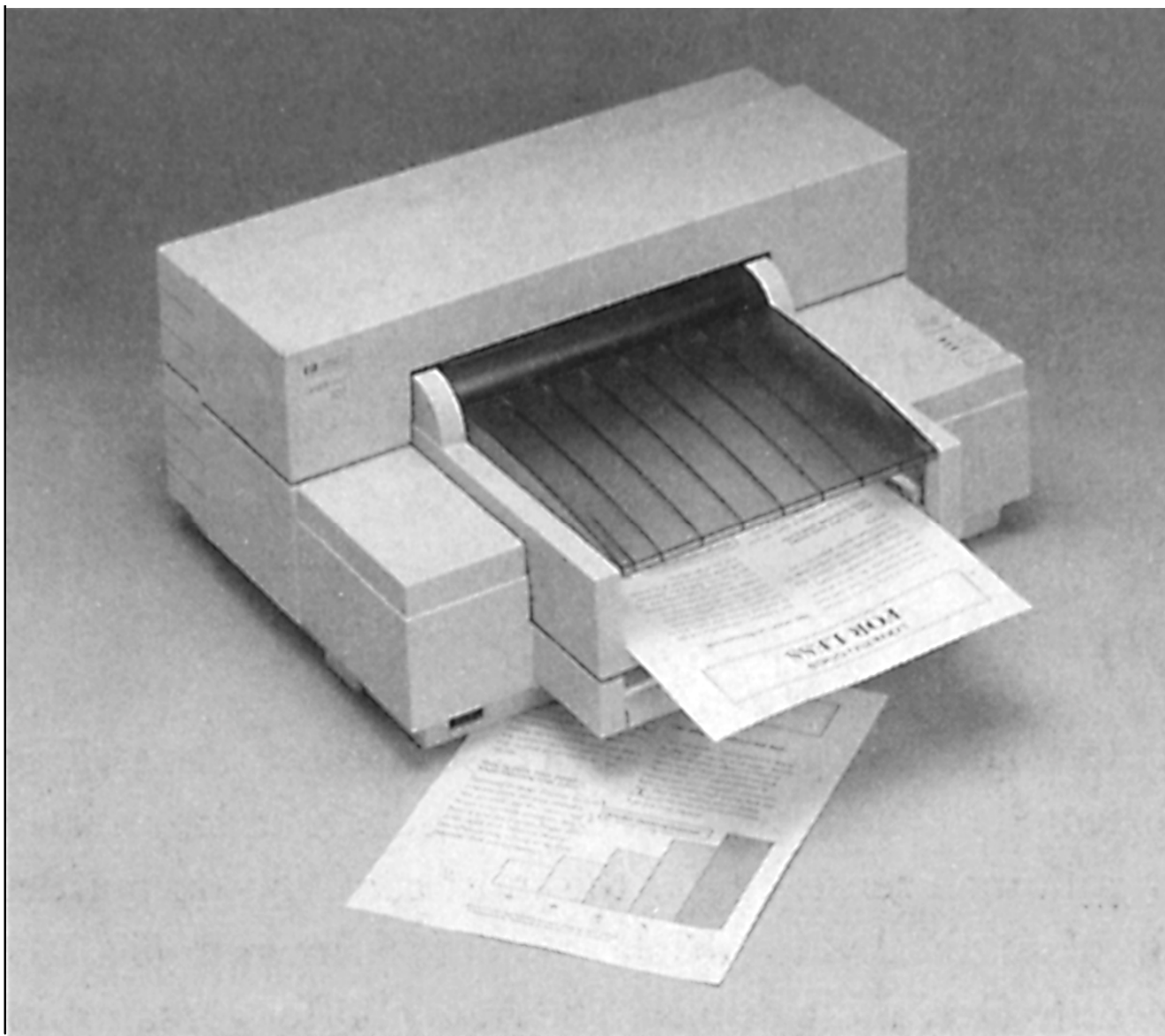
The pen and tablet constitute a different sort of pointing device. The display screen, usually a flat panel screen such as an LCD, is overlaid with a transparent mesh of wires. Each line in the  $x$  and  $y$  dimension is individually pulsed in sequence. When the pen touches the screen, it either capacitively senses or electromagnetically modifies pulses on the  $x$  and  $y$  lines that cross under it. From knowledge of the timing of the pulses, the associated hardware determines the specific  $x$  and  $y$  lines. The hardware reports this position information to the computer as an absolute ( $x, y$ ), instead of relative ( $\delta x, \delta y$ ), value as used by the mouse and the TrackPoint II™. Sampling rates are about the same—200 points per second.

See [Table 135.3](#) for typical human-machine data communications rates.

**Table 135.3** Human-to-Machine Data Rates

Human	Data Rate	Limited by	Machine
Read text at 1000 wpm	.083 kB/s	H	Alphanumeric display
Visual pattern recognition	30–70 MB/s	M	Graphic display
Type at 100 wpm	.008 kB/s	H	Keyboard
Pointing/selecting	1 kB/s	H	Mouse; TrackPoint II™; tablet
Hard copy	1–100 kB/s	M	Printer

The Hewlett-Packard DeskJet printers are the best-selling monochrome and color-optional printers in the world. (Photo courtesy of Hewlett-Packard.)



#### AFFORDABLE DESKJET PRINTERS FOR THE HOME MARKET

Hewlett-Packard Company has introduced the HP DeskJet 540 printer for PCs and the HP DeskWriter 540 printer for Macintosh. These two printers replace best-selling monochrome and color-optional printers previously offered by HP. They are targeted for the rapidly growing home market and offer the best black print quality at an affordable price. HP DeskJet printers offer color as a standard or as an option. Affordable color upgrade kits for these printers give HP the distinction of having the lowest price available for color printing.

HP's ColorSmart technology, which is used in the printers, renders vivid, vibrant color automatically. ColorSmart uses object identification to recognize text, graphics, and photographic elements separately, selecting the optimal color or grayscale tone for each element. By reducing the process to a single step, ColorSmart has revolutionized color printing in much the same way that autofocus cameras have reduced the expertise required for 35 mm photography.

## Communications-Oriented I/O

Communications-oriented I/O devices make up the computer's vehicle for accessing computer networks to share resources and send messages (e.g., the Internet). See Tanenbaum [1988] for more information about computer networks. The "communications device" is an I/O adapter interfacing with a communications channel(s). A channel is a single path or line over which data is transferred. Channels are either simplex (a single, strictly one-way path), half duplex (a shared path used part-time in each direction) or full duplex (two simplex paths connected in opposite directions). The media for the channels can be twisted-pair wires, coaxial cables, fiber optic cables, radio waves, and so forth. See Table 135.4 for a few examples of communications devices. The I/O behavior of the I/O device reflects the characteristics of the channel. For example, low-speed communications over a single telephone modem can easily be handled by a background interrupt-driven device driver, whereas FCS communications at 100 MB/s requires a complex, high-speed adapter performing DMA transfers over a high-speed bus to a high-performance memory system.

**Table 135.4** Machine-to-Machine I/O

Medium	Bandwidth	Mode*
Telephone modem	1.2–19.2 kb/s	C
Wireless wide-area net	4–19.2 kb/s	P
Wireless LAN	200 kb/s	P
Ethernet	10 Mb/s	P
Token ring	16 Mb/s	P
FDDI	100 Mb/s	P
FCS	100 Mb/s	C or P
ATM	51.84 Mb/s–2.488 Gb/s	C or P

\*C—connection; P—packet

## Storage-Oriented I/O

Storage-oriented I/O devices exploit physical phenomena to provide high-capacity, inexpensive, nonvolatile data storage. Table 135.5 lists several such storage devices. With the exception of flash memory, all provide reliable permanent data storage with access times much slower than DRAM. Since flash memories are based on semiconductor memory technology, they provide a storage subsystem with speeds comparable to main memory. Unfortunately, this technology suffers from write wear-out. In general, storage devices interface with computers as a DMA device with a long access time and a moderate transfer rate. Once they start to transfer, they need to continue transferring. Therefore, most of the adapter cards that support such devices have an onboard buffer to help match the speed differences and in some cases act as a cache.

**Table 135.5** Machine-to-Storage I/O

Device	Data Rate	Access Time	Latency	Formatted Capacity
Floppy disk	.125–1 MB/s	50–150 ms	100–200 ms <sup>1</sup>	.36–2.88 MB
Flash memory	10 MB/s	.0001 ms	—	—
Disk drive	1–10 MB/s	8–20 ms	7–16 ms	.02–2 GB
CD ROM	.2–5 MB/s	200–500 ms	50–150 ms <sup>2</sup>	.6–2 GB
Tape	.2–2 MB/s	—	30–90 s	2 GB

<sup>1</sup>500 ms spin up time

<sup>2</sup>6 s spin up time

## Defining Terms

**Device driver:** Modern operating systems and applications are written to be portable; that is, they are written to run on many computers and with many different I/O configurations and with I/O devices that did not exist when the software was written. Therefore, all I/O device-specific code needed to use a device is written as a separate program, either by a device vendor or the user, thus permitting new devices and upgrades to old devices without requiring changes to the operating system or the application. This separate program is called a **device driver**.

**Direct memory access (DMA):** The method of using hardware, either the I/O adapter or a specialized controller, to transfer I/O data directly between the memory and the I/O device without the direct involvement of the software.

**Display controller:** Digital electronics that controls access to a display's frame buffer.

**Frame buffer:** Each dot on a display screen is described by a small block of data. The information stored there ranges from whether the dot should be present to intensity information for all three colors. This information is stored in a memory called a **frame buffer**.

**Interrupt:** An interrupt is a signal from an I/O adapter to the software indicating that a hardware event has occurred. The event is usually expected by the hardware, and the interrupt causes the software to start executing a higher-priority task.

**I/O adapter:** Hardware interface between I/O bus and I/O device.

**Liquid crystal display (LCD):** A light, low-power, flat panel display used primarily in laptop and smaller computers.

## References

- Comerford, R. 1994. The multimedia drive. *IEEE Spectrum*. 31(4):77–83.
- Goupille, P.-A., translated by Loadwick, G. M. 1993. *Introduction to Computer Hardware and Data Communications*. Prentice Hall, Hertfordshire, UK.
- Hennessy, J. L. and Patterson, D. A. 1990. *Computer Architecture: A Quantitative Approach*. Morgan Kaufmann, San Mateo, CA.
- Mee, C. D. and Daniel, E. D. 1990. *Magnetic Recording Handbook: Technology and*

*Applications*. McGraw-Hill, New York.

Peddie, J. 1994. *High Resolution Graphics Display Systems*. Wincrest/McGraw-Hill, New York.

Selker, T. and Rutledge, J. D. 1990. Force-to-motion functions for pointing. *INTERACT '90*. 701–705.

Sherr, S. 1979. *Electronic Displays*. John Wiley & Sons, New York.

Shiva, S. G. 1985. *Computer Design and Architecture*. Little, Brown & Company, Boston, MA.

Tanenbaum, A. S. 1988. *Computer Networks*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.

## **Further Information**

For information on a particular computer's I/O subsystem, the best resource is the computer manufacturer's documentation. For more general technical information, the various IEEE Computer Societies are a good source: *Transactions on Computers*, *Computer* magazine, *Design and Test*, and so forth. For information on storage, the *IEEE Transactions on Magnetics* is a good source. Information on I/O directed at more general audiences can be found in McGraw-Hill's *BYTE* magazine and in Ziff-Davis's *PC Magazine*.

Chen, P. M. "Memory and Mass Storage Systems"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Memory and Mass Storage Systems

---

## 136.1 Aspects of Storage Devices

## 136.2 Types of Storage Devices

Solid-State Storage Devices • Magnetic Storage Devices • Optical Storage Devices

## 136.3 Storage Organization

### **Peter M. Chen**

*University of Michigan*

Memory and mass storage systems compose one of the core components of a computer system. The basic task of a computer is to process data, and memory and mass storage systems provide a fundamental aspect of this data processing—that of remembering information.

To appreciate how fundamental the storing and retrieving of information is to computers, let us examine a typical transaction at a bank's automated teller machine (ATM). The customer first feeds his ATM card into the bank machine. The magnetic strip on the ATM card stores the account number of the customer. The ATM contains a storage device that can read this magnetic strip and feed it into the bank's computer. The bank's computer runs a program to interact with the customer; this program typically resides in the computer's memory. The screen that displays the menus and information for the customer uses memory to remember what to display. The customer then interacts with the bank's computer, providing a password and one or more account transactions. The password that the customer enters is verified by comparison to the password stored in the computer. Account transactions manipulate the account balances; these too are stored in the computer. In addition, the computer stores a log of all transactions. Video cameras are often present at the ATM to capture the faces of ATM users. These video cameras and tapes are also a type of memory and mass storage.

As is evident from the above example, computers need memory and storage to function. Memory and storage devices store the instructions that computers follow, the databases that computers manipulate, and the screens that computers display.

*Memory* typically refers to devices that are closer to the processor and can be accessed quickly, whereas *storage* typically refers to devices that are farther away. In the rest of this chapter the term *storage* will refer to all of these devices.

## 136.1 Aspects of Storage Devices

---

Storage devices are used throughout the computer system. No single type of storage device is appropriate for all of these uses; rather, there is a wide variety of storage devices. Choosing an

appropriate device for a particular purpose requires understanding the ways that devices differ and weighing the trade-offs involved in using various devices [Chen and Patterson, 1993] .

*Performance* refers to the speed of a storage device and can be broken down into two parts, **throughput** and **latency**. Throughput is the rate at which a device can accomplish work; this work typically is storing and retrieving data. Throughput is often expressed as bytes per second or requests per second. Latency is the time it takes to do a portion of work. The ideal storage device has high throughput and low latency.

*Reliability* refers to the rate at which the storage device fails; this can also be inverted to express the expected time between failures. For example, devices may quote failure rates of 1 error every 100 trillion accesses or 1 failure every 10 years.

*Capacity* refers to the amount of data that a device can store. Closely related to this concept is cost, since a customer can usually buy more devices to increase capacity. Cost is related to all metrics, since it is usually possible to get more by spending more.

*Volatility* refers to whether or not a device can retain information after power is turned off. Volatile devices need power to store information; nonvolatile devices do not. Of course, both volatile and nonvolatile devices require power to store new information or retrieve old information. Volatility can be viewed as a component of reliability, since a power failure is one way to cause a volatile device to fail. Devices discussed here can be assumed to be nonvolatile unless otherwise noted.

*Writable* refers to whether a device can store new information. Almost all applications require the ability to read a storage device, but many do not require the ability to write new information to the device. Devices discussed here can be assumed to be writable unless otherwise noted.

**Random access** refers to the ability of a device to quickly store and retrieve any information in any order. Some devices only access information **sequentially**, that is, in an order that depends on the location of the stored information. Other devices can access information randomly, that is, fairly independently of where the data are stored.

Table 136.1 provides a summary of the ways that storage devices differ.

**Table 136.1** Aspects of Storage Devices

Type	Summary Question	Units	Examples
Throughput	How much work can be done in a given time?	Bytes/second, requests/second	This device can sustain a throughput of 5 million bytes per second.
Latency	How long does it take to do work?	Seconds	This device responds with an average latency of 10 milliseconds.
Reliability	What is the frequency at which this device fails?	1/second, 1/access	This device returns the wrong information once every 100 trillion accesses. You can expect this device to work for 10 years before failing.
Capacity	How much information can this	Bytes	This device can store 1 billion



	device store?		bytes of information.
Cost	How much does this device cost?	Dollars	This device costs \$1000.
Volatility	Does this device lose information if power is turned off?	Yes or no	This device will retain information for 100 years without power.
Writable	Can this device store new information?	Yes or no	This device can store new information, but it cannot erase information once it is stored.
Random-access	Can this device access information independently of how that information was stored?	Yes or no	This device will read information 1000 times faster if the data are read in consecutive order.

---

## 136.2 Types of Storage Devices

---

Numerous types of storage devices exist, and new ones are constantly being invented. Current storage devices can be roughly categorized into three groups—solid-state, magnetic, and optical—but the volatile nature of this field makes it difficult to predict what methods will be used to store and retrieve data in ten years.

### Solid-State Storage Devices

Solid-state storage devices use integrated circuit technology to store information. For most types of solid-state memories, the information is stored by manipulating charges, that is, electrons. These electrons may be stored in explicit capacitors or implicit capacitors (such as those that cause propagation delay), or they may be trapped in transistor gates. Devices that store charge in capacitors are volatile because capacitors allow charge to slowly leak away; devices that trap charge in gates are typically nonvolatile. For a few types of solid-state memories—specifically, ROM (read-only memory) and PROMS (programmable read-only memory)—devices have a set of connections, and the pattern of these electrical connections (fuses) stores information. Solid-state storage devices are currently by far the fastest type of storage.

DRAM (dynamic random access memory) is currently the densest type of solid-state storage device. DRAMs use only one transistor and one capacitor to store one **bit** (binary digit) of information [Wakerly, 1994]. Though they are very dense, DRAM chips are relatively slow for solid-state memory. The charge stored in the capacitor slowly leaks away. To store information for more than a few milliseconds, the charge on these capacitors must be periodically restored; this process is called *refresh*. Because of the high density of DRAMs, they are typically used to store the main memory of a computer system.

SRAM (static random access memory) is optimized for speed rather than density and uses four to six transistors to store a bit of information. Unlike DRAMs, SRAMs do not require periodic refresh. Because of their high speed, SRAMs typically store the data that the processor uses most

frequently.

There are several types of solid-state storage devices that are designed primarily for retrieving data. These include ROM (read-only memory), PROM (programmable read-only memory), EPROM (erasable, programmable read-only memory), and EEPROM (electrically erasable, programmable read-only memory). ROMs and PROMs store information with a set of electrical connections; they can be set only once. EPROMs and EEPROMs store data in the floating gates of transistors; this allows them to be erased and then written with fresh information. However, erasing and writing data is very slow in EPROMs and EEPROMs (many seconds to several minutes) and may only be done a limited number of times (such as 10 000).

## Magnetic Storage Devices

Magnetic storage devices store most of the information for nearly all computer systems, from personal computers to large commercial computers. These storage devices store information on a thin layer of magnetic coating. A bit of information is stored by changing the polarity of tiny magnetic regions in this coating. A **read/write head** creates the magnetic field that induces this change in polarity. Because the state is stored in the polarity of magnets, magnetic storage devices are nonvolatile. In some magnetic storage devices the **storage media** may even be removed from the actual device, allowing the user to multiply greatly the amount of data that can be accessed by a single device.

Magnetic storage devices deposit the magnetic coating on distinct substrates. A hard disk drive deposits the coating on rigid platters. These platters spin underneath an electromagnetic read/write disk head. Because a hard disk is rigid and hermetically sealed, it can use very tight tolerances to pack a high number of bits per square inch and to spin the platters at high speeds. As a result, the hard disk currently has the best performance of any magnetic storage device. Hard disks are the main storage device used in nearly all computer systems.

Floppy disks deposit magnetic coating on a flexible platter. The flexible substrate and unsealed nature of floppy disk drives create looser tolerances than hard disks, which lowers density and performance. Because the floppy disk is inexpensive and can be removed from its drive, it serves as an ideal medium for exchanging data.

Magnetic tapes deposit the magnetic coating on a spool of flexible tape. Examples of machines that use magnetic tape include VCR machines and audio cassette players. The flexible tape is rolled up, which has two significant effects. First, accessing data requires winding through the tape until the head reaches the location on the tape where the data are stored. This winding makes the speed of accessing data in nonsequential locations prohibitively slow, so tapes are read and written sequentially. Second, rolling the tape enables extraordinarily high volumetric densities compared to devices that use platters. Magnetic tapes are also removable, which further decreases their effective cost. Because of their low cost and slow performance, magnetic tapes are used primarily to archive unused data, though this may be changing as the amount of data needed by applications increases.

# Optical Storage Devices

Optical storage devices use lasers to write and/or read data, either alone or in conjunction with magnetic disk technology. Lasers can be focused to read or write a very small spot on the disk, which enables the optical device to store information very densely on the medium. Current optical storage devices allow the removal of the medium that contains the data, which makes them particularly useful for archiving large amounts of infrequently used data. The main disadvantage of optical devices is their speed—magnetic disks are currently several times faster than the fastest optical device.

There are several methods for writing and reading optical storage devices. WORM (write-once, read-many) disk drives were an early type of optical disk. As the name implies, each spot on this device can only be written once. The device writes data by using a laser to burn a pit in the medium at various locations. Future-read accesses sense how the laser is reflected off those pits. CD-ROMs (compact disk read-only memory) are similar to WORMs but are written en masse by stamping a pattern of pits on the disk. These disks cannot be modified; they are used throughout the music and computer industry to inexpensively distribute large quantities of information.

Magneto-optical disk drives use both magnets and optics to write data. As in magnetic disks, the drive uses a magnetic write-head to change the polarity of regions on the magnetic medium. To achieve higher densities than magnetic disk drives, magneto-optical disk drives use a laser to heat up a small part of the medium; only the heated portion of the medium is actually changed by the magnet. To read the stored information, magneto-optical disks sense how the magnetic polarities change the reflection of the laser. Current magneto-optical disks require additional time to prepare the medium before writing new information; this slows down performance considerably.

Phase-change optical disks use relatively new technology and were designed to improve performance over magneto-optical disks. Phase-change optical disks use lasers to heat small regions on the medium; this heat changes the region into one of two states. The state of a region can then be read by measuring how the region reflects laser beams.

Table 136.2 shows characteristics for a range of storage devices.

**Table 136.2** Characteristics for a Range of Storage Devices

	SRAM (1 chip)	DRAM (1 chip)	Magnetic Floppy Disk	Magnetic Hard Disk	Magnetic Tape	Magneto-Optical Disk
Technology	Solid-state	Solid-state	Magnetic	Magnetic	Magnetic	Magnetic and optical
Sustainable throughput (bytes per second)	10 <sup>7</sup> –10 <sup>8</sup>	10 <sup>7</sup>	10 <sup>5</sup> –10 <sup>6</sup>	10 <sup>6</sup> –10 <sup>7</sup>	10 <sup>6</sup> –10 <sup>7</sup>	10 <sup>5</sup> –10 <sup>6</sup>
Latency for random access (seconds)	10 <sup>–9</sup> –10 <sup>–7</sup>	10 <sup>–8</sup> –10 <sup>–7</sup>	10 <sup>–2</sup> –10 <sup>–1</sup>	10 <sup>–3</sup> –10 <sup>–1</sup>	10 <sup>1</sup> –10 <sup>3</sup>	10 <sup>–2</sup> –10 <sup>–1</sup>
Capacity (bytes)—for removable media, assume 1 medium	10 <sup>3</sup> –10 <sup>6</sup>	10 <sup>5</sup> –10 <sup>7</sup>	10 <sup>6</sup> –10 <sup>7</sup>	10 <sup>8</sup> –10 <sup>10</sup>	10 <sup>8</sup> –10 <sup>10</sup>	10 <sup>8</sup> –10 <sup>10</sup>

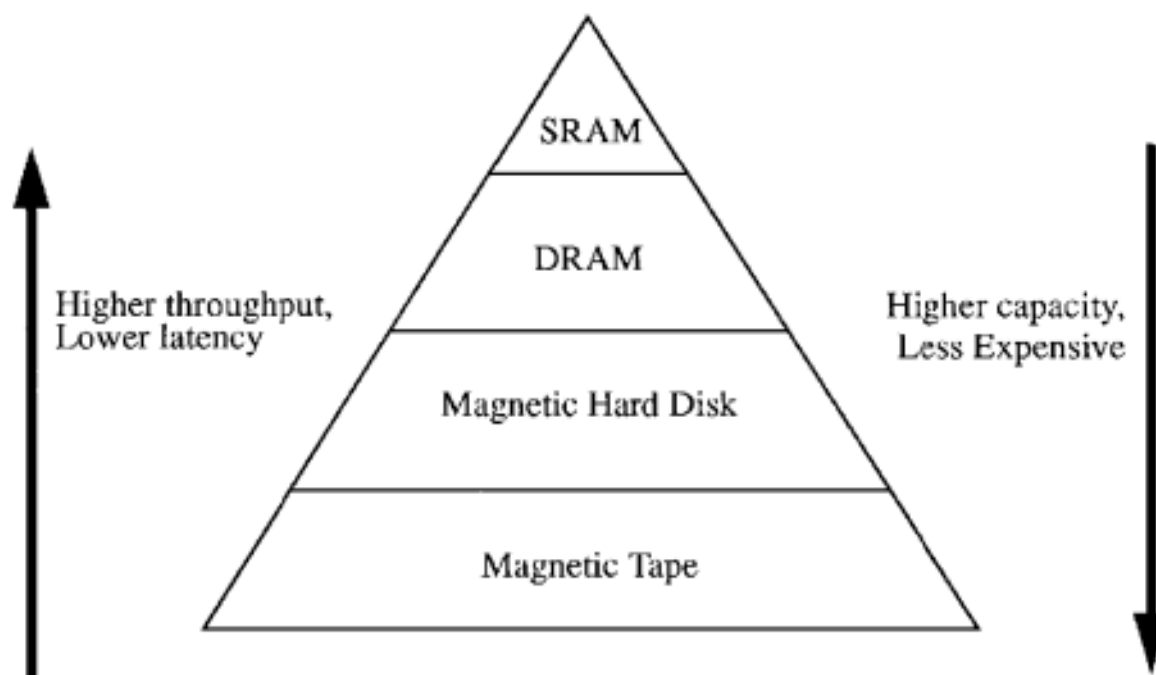
Cost per byte (dollars)—for removable media, assume 100 media per drive	$10^{-4}$ – $10^{-3}$	$10^{-5}$ – $10^{-4}$	$10^{-7}$ – $10^{-6}$	$10^{-7}$ – $10^{-6}$	$10^{-9}$ – $10^{-7}$	$10^{-7}$ – $10^{-6}$
Volatile	Yes	Yes	No	No	No	No
Removable	No	No	Yes	No	Yes	Yes

These approximate numbers for 1994 give a rough indication of aspects of various storage technologies. The rapid pace of innovation in the storage area can lead to changes of up to 50% per year, yet these figures help us to appreciate the diversity of available storage technologies.

## 136.3 Storage Organization

A typical computer system contains many types of storage devices, ranging from very fast and expensive devices such as SRAM, to slow, inexpensive devices such as magnetic tape. They are almost always organized into a **hierarchy** of levels, as in Fig. 136.1[Katz, 1992]. The goal of such an organization is to achieve the speed of the fastest, most expensive device, but to do so at the cost of the slowest, least expensive device. To achieve low average cost per **byte**, the hierarchy allocates more bytes of storage to the cheaper devices. To accomplish high average performance, computer systems store the most frequently used data in the fastest devices. Since average access time is weighted by the frequency of access, the access time of the fastest, most frequently used devices dominates the overall performance.

**Figure 136.1** Storage hierarchy.



This logical organization corresponds well to how devices are connected to the computer system. Devices that are close to the computer's processing unit can be accessed more quickly than devices that are far away. It is thus natural to place fast, frequently accessed devices such as SRAM and

DRAM close to the processing unit. Devices that are far away take longer to access, but the greater distance gives more room for expansion. It is thus natural to place devices that store more information, such as magnetic disk and tape, farther away from the processing unit.

## Defining Terms

**Bit:** One binary digit of information; a bit can store a 0 or 1.

**Byte:** Eight bits of information. This is enough to hold one alphabetic character.

**Latency:** The length of time a system takes to do work, also called *response time*.

**Random access:** Describes a storage device that can access any of its locations in approximately the same amounts of time.

**Read/write head:** The portion of a storage device that changes and/or senses the state of the storage media.

**Sequential access:** Describes a storage device that takes significantly longer to access nonconsecutive information.

**Storage hierarchy:** An organization of multiple storage devices that combines a large amount of a slow, cheap device with a small amount of a fast, expensive device. A storage hierarchy may consist of many types of devices.

**Storage medium (or medium):** The part of the storage device that physically stores data. Some devices allow this medium to be removed from the storage device.

**Throughput:** The rate at which a system can accomplish work, also called *bandwidth*.

## References

Chen, P. M. and Patterson, D. A. 1993. Storage performance—metrics and benchmarks. *Proc. IEEE*. 81(8):1151–1165.

Katz, R. H. 1992. High-performance network and channel-based storage. *Proc. IEEE*. 80(8): 1238–1261.

Wakerly, J. F. 1994. Memory. In *Digital Design: Principles and Practices*, 2nd ed., pp. 723–762. Prentice Hall, Englewood Cliffs, NJ.

White, R. M. 1980. Disk-storage technology. *Sci. Am.* 243(2):138–148.

## Further Information

In the U.S. the most relevant professional organizations that sponsor publications and conferences related to memory and mass storage systems are the Institute of Electrical and Electronics Engineers (IEEE) and the Association for Computing Machinery (ACM). In particular, *ACM Transactions on Computer Systems*, *IEEE Transactions on Computers*, *IEEE Journal of Solid-State Circuits*, IEEE symposia on mass storage systems, ACM symposia on operating systems principles, IEEE/ACM international conferences on architectural support for programming languages and operating systems, and IEEE/ACM international symposia on computer architecture are excellent sources for up-to-date technical information on designing, building, and using memory and mass storage systems.

von Maltzahn, W. W. "Measurement and Instrumentation"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



The HP E5050A Colloid Dielectric Probe is a Hewlett Packard instrumentation system used to measure permittivity on colloidal liquids. It comprises a colloidal dielectric probe that can be immersed into the liquid material; a general purpose precision LCR-meter for sampling data, eliminating nonlinearities, and calculating LCR component values; and a personal computer and monitor for data processing, data display, and data storage. Most modern precision instrumentation systems consist of similar components: a sensor/transducer unit, a signal processing unit, and a computer. (Photo courtesy of Hewlett-Packard Company.)

# Measurement and Instrumentation

---

**Wolf W. von Maltzahn**

*University of Karlsruhe*

**137 Sensors and Transducers** *R. L. Smith*

Physical Sensors • Chemical Sensors • Biosensors • Microsensors

**138 Measurement Errors and Accuracy** *S. J. Harrison*

Measurement Errors, Accuracy, and Precision • Estimating Measurement Uncertainty • Propagation of Measurement Uncertainty • Uncertainty Analysis in Experimental Design

**139 Signal Conditioning** *S. A. Dyer*

Linear Operations • Nonlinear Operations

**140 Telemetry** *S. Horan*

Telemetry Systems • Frame Telemetry • Packet Telemetry

**141 Recording Instruments** *W. Owens*

Types of Recording Instruments • Methods of Data Recording • Future Directions: Distributed Data Acquisition and Recording

**142 Bioinstrumentation** *W. W. von Maltzahn and K. Meyer-Waarden*

Basic Bioinstrumentation Systems • Applications and Examples • Summary

HUMAN BEINGS, CONFINED TO THE WORLD inside of their skin, experience the world outside through the five senses. These senses are extremely sophisticated and exceptionally sensitive to physical events important to our survival and are surprisingly insensitive to almost all other events. To extend the sensitivity of our five senses and to experience physical phenomena outside the reach of our senses, we have invented and developed ever more complicated instruments and devices. For instance, we now probe the world of atoms and molecules with the atomic force microscope and explore outer space with the Hubble Space Telescope. Numerous other instruments measure the domains between these microscopic and macroscopic worlds. In addition to extending our knowledge of the physical world, we want to harness nature's forces for our own purposes. Modern methods of controlling these forces, such as controlling the combustion cycle of an automobile engine, depend on measuring many process variables and determining appropriate responses. Special sensor-transducer units, sensitive to one variable and insensitive to others, measure such process variables and convert them to an electrical voltage or current. Once available as electrical signals, they can be amplified, filtered, combined with other signals, or otherwise processed to yield a control signal for an actuator, to drive an output device, or to be stored on magnetic tape or optical disk.

Fundamental to the scientific understanding and control of the physical world are measurements. A measurement determines the value or magnitude of a physical quantity or variable and expresses the result in terms of a number and a unit. The number reflects the value of the measurement, while



the unit communicates the scale of the measurement. The scientific and engineering communities use the SI system of units, although the English system of units continues to be used in some disciplines.

Almost all measurements require instruments to determine the value of a physical quantity. Such instruments may be as simple as a measuring tape, pendulum, or cup or as complicated as the Hubble Space Telescope. Complicated instruments may use elaborate schemes to transform the physical quantity to be measured into a numerical value. Scientific research instruments require high accuracy, precision, and reliability, while their cost and ease of operation is usually not important. Medical and diagnostic instruments also need to be highly accurate and precise, but their cost is a major factor for health care providers and their ease of operation is important to clinicians. In contrast, instruments in commercial products often require neither high accuracy nor precision.

A sensor interacts with the physical quantity to be measured, either directly or indirectly. Sensors are designed with three main goals in mind: to provide an output signal proportional to the measured physical quantity, to be insensitive to other physical quantities, and to minimize the unavoidable disturbance of the measured physical quantity and its environment. A transducer changes one form of energy into another one. In modern instrumentation systems the output of a transducer is typically an electrical voltage or current. Sensors and transducers often form one physical unit. Signal conditioning, in general, refers to any linear or nonlinear process that changes an analog or digital signal. Analog signal conditioning units may include amplifiers, filters, rectifiers, triggers, comparators, or wave shapers, and digital signal processing consists mostly of process algorithms implemented on microprocessors. Analog or digital signals can be stored on magnetic tape, displayed on a liquid crystal display, recorded on a strip chart recorder, or transmitted wirelessly to another device (telemetry). The topic of bioinstrumentation deals with measuring and monitoring electrophysiological variables on living human beings. The engineer designing biomedical instruments needs to know the physical and engineering principles behind these measurements and also understand how living biological systems work.

Unfortunately, this introduction to measurement and instrumentation is brief and incomplete; more complete information is available to the reader in the reference materials listed at the end of each chapter.

Smith, R. L. "Sensors and Transducers"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## 137.1 Physical Sensors

Temperature Sensors • Displacement and Force Sensors • Optical Radiation Sensors

## 137.2 Chemical Sensors

Ion-Selective Electrode • Gas Chromatograph

## 137.3 Biosensors

Immunosensor • Enzyme Sensor

## 137.4 Microsensors

### Rosemary L. Smith

*University of California, Davis*

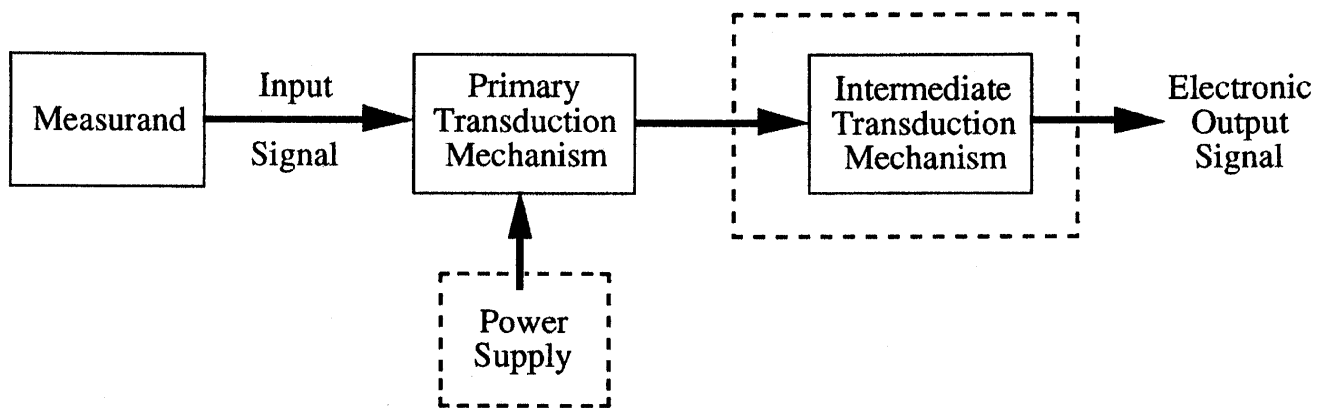
**Sensors** are critical components in all measurement and control systems. Along with the always-present need for sensors in science and medicine, the demand for sensors in automated manufacturing and processing is rapidly growing. In addition, small, inexpensive sensors are finding their way into all sorts of consumer products, from children's toys to dishwashers to automobiles. Because of the vast variety of useful things to be sensed and sensor applications, sensor engineering is a multidisciplinary and interdisciplinary field of endeavor. This chapter introduces some basic definitions, concepts, and features of sensors and illustrates them with several examples. The reader is directed to the references and further information for more details and examples.

The terms *sensor*, **transducer**, *meter*, *detector*, and *gage* are often used synonymously. Generally, a transducer is defined as a device that converts energy from one form to another. However, the most widely used definition for sensor is that which has been applied to electrical transducers by the Instrument Society of America [[ANSI, 1975](#)]: "Transducer—A device which provides a usable output in response to a specified measurand." The measurand can be any physical, chemical, or biological property or condition to be measured. A usable output refers to an optical, electronic, or mechanical signal. Following the advent of the microprocessor, a usable output has come to mean an electronic output signal. For example, a mercury thermometer produces a change in volume of mercury in response to a temperature change via thermal expansion. The output is the change in height of the mercury column and not an electrical signal. A thermometer is still a useful sensor since humans sense the change in mercury height using their eyes as a secondary transducing element. However, in order to produce an electrical signal for use in a control loop, the height of the mercury has to be converted to an electrical signal, for example, using optical or capacitive effects. However, there are more direct temperature-sensing methods, that is, methods in which an electrical output is produced in response to a change in temperature.

An example is given in the next section, on physical sensors.

Most, but not all, sensors are transducers, employing one or more transduction mechanisms to produce a usable output signal. Sometimes sensors are classified as direct or indirect sensors according to how many transduction mechanisms are required to produce the desired output signal. Most commercial gas flow rate sensors are indirect sensors. Flow rate is usually inferred from other measurements, such as the displacement of an object placed in the flow stream (e.g., rotometers), the temperature of the gas measured downstream from a hot element (anemometer), or the difference in pressure measured at two points along the flow path. Figure 137.1 depicts a typical sensor block diagram identifying the measurand and associated input signal, the primary and intermediate transduction mechanisms, and the electronic output signal. Active sensors require an external power source in order to produce a usable output signal. A piezoresistor, for example, is a resistor that changes value when strained. But in order to sense the change in resistance, a current is passed through the resistor and the measured output voltage is related to resistance by Ohm's law:  $V = I \cdot R$ . Table 137.1 is a  $6 \times 6$  matrix of the more commonly employed physical and chemical transduction mechanisms. Detailed descriptions of most of these mechanisms can be found in college-level physics textbooks.

**Figure 137.1** Sensor block diagram. Active sensors require input power to accomplish transduction. Many sensors employ multiple transduction mechanisms in order to produce an electronic output in response to the measurand.



**Table 137.1** Physical and Chemical Transduction Properties

Primary Signal	Secondary Signal					
	Mechanical	Thermal	Electrical	Magnetic	Radiant	Chemical
Mechanical	(Fluid) mechanical and acoustic effects (e.g., diaphragm, gravity balance, echo sounder)	Friction effects (e.g., friction calorimeter), cooling effects (e.g., thermal flow meters)	Piezoelectricity; piezoresistivity; resistive, capacitive, and inductive effects	Magnetomechanical effects (e.g., piezomagnetic effect)	Photoelastic systems (stress-induced birefringence), interferometers, Sagnac effect, Doppler effect	
Thermal	Thermal expansion (bimetal strip, liquid-in-glass, and gas thermometers; resonant frequency), radiometer effect (light mill)		Seebeck effect, thermoresistance, pyroelectricity, thermal (Johnson) noise		Thermo-optical effects (e.g., in liquid crystals), radiant emission	Reaction activation (e.g., thermal dissociation)
Electrical	Electrokinetic and electromechanical effects (e.g., piezoelectricity, electrometer, Ampere's law)	Joule (resistive) heating, Peltier effect	Charge collectors, Langmuir probe	Biot-Savart's law	Electro-optical effects (e.g., Kerr effect, Pockels effect, electroluminescence)	Electrolysis, electromigration
Magnetic	Magnetomechanical effect (e.g., magnetorestriction, magnetometer)	Thermomagnetic effects (e.g., Righi-Leduc effect), galvanomagnetic effects (e.g., Ettingshausen effect)	Thermomagnetic effects (e.g., Ettingshausen-Nernst effect), alvano-magnetic effects (e.g., Hall effect, magnetoresistance)		Magneto-optical effects (e.g., Faraday effect, Cotton-Mouton effect)	
Radiant	Radiation pressure	Bolometer, thermopile	Photoelectric effects (e.g., photovoltaic effect, photoconductive effect)		Photorefractive effects, optical bistability	Photosynthesis, dissociation
Chemical	Hygrometer, electrodeposition cell, photoacoustic effect	Calorimeter, thermal conductivity cell	Potentiometry, conductimetry, amperometry, flame ionization, Volta effect, gas sensitive field effect	Nuclear magnetic resonance	(Emission and absorption) spectroscopy, chemiluminescence	

Source: Grandke, T. and Hesse, J. Volume 1: Fundamentals and general aspects. In *Sensors: A Comprehensive Survey*, ed. W. Gopel, J. Hesse and J. H. Zemel. VCH, Weinheim, Germany. With permission.

In choosing a particular sensor for a given application, many factors must be considered. These deciding factors or specifications can be divided into three major categories: environmental factors, economic factors, and sensor performance. The most commonly encountered factors are listed in [Table 137.2](#), although not all of these may be pertinent to a particular application. Most of the environmental factors determine the *packaging* of the sensor—which refers to the encapsulation or insulation that provides protection or isolation—and the input/output leads, connections, and cabling. The economic factors determine the type of manufacturing and materials used in the sensor and to some extent the quality of the materials. For example, a very expensive sensor may be cost-effective if it is used repeatedly, operates reliably for very long periods of time, or has exceptionally high performance. On the other hand, a disposable sensor, such as is desired in many medical applications, should be inexpensive. The performance requirements of the sensor are usually the specifications of primary concern. The most important parameters are **sensitivity**, **stability**, and **repeatability**. Normally, a sensor is only useful if all three of these parameters are tightly specified for a given range of measurand and time of operation. For example, a highly sensitive device is not useful if its output signal drifts greatly during the measurement time, and the data obtained are not reliable if the measurement is not repeatable. Other output signal characteristics, such as selectivity and linearity, can often be compensated for by using additional, independent sensors or with signal-conditioning circuits. In fact, most sensors have a response to temperature, since most transduction mechanisms are temperature-dependent.

**Table 137.2** Factors in Sensor Selection

Environmental Factors	Economic Factors	Sensor Performance
Temperature range	Cost	Sensitivity
Humidity effects	Availability	Range
Corrosion	Lifetime	Stability
Size	Performance	Repeatability
Overrange protection		Linearity
Susceptibility to EM interferences		Error
Ruggedness		Response time
Power consumption		Frequency response
Self-test capability		

Sensors are most often classified by the type of measurand—that is, physical, chemical, or biological. This approach provides a much simpler means of classification than by transduction mechanism or output signal (e.g., digital or analog), since many sensors use multiple transduction mechanisms and the output signal can always be processed, conditioned, or converted by a circuit so as to cloud the definition of output. A description of each class and examples are given in the following sections. In section 137.4, **microsensors** are introduced with some examples.

## 137.1 Physical Sensors

Physical measurands include temperature, strain, force, pressure, displacement, position, velocity, acceleration, optical radiation, sound, flow rate, viscosity, and electromagnetic fields. Referring to [Table 137.1](#), note that all but those transduction mechanisms listed in the chemical row are used in the design of physical sensors. Clearly, physical sensors compose a very large proportion of all sensors. It is impossible to illustrate all of them, but three measurands stand out in terms of their widespread application: temperature, displacement (or associated force), and optical radiation sensors.

### Temperature Sensors

Temperature is an important parameter in many control systems, most familiarly in environmental control systems. Several distinct transduction mechanisms have been employed. The mercury thermometer was mentioned earlier as a non-electronic sensor. The most commonly used electronic temperature sensors are thermocouples, thermistors, and resistance thermometers. Thermocouples operate according to the Seebeck effect, which occurs at the junction of two dissimilar metal wires, for example, copper and constantan. A voltage difference is generated at the hot junction due to the difference in the energy distribution of thermally energized electrons in each metal. A temperature gradient produces the requisite, albeit minute, current that enables the measurement of the junction voltage across the cool ends of the two wires. The voltage changes linearly with temperature over a given range, depending on the choice of metals. To minimize measurement error, the cool end of the couple must be kept at a constant temperature and the voltmeter must have a high input impedance. Connecting thermocouples in series creates a

thermopile that produces an output voltage approximately equal to a single thermocouple voltage multiplied by the number of couples.

The resistance thermometer relies on the increase in resistance of a metal wire with increasing temperature. As the electrons in the metal gain thermal energy, they move about more rapidly and undergo more frequent collisions with each other and the atomic nuclei. These scattering events reduce the mobility of the electrons, and since resistance is inversely proportional to mobility, the resistance increases. Resistance thermometers consist of a coil of fine metal wire. Platinum wire gives the largest linear range of operation. To determine the resistance indirectly, a constant current is supplied and the voltage is measured. A direct measurement can be made by placing the resistor in the sensing arm of a Wheatstone bridge and adjusting the opposing resistor to "balance" the bridge, which produces a null output. A measure of the sensitivity of a resistance thermometer is its temperature coefficient of resistance:  $TCR = (\Delta R/R)(1/\Delta T)$ , in units of percent resistance per degree of temperature.

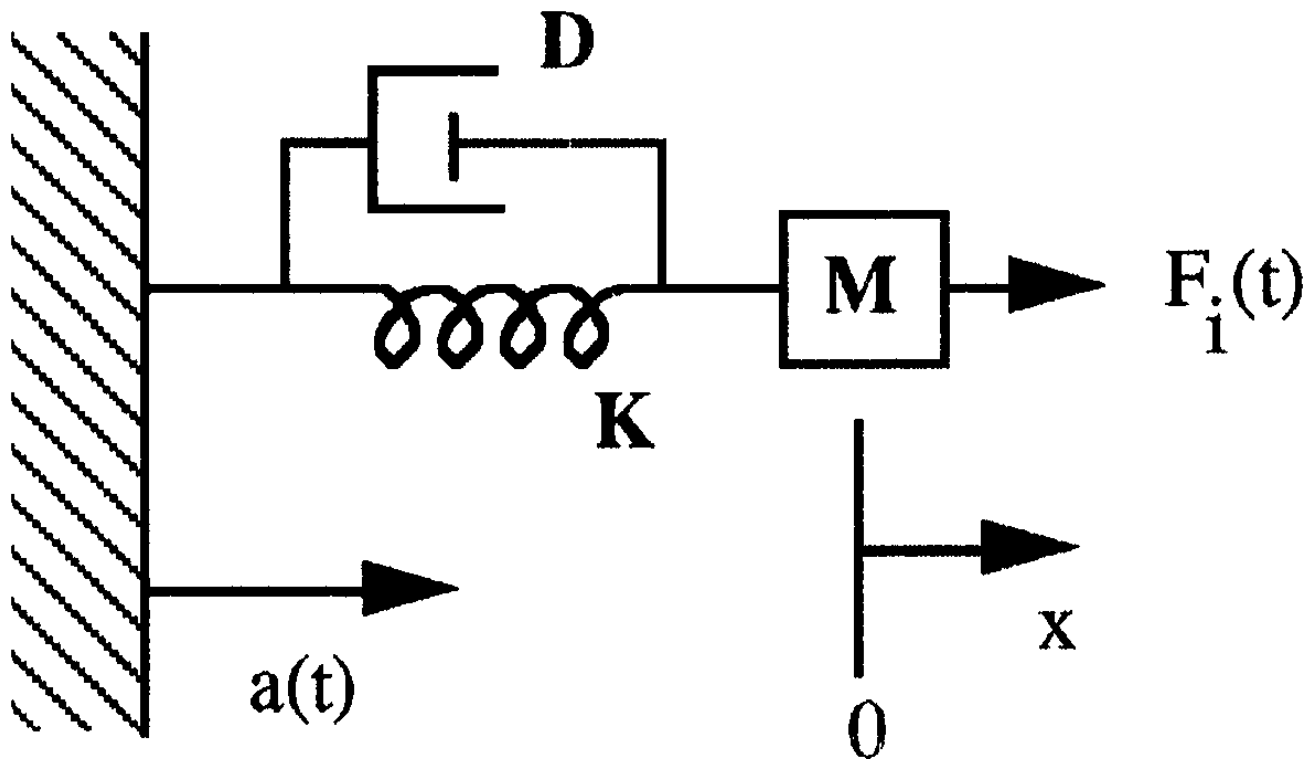
Thermistors are resistive elements made of semiconductive materials and have a negative coefficient of resistance. The mechanism governing the resistance change of a thermistor is the increase in the number of conducting electrons with an increase in temperature due to thermal generation; that is, the electrons that are tightly bound to the nucleus by coulombic attraction gain sufficient thermal energy to break away from the nucleus and become influenced by external fields. Thermistors can be measured in the same manner as resistance thermometers, but thermistors have up to 100 times higher TCR values.

## Displacement and Force Sensors

Many types of forces are sensed by the displacements they create. For example, the force due to acceleration of a mass at the end of a spring will cause the spring to stretch and the mass to move. This phenomenon is illustrated in [Fig. 137.2](#). The displacement,  $x$ , of the mass,  $M$ , from its zero acceleration position, is governed by the force generated by acceleration ( $F = M \cdot a$ ) and the restoring force of the spring ( $F = K \cdot x$ ). For a constant negative acceleration,  $-g$ , the steady state displacement,  $x$ , is equal to  $M \cdot g/K$ . Therefore, measurement of the displacement of a known mass can determine its acceleration.

**Figure 137.2** Schematic diagram of a mechanical oscillator made up of a mass at the end of a spring. The displacement of the mass,  $x(t)$ , in response to an acceleration,  $a(t)$ , is given by  $F_i = -M \cdot a(t) = M\ddot{x} + D\dot{x} + Kx$ , where  $K$  is the spring constant,  $D$  is viscous drag,  $M$  is the mass, and  $F_i$  is the force exerted on the mass. For a constant negative acceleration,  $-g$ , the steady state solution to this differential equation yields  $x = Mg/K$ . The displacement of a known mass, therefore, can be used to determine its acceleration.

**Figure 137.2**



Another example of force sensed by displacement is the displacement of the center of a deformable membrane due to a difference in pressure across it. Both the pressure sensor and the accelerometer examples use multiple transduction mechanisms to produce an electronic output. The primary mechanism converts force to displacement (mechanical to mechanical). An intermediate mechanism is used to convert displacement to an electrical signal (mechanical to electrical). Displacement can be determined by measuring the capacitance between two plates that move with respect to each other. The gap between the two plates is the displacement, and the associated capacitance is given by  $C = \text{area} \times \text{dielectric constant} / \text{gap length}$ . The ratio of plate area to gap length must be greater than 100, since most dielectric constants are on the order of  $1 \cdot 10^{-13}$  farads/cm and capacitance is readily resolvable to only about  $10^{-11}$  farads. This is because measurement leads and contacts create parasitic capacitances of about  $10^{-12}$  farads. If the capacitance is measured at the generated site by an integrated circuit (see the discussion of microsensors), capacitances as small as  $10^{-15}$  farads can be measured. Displacement is also commonly measured by the movement of a ferromagnetic core inside an inductor coil. The displacement produces a change in inductance that can be measured by placing the inductor in an oscillator circuit and measuring the change in frequency of oscillation.

The most commonly used force sensor is the strain gage. It consists of metal wires that are fixed to an immobile structure at one end and to a deformable element at the other. The resistance of the wire changes as it undergoes strain; that is, resistance changes in length, since the resistance of a wire is  $R = \text{resistivity} \times \text{length} / \text{cross-sectional area}$ . The wire's resistivity is a bulk property of the metal, which is a constant for constant temperature. For example, a strain gage can be used to measure acceleration by attaching both ends of the wire to a cantilever beam, with one end of the



wire at the attached beam end and the other at the free end. The cantilever-beam free end moves in response to an applied force, such as the force due to acceleration, which produces strain in the wire and a subsequent change in resistance. The sensitivity of a strain gage is described by the unitless gage factor,  $G = (\Delta R/R)/(\Delta L/L)$ . For metal wires, gage factors typically range from 2 to 3. Semiconductors are known to exhibit *piezoresistivity*, a change in resistance in response to strain that involves a large change in resistivity in addition to the change in linear dimension. Piezoresistors have gage factors as high as 130. Piezoresistive strain gages are frequently used in microsensors.

## Optical Radiation Sensors

The intensity and frequency of optical radiation are parameters of growing interest and utility in consumer products, such as the video camera and home security systems, and in optical communications systems. The conversion of optical energy to electronic signals can be accomplished by several mechanisms; however, the most commonly used is the photogeneration of electrons in semiconductors. The most often used device is the PN junction photodiode. The construction of this device is very similar to that of the diodes used in electronic circuits as rectifiers. The diode is operated in reverse bias, where very little current normally flows. When light is incident on the structure and is absorbed in the semiconductor, energetic electrons are produced. These electrons flow in response to the electric field sustained internally across the junction, producing an externally measurable current. The current magnitude is proportional to the light intensity and also depends on the frequency of the light.

Sensing infrared radiation (IR) by means of photogeneration requires very small band-gap semiconductors, such as HgCdTe. Since electrons in semiconductors are also generated by thermal energy (e.g., thermistors), these IR sensors require cryogenic cooling in order to achieve reasonable sensitivity. Another means of sensing IR is to first convert the optical energy to heat and then measure the temperature change. Accurate and highly sensitive measurements require passive sensors, such as a thermopile or a pyroelectric sensor, since active devices will generate heat themselves.

## 137.2 Chemical Sensors

---

Chemical measurands include ion concentration, chemical composition, rate of reactions, reduction-oxidation potentials, and gas concentration. The last row of [Table 137.1](#) lists some of the transduction mechanisms that have been, or could be, employed in chemical sensing. Two examples of chemical sensors are described here: the ion-selective electrode (ISE) and the gas chromatograph. These sensors were chosen because of their general use and availability and because they illustrate the use of a primary (ISE) versus a primary plus intermediate (gas chromatograph) transduction mechanism.

### Ion-Selective Electrode

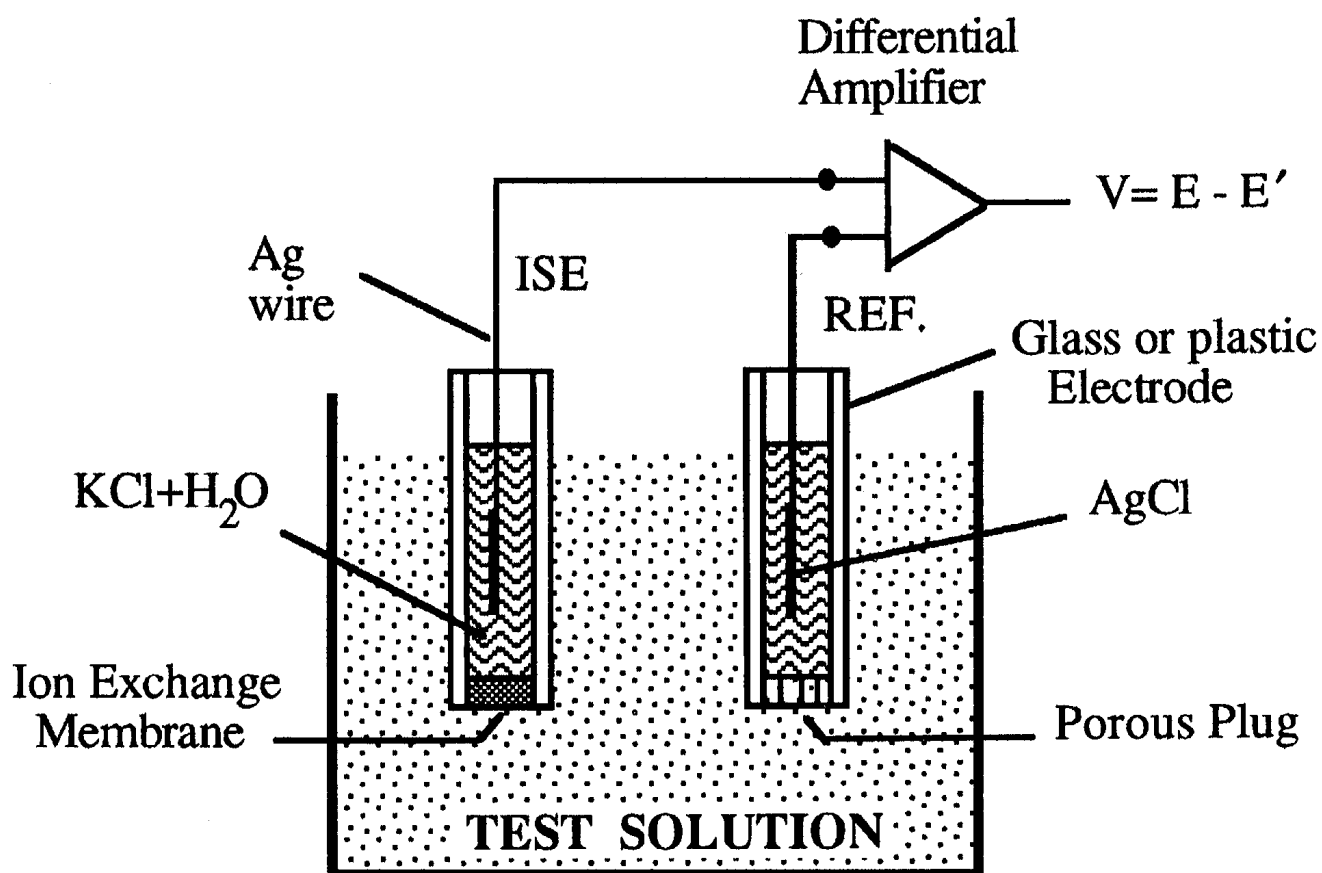
As the name implies, ISEs are used to measure the concentration of a specific ion concentration in

a solution of many ions. To accomplish this, a membrane that selectively generates a potential dependent on the concentration of the ion of interest is used. The generated potential is usually an equilibrium potential, called the *Nernst potential*, and develops across the interface of the membrane with the solution. This potential is generated by the initial net flow of ions (charge) across the membrane in response to a concentration gradient. From that moment on, the diffusional force is balanced by the generated electric force, and equilibrium is established. This potential is very similar to the built-in potential of a PN junction diode. The ion-selective membrane acts in such a way as to ensure that the generated potential is dependent mostly on the ion of interest and negligibly on any other ions in solution. This condition is brought about by enhancing the exchange rate of the ion of interest across the membrane, so it is the fastest moving and therefore the species that generates and maintains the potential.

The most familiar ISE is the pH electrode. In this device the membrane is a sodium glass that possesses a high exchange rate for  $H^+$ . The generated Nernst potential,  $E$ , is given by the expression:  $E = E_0 + (RT/F) \ln[H^+]$ , where  $E_0$  is a constant for constant temperature,  $R$  is the gas constant, and  $F$  is the Faraday constant. The pH is defined as the negative of the  $\log[H^+]$ ; therefore  $pH = (E_0 - E)(2.3)F/RT$ . One pH unit change corresponds to a tenfold change in the molar concentration of  $H^+$  and a 59 mV change in the Nernst potential at room temperature. Other ISEs have the same type of response, but specific to a different ion, depending on the choice of membrane. Some ISEs employ ionophores trapped inside a polymeric membrane. An ionophore is a molecule that selectively and reversibly binds with an ion and thereby creates a high exchange rate for that particular ion.

The typical ISE consists of a glass or plastic tube with the ion-selective membrane closing the end of the tube that is immersed into the test solution (see [Fig. 137.3](#)). The Nernst potential is measured by making electrical contact to either side of the membrane. This contact is made by placing a fixed-concentration, conductive filling solution inside the tube and placing a wire into the solution. The other side of the membrane is contacted by a reference electrode placed inside the same solution under test. The reference electrode is constructed in the same manner as the ISE, but it has a porous membrane or leaky plug that creates a liquid junction between its inner filling solution and the test solution. That junction is designed to have a potential that is invariant with changes in concentration of any ion in the test solution. The reference electrode, the solution under test, and the ISE form an electrochemical cell. The reference electrode potential acts like the ground reference in electric circuits, and the ISE potential is measured between the two wires emerging from the respective two electrodes. The details of the mechanisms of transduction in ISEs are beyond the scope of this article. The reader is referred to the book by Bard and Faulkner [1980] or the book by Koryta [1975].

**Figure 137.3** An electrochemical cell comprising an ion-selective electrode (ISE) and reference electrode in an ionic solution. The ion exchange membrane of the ISE generates a potential,  $E$ , that is proportional to the log of the concentration of the ion of interest.



## Gas Chromatograph

Molecules in gases have thermal conductivities that are dependent on their masses; therefore, a pure gas can be identified by its thermal conductivity. One way to determine the composition of a gas is to first separate it into its components and then measure the thermal conductivity of each. A gas chromatograph does exactly that. The gas flows through a long narrow column, which is packed with an adsorbent solid (for gas-solid chromatography), wherein the gases are separated according to the retentive properties of the packing material for each gas. As the individual gases exit the end of the tube one at a time, they flow over a heated wire. The amount of heat transferred to the gas depends on its thermal conductivity. The gas temperature is measured a short distance downstream and compared to a known gas flowing in a separate sensing tube. The temperature is related to the amount of heat transferred and can be used to derive the thermal conductivity according to thermodynamic theory and empirical data. This sensor required two transductions: a chemical-to-thermal energy transduction, followed by a thermal-to-electrical transduction.

## 137.3 Biosensors

---

**Biosensors** respond to biological measurands, which are biologically produced substances, such as antibodies, glucose, hormones, and enzymes. Biosensors are not the same as biomedical sensors, which are any sensors used in biomedical applications, such as blood pressure sensors or electrocardiogram electrodes. Many biosensors are biomedical sensors; however, the former are also used in industrial applications, for example, the monitoring and control of fermentation reactions. [Table 137.1](#) does not include biological signals as primary signals because they can be classified as either chemical or physical in nature. Biosensors are of special interest because of the very high selectivity of biological reactions and binding. However, detection of that reaction or binding is often elusive. A very familiar commercial biosensor is the in-home pregnancy test sensor, which detects the presence of human growth factor in urine. That device is a nonelectronic sensor since the output is a color change that your eye senses. In fact, most biosensors require multiple transduction mechanisms to arrive at an electrical output signal. Two examples are given below: an immunosensor and an enzyme sensor.

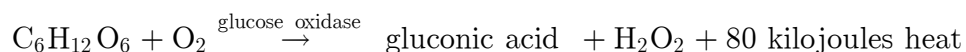
### Immunosensor

Most commercial techniques for detecting antibody-antigen binding utilize optical or x-radiation detection. An optically fluorescent molecule or radioisotope is attached or tagged to the species of interest in solution. The complementary binding species is chemically attached to a glass substrate or glass beads, which are packed into a column. The solution containing the tagged species of interest, say the antibody, is passed over the antigen-coated surface, where the two selectively bind. The nonbound fluorescent molecules or radioisotopes are washed away, and the antibody concentration is determined by fluorescence spectroscopy or with a scintillation counter, respectively. These sensing techniques are quite costly and bulky, and therefore other biosensing mechanisms are rapidly being developed and commercialized. One technique uses the change in the permittivity or index of refraction of the bound antibody-antigen complex in comparison to an unbound surface layer of antigen. The technique utilizes surface plasmon resonance as the mechanism for detecting the change in permittivity. Surface plasmons are charge density oscillations that propagate along the interface between a dielectric and a thin metal film. They can be generated by the transfer of energy from an incident, transverse magnetic (TM) polarized light beam. Plasmon resonance, or maximum energy transfer, occurs when the light beam is incident at a specific angle that depends on, and is extremely sensitive to, the dielectric permittivity at the interface on either side of a given metal film. At resonance, the optical beam intensity reflected from the metal film drops to a minimum; therefore, the resonance angle can be determined by recording reflected intensity (e.g., with a photodiode), versus incident light beam angle. The binding of antibody to an antigen layer attached to the metal film will produce a change in the resonance angle.

### Enzyme Sensor

Enzymes selectively react with a chemical substance to modify it, usually as the first step in a chain of reactions to release energy (metabolism). A well-known example is the selective reaction of glucose oxidase (enzyme) with glucose to produce gluconic acid and peroxide, according to the

following formula:



An enzymatic reaction can be sensed by measuring the rise in temperature associated with the heat of reaction or by the detection and measurement of by-products. In the glucose example, the reaction can be sensed by measuring the local dissolved peroxide concentration. This is done via an electrochemical analysis technique called *amperometry* [Bard and Faulkner, 1980]. In this method a potential is placed across two inert metal wire electrodes immersed in the test solution, and the current that is generated by the reduction/oxidation reaction of the species of interest is measured. The current is proportional to the concentration of the reducing/oxidizing species. A selective response is obtained if no other available species has a lower redox potential. Because the selectivity of peroxide over oxygen is poor, some glucose-sensing schemes employ a second enzyme called *catalase*, which converts peroxide to oxygen and hydroxyl ions. The latter produces a change in the local pH. As described earlier, an ISE can then be used to convert the pH to a measurable voltage. In this latter example, glucose sensing involves two chemical-to-chemical transductions followed by a chemical-to-electrical transduction mechanism.

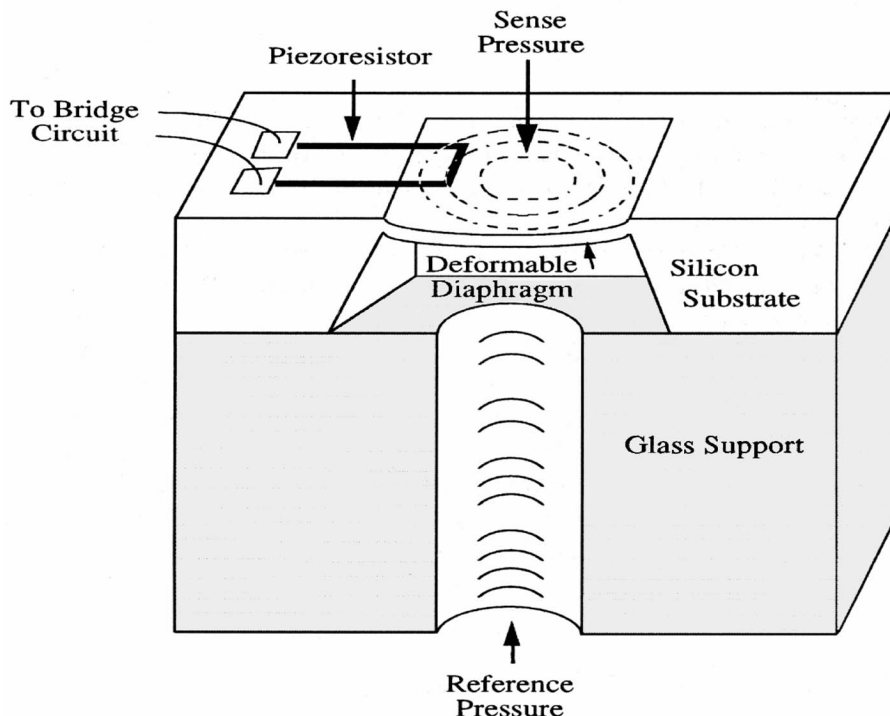
## 137.4 Microsensors

---

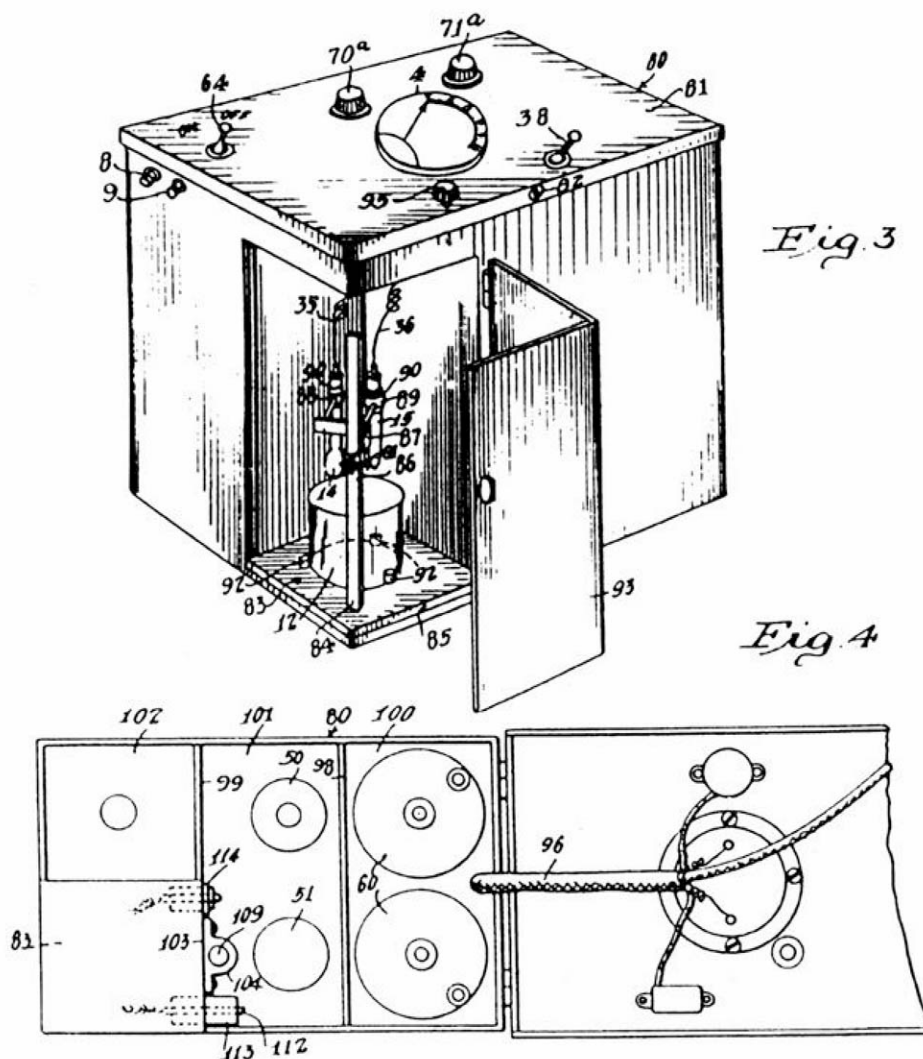
Microsensors are sensors that are manufactured using integrated circuit fabrication technologies and/or micromachining. Integrated circuits are fabricated using a series of process steps that are done in batch fashion, meaning that thousands of circuits are processed together at the same time in the same way. The patterns that define the components of the circuit are photolithographically transferred from a template to a semiconducting substrate using a photosensitive organic coating. The coating pattern is then transferred into the substrate or into a solid-state thin film coating through an etching or deposition process. Each template, called a *mask*, can contain thousands of identical sets of patterns, with each set representing a circuit. This "batch" method of manufacturing is what makes integrated circuits so reproducible and inexpensive. In addition, photoreduction enables one to make extremely small features, on the order of microns, which is why this collection of process steps is referred to as *microfabrication*. The resulting integrated circuit is contained in only the top few microns of the semiconductor substrate and the submicron thin films on its surface. Hence, integrated circuit technology is said to consist of a set of planar microfabrication processes. **Micromachining** refers to the set of processes that produces three-dimensional microstructures using the same photolithographic techniques and batch processing as for integrated circuits. Here, the third dimension refers to the height above the substrate of the deposited layer or the depth into the substrate of an etched structure. Micromachining produces third dimensions in the range of 1–500 microns (typically). The use of microfabrication to manufacture sensors produces the same benefits as it does for circuits: low cost per sensor, small size, and highly reproducible behavior. It also enables the integration of signal conditioning, compensation circuits, and actuators—that is, entire sensing and control systems—which can dramatically improve sensor performance for very little increase in cost. For these reasons there is a great deal of research and development activity in microsensors. The first

microsensors were integrated circuit components, such as semiconductor resistors and PN junction diodes. The piezoresistivity of semiconductors and optical sensing by the photodiode were discussed earlier. Diodes are also used as temperature-sensing devices. When forward biased with a constant diode current, the resulting diode voltage increases approximately linearly with increasing temperature. The first micromachined microsensor to be commercially produced was the silicon pressure sensor. It was invented in the mid-to-late 1950s at Bell Labs and commercialized in the 1960s by General Electric, Endevco, and Fairchild Control Division (now Foxboro/ICT, Inc.). This device contains a thin silicon diaphragm ( $\approx 10$  microns) that deforms in response to a pressure difference across it (Fig. 137.4). The deformation produces two effects: a position-dependent displacement, which is maximum at the diaphragm center, and position-dependent strain, which is maximum near the diaphragm edge. Both of these effects have been used in microsensors to produce an electrical output that is proportional to differential pressure. The membrane displacement is sensed capacitively, as previously described, in one type of pressure sensor. In another, the strain is sensed by placing a piezoresistor, fabricated in the same silicon substrate, along one edge of the diaphragm. The two leads of the piezoresistor are connected to a Wheatstone bridge. The latter type of sensor is called a *piezoresistive pressure sensor* and is the more common type of commercial pressure microsensor. Pressure microsensors constituted about 5% of the total U.S. consumption of pressure sensors in 1991. Most of them are used in the medical and automotive industry because of their low cost and small, rugged construction. Many other types of microsensors are available commercially, including accelerometers, flow rate sensors, gas sensors, and biosensors.

**Figure 137.4** Schematic cross section of a silicon piezoresistive pressure sensor. A differential pressure deforms the silicon diaphragm, producing strain in the integrated piezoresistor. The change in resistance is measured via a Wheatstone bridge.







# APPARATUS FOR TESTING ACIDITY

Arnold O. Beckman and Henry E. Fracker

Patented October 27, 1936

#2,058,761

An excerpt:

In accordance with the present invention, we substitute for the delicate and fragile galvanometer, previously considered necessary, a simple and mechanically rugged milliammeter in combination with a specially designed vacuum tube amplifier employing standard radio tubes and energized from dry cells. The entire apparatus can be housed in a compact portable case, it is extremely accurate and is substantially foolproof in operation so that it can be manipulated by unskilled operators without danger of it being ruined or being thrown out of adjustment.

This was the modern pH meter. Using vacuum-tube DC amplifiers (a relatively new field) they were able to measure potentials in a circuit of very high resistance, making determination of hydrogen ion concentration (or pH) much easier and reliable. Beckman Instruments, Inc., started because of this invention, today still produces pH meters, probes and other modern instrumentation products. (© 1995, DewRay Products, Inc. Used with permission.)

## Defining Terms

**Biosensor:** A sensor that responds to biologically produced substances, such as enzymes, antibodies, and hormones.

**Micromachining:** The set of processes that produce three-dimensional microstructures using sequential photolithographic pattern transfer and etching or deposition in a batch-processing method.

**Microsensor:** A sensor that is fabricated using integrated circuit and micromachining technologies.

**Repeatability:** The ability of a sensor to reproduce output readings for the same value of measurand, when applied consecutively and under the same conditions.

**Sensitivity:** The ratio of the change in sensor output to a change in the value of the measurand.

**Sensor:** A device that produces a usable output in response to a specified measurand.

**Stability:** The ability of a sensor to retain its characteristics over a relatively long period of time.

**Transducer:** A device that converts one form of energy to another.

## References

- ANSI. 1975. *Electrical Transducer Nomenclature and Terminology*. ANSI Standard MC6.1-1975. (ISA S37.1). Instrument Society of America, Research Triangle Park, NC.
- Bard, A. J. and Faulkner, L. R. 1980. *Electrochemical Methods: Fundamentals and Applications*. John Wiley & Sons, New York.
- Carstens, J. R. 1993. *Electrical Sensors and Transducers*. Regents/Prentice Hall, Englewood Cliffs, NJ.
- Cobbold, R. S. C. 1974. *Transducers for Biomedical Measurements: Principles and Applications*. John Wiley & Sons, New York.
- Fraden, J. 1993. *AIP Handbook of Modern Sensors: Physics, Designs, and Applications*. American Institute of Physics, New York.
- Grandke, T. and Ko, W. H. 1989. Volume 1: Fundamentals and general aspects. In *Sensors: A Comprehensive Survey*, ed. W. Gopel, J. Hesse, and J. N. Zemel. VCH, Weinheim, Germany.
- Janata, J. 1989. *Principles of Chemical Sensors*. Plenum Press, New York.
- Koryta, J. 1975. *Ion-Selective Electrodes*. Cambridge University Press, Cambridge, UK.
- Norton, H. N. 1989. *Handbook of Transducers*. Prentice Hall, Englewood Cliffs, NJ.



## Further Information

*Sensors: A Comprehensive Survey*. Gopel, W., Hesse, J., and Zemel, J. N. (Eds.) VCH, Weinheim, Germany.

Volume 1: Fundamentals and General Aspects, 1989

Volume 2, 3: Mechanical Sensors, 1991

Volume 4: Thermal Sensors, 1990

Volume 5: Magnetic Sensors, 1989

Volume 6: Optical Sensors, 1991

Volume 7, 8: Chemical and Biochemical Sensors, 1990

*Sensors and Actuators* is a technical journal devoted to solid-state sensors and actuators. It is published bimonthly by Elsevier Press in two volumes: **Volume A: Physical Sensors** and **Volume B: Chemical Sensors**.

The International Conference on Solid-State Sensors and Actuators is held every two years, hosted in rotation by the U.S., Japan, and Europe. It is sponsored in part by the Institute of Electrical and Electronic Engineers (IEEE). The *Digest of Technical Papers* is published by and available through IEEE, Piscataway, NJ.

Harrison, S. J. "Measurement Errors and Accuracy"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Measurement Errors and Accuracy

---

[138.1 Measurement Errors, Accuracy, and Precision](#)

[138.2 Estimating Measurement Uncertainty](#)

[138.3 Propagation of Measurement Uncertainty](#)

[138.4 Uncertainty Analysis in Experimental Design](#)

**Steve J. Harrison**

*Queen's University*

Engineers are increasingly being asked to monitor or evaluate the efficiency of a process or the performance of a device. Results are often derived from the combination of values determined from a number of individual measurements. In the planning of an experiment, an attempt is usually made to minimize the error associated with the experimental measurements. Unfortunately, every measurement is subject to error, and the degree to which this error is minimized is a compromise between the (overall) accuracy desired and the expense required to reduce the error in the component measurements to an acceptable value.

Good engineering practice dictates that an indication of the error or uncertainty should be reported along with the derived results. In the worst case, the error in a derived result could be treated as the sum of the errors of the component measurements. Implicit in this assumption is that the worst-case errors will occur simultaneously and in the most detrimental fashion. This condition is unlikely, and therefore such an analysis usually results in an overprediction of the error in a derived result.

A more realistic estimate of error was presented by Kline and McClintock [1953] based on single-sample uncertainty analysis. This analysis will be described later in this chapter, but first let us undertake to describe measurement errors. The following discussion is meant to provide an insight into measurement uncertainty rather than a rigorous treatment of the theoretical basis. Readers are encouraged to consult the many references [e.g., ANSI/ASME, 1985] dealing with these issues if they consider it necessary.

### 138.1 Measurement Errors, Accuracy, and Precision

---

**Measurement error** may be defined as the difference between the true value and the measured value of the quantity:

$$e \equiv x - x_{\text{true}}$$

where

$e$  is the error in  $x$

$x$  is a measured or observed value of some physical quantity

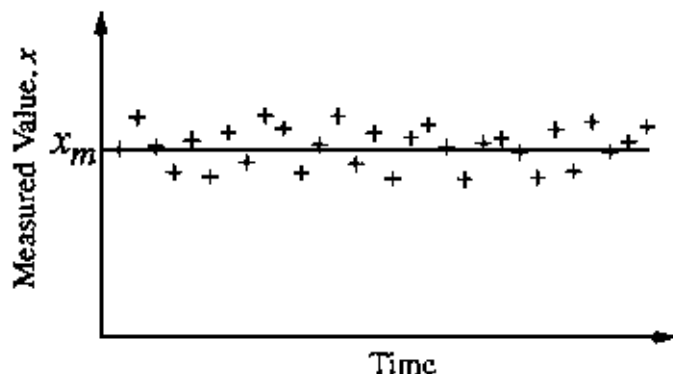
$x_{\text{true}}$  is the actual value

The errors that occur in an experiment are usually categorized as mistakes or recording errors, systematic or fixed errors, and accidental or random errors.

- Mistakes or recording errors are usually the result of measurement errors (e.g., the observer reads 10.1 instead of 11.1 units on the scale of a meter). It is assumed that careful experimental practices will minimize the occurrence of this type of error.
- **Systematic errors** or **bias errors** are errors that persist and cannot be considered due entirely to chance. Systematic errors may result from incorrect instrument calibrations and relate to instrument accuracy (the ability of the instrument to indicate the true value).
- **Random errors** cause readings to take random values on either side of some mean value. They may be due to the observer or the instrument and are revealed by repeated observations. They are disordered in incidence and variable in magnitude.

In measurement systems, **accuracy** generally refers to the closeness of agreement of a measured value and the true value. All measurements are subject to both systematic (bias) and random errors to differing degrees, and consequently the true value can only be estimated. To illustrate the above concepts, consider the case shown in Fig. 138.1, where measurements of a fixed value are taken over a period of time.

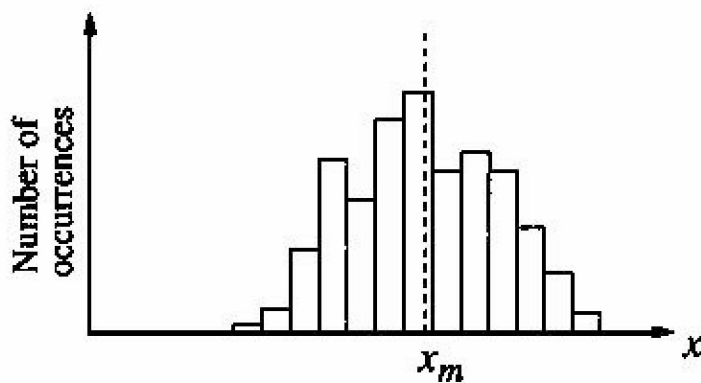
**Figure 138.1** Repeated measurements of a fixed value.



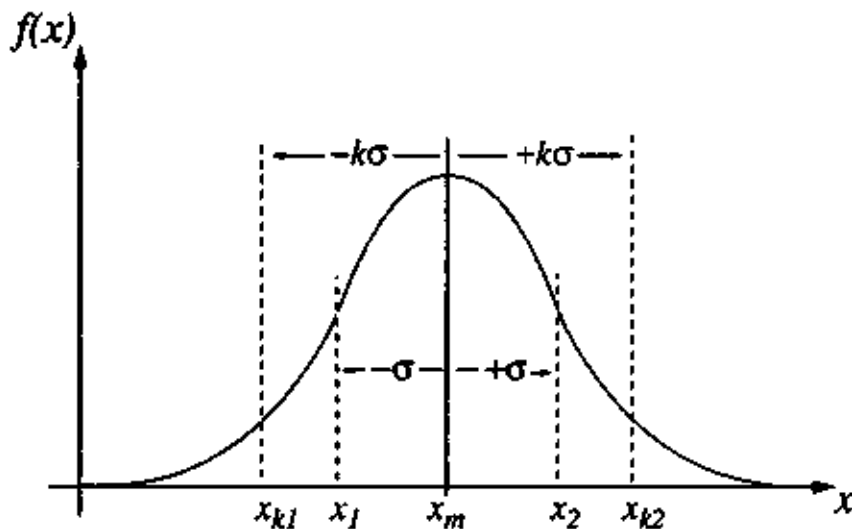
As illustrated, the measured values are scattered around some mean value,  $x_m$ . In effect, if many measurements of the quantity were taken, we could calculate a mean value of the fixed quantity. If we further grouped the data into ranges of values, it would be possible to plot the frequency of occurrence in each range as a histogram (Fig. 138.2) If a large number,  $n$ , of measurements were recorded (i.e., as  $n$  goes to infinity), a plot of the **frequency distribution** of the measurements,

$f(x)$ , could be constructed (Fig. 138.3). Figure 138.3 is often referred to as a plot of the probability density function, and the area under the curve represents the probability that a particular value of  $x$  (the measured quantity) will occur. The total area under the curve has a value of 1, and the probability that a particular measurement will fall within a specified range (e.g., between  $x_1$  and  $x_2$ ) is determined by the area under the curve bounded by these values. Figure 138.3 indicates that there is a likelihood of individual measurements being close to  $x_m$ , and that the likelihood of obtaining a particular value decreases for values farther away from the mean value,  $x_m$ .

**Figure 138.2** Histogram of measured values.



**Figure 138.3** Frequency distribution of repeated measurements.



The frequency distribution shown in Fig. 138.3 corresponds to a **Gaussian or normal distribution curve** [Bolz and Tuve,1987], the form generally assumed to represent random measurement uncertainty. There is no guarantee that this symmetrical distribution, indicating an

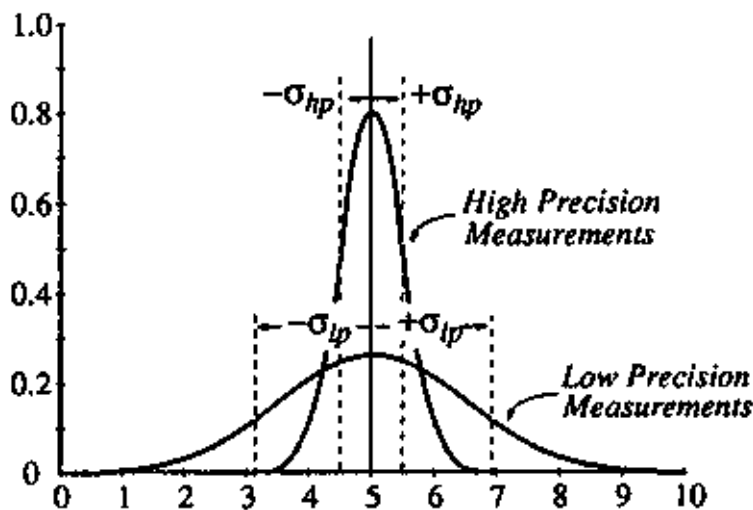
equal probability of measurements falling above or below the mean value,  $x_m$ , will occur, but experience has shown that the normal distribution is generally suitable for most measurement applications. Other distributions could be determined based on taking large numbers of measurements and plotting the results [Bevington, 1969], but generally this information is unavailable.

In analyzing these results we may apply standard statistical tools to express our confidence in the determined value based on the probability of obtaining a particular result. If experimental errors follow a normal distribution, then a widely reported value is the **standard deviation**,  $\sigma$  [Bevington, 1969]. There is a 68% (68.27%) probability that an observed value  $x$  will fall within  $\pm\sigma$  of  $x_m$  (Fig. 138.3). Other values are often reported for  $\pm 2\sigma$  (probability 95.45%) or  $\pm 3\sigma$  (probability 99.73%). Values of probabilities are tabulated [Bolz and Tuve, 1987] for ranges from  $x_{k1}$  to  $x_{k2}$ , or  $x_m \pm k\sigma$ , as illustrated in Fig. 138.3.

In reporting measurements, an indication of the probable error in the result is often stated based on an absolute error prediction [e.g., a temperature of  $48.3 \pm 0.1^\circ \text{C}$  (based on a 95% probability)] or on a relative error basis [e.g., voltage of  $9.0 \text{ V} \pm 2\%$  (based on a 95% probability)]. The choice of probability value corresponding to the error limits is arbitrary, but a value of 95%, corresponding to  $\pm 2\sigma$ , appears to be widely used.

When considering the results in Fig. 138.3, lacking other information, our best estimate of the true value of the measured quantity is the mean (or average) of the measured values. Based on the previous discussions we may characterize measurements and instrumentation as being of high or low **precision**. This is illustrated in Fig. 138.4, where two probability distributions are plotted. As may be seen, the probability of obtaining values near the mean is greater for the high-precision measurement (i.e., the measurements are highly repeatable or precise). The low-precision measurements have a wider distribution and are characterized by a greater standard deviation,  $\pm\sigma_{lp}$ , compared with the high-precision measurements,  $\pm\sigma_{hp}$ .

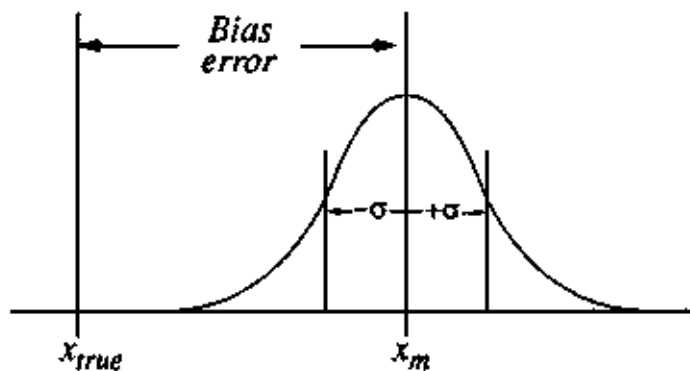
**Figure 138.4** Frequency distributions for instruments of varying precision.



It is worth noting that while the low-precision measurement is not highly repeatable, if the mean value of a large number of measurements is calculated, then the value  $x_m$  should be the same as that determined from the precise measurement. Therefore, in the absence of bias or systematic error, the mean of a large sample of low-precision measurements theoretically indicates the true value.

The previous discussion, illustrating the concept of random measurement error, has not addressed the effects of systematic or fixed (bias) errors. In reality, even though a high probability exists that an individual measurement will be close to the mean value, there is no guarantee that the value of the mean of the large sample of measurements will be the true value (Fig. 138.5). In effect, a particular measurement and measurement instrument may be highly repeatable (or precise) but in error (i.e., not accurate). Systematic or bias errors may arise from either the observer or the instrument. They may be constant or vary in some regular way. An example of a source of systematic error is a pressure gauge needle that is bent (i.e., not zeroed).

**Figure 138.5** Measurements with bias errors in addition to random errors.



Bias errors may be significant but, by their nature, should be identifiable and their effects accounted for through careful calibration and comparison with other instruments. Recently, methods for combining bias and random error estimates into an overall error estimate have been presented [ANSI/ASME, 1985].

## 138.2 Estimating Measurement Uncertainty

In the previous discussion we considered multi-sample experiments of a single parameter.

**Multi-sample experiments** are those in which, for a given set of the independent experimental variables, the readings are taken many times. If we could repeat our tests many times, with many observers and a variety of instruments, we could apply statistics to determine the reliability of the results as in the previously discussed methods.

In engineering we are often forced to conduct "single-sample" experiments, and therefore we cannot estimate uncertainty by observing repeated tries. Single-sample experiments are those in which, for a given set of experimental conditions, the readings are taken only once. These are typical in engineering, where financial or time constraints limit the number of repetitions of a

particular test. In these cases the **uncertainty** in the measurements can usually only be estimated.

In estimating the uncertainty in a measurement, the experimenter must rely on judgment based on experience. Only an estimate of what would happen (if a large number of observations were made) can be made. In this case we refer to an "uncertainty distribution" rather than a "frequency distribution." The uncertainty distribution is the distribution of error which the experimenter believes would be found in a given variable if the variable were sampled a great many times. The experimenter also expresses the degree of confidence in the stated uncertainty based on "odds," in a manner analogous to the standard deviation. Therefore, odds of approximately 2 to 1 (e.g., the odds are 2 to 1 that the error on a particular reading will be within a specified interval) roughly corresponds to an interval of  $\pm\sigma$  (i.e., a probability of 68%). Therefore, based on the experience of the experimenter, an estimate of the uncertainty associated with a particular measurement can be made [i.e.,  $m \pm \omega$  ( $b$  to 1)], where  $m$  is the mean value,  $\omega$  is the uncertainty interval, and  $b$  is the odds]. The experimenter is wagering  $b$  to 1 that the error is less than  $\omega$  [e.g., voltage =  $12 \pm 1$  V (10 to 1 odds)].

### 138.3 Propagation of Measurement Uncertainty

---

As previously stated, results are often derived from the combination of values determined from a number of individual measurements. The propagation of uncertainty is defined as the way in which uncertainties in the variables affect the uncertainty in the results. Kline and McClintock [1953] have presented a technique for determining the propagation of errors in a derived result for single-sample experiments. In this technique a best estimate of the uncertainty in each variable is made by the researcher assuming odds (e.g., 20 to 1) that a measured value will fall within the uncertainty interval. Therefore, assuming that a desired result is derived from  $n$  independent variables or simultaneous measurements [e.g.,  $R = R(X_1, X_2, X_3, \dots, X_n)$ ] and letting  $\omega_1, \omega_2, \dots, \omega_n$  be the uncertainties in the independent variables, we can derive an expression for the uncertainty in the result,  $\omega_R$ .

If the same uncertainty estimate is used for all the component variables in the analysis (i.e., all the uncertainties must be given based on the same odds), the uncertainty in the result will be given by

$$\omega_R = \pm \left[ \left( \frac{\partial R}{\partial X_1} \omega_1 \right)^2 + \left( \frac{\partial R}{\partial X_2} \omega_2 \right)^2 + \left( \frac{\partial R}{\partial X_3} \omega_3 \right)^2 + \dots + \left( \frac{\partial R}{\partial X_n} \omega_n \right)^2 \right]^{1/2}$$

In certain instances (such as a product form of the equation, i.e.,  $R = X_1^a \cdot X_2^b \cdot X_3^c \dots X_n^m$ ), this equation can be reduced to

$$\frac{\omega_R}{R} = \pm \left[ \left( a \cdot \frac{\omega_1}{X_1} \right)^2 + \left( b \cdot \frac{\omega_2}{X_2} \right)^2 + \left( c \cdot \frac{\omega_3}{X_3} \right)^2 + \dots + \left( m \cdot \frac{\omega_n}{X_n} \right)^2 \right]^{1/2}$$



where  $\omega_R/R$  is the relative uncertainty in the result,  $R$  [Moffat, 1988].

## 138.4 Uncertainty Analysis in Experimental Design

The possible experimental errors should be examined before conducting an experiment to ensure the best results. The experimenter should estimate the uncertainties in each measurement. Uncertainty propagation in the result depends on the squares of the uncertainties in the independent variables; thus the squares of the large uncertainty values dominate the result. Thus, in designing an experiment, very little is gained by reducing the small errors (uncertainties). A series of measurements with relatively large uncertainties could produce a result with an uncertainty not much larger than that of the most uncertain measurement. Thus to improve the overall experimental result, the large uncertainties must be reduced [Moffat, 1985, 1988].

**Example.** Consider an experiment to measure the thermal efficiency,  $\eta$ , of a solar collector for heating water. During testing, the solar collector is connected in series with an electric reference heater, and cooling water is circulated through both. Values of  $\eta$  are obtained by individual measurements of: the temperature rise across the solar collector,  $\Delta T_c$ , and reference heater,  $\Delta T_h$ ; the power input to the reference heater,  $P_h$ ; and the total solar energy incident on the collector surface,  $G_i$ . The thermal efficiency,  $\eta$ , is determined according to

$$\eta = \frac{P_h}{G_i} \cdot \left( \frac{\Delta T_c}{\Delta T_h} \right)$$

To estimate the relative uncertainty in the measured value of  $\eta$ , the uncertainties in the component measurements must be estimated. The electrical power input to the reference heater was measured with a power transducer with an estimated uncertainty of  $\pm 2\%$  of reading. The measurement of solar radiation introduced the largest uncertainty and was estimated at  $\pm 10\%$  of reading. Errors in the measurement of temperature rise across the heater and collector were estimated at  $\pm 0.1^\circ \text{C}$ .

Therefore, the uncertainty values are

$$\omega P_h = \pm 0.02 \times P_h$$

$$\omega G_i = \pm 0.10 \times G_i$$

$$\omega \Delta T = \pm 0.1^\circ \text{C}$$

These values are based on best estimates and represent the odds of 20 to 1 that measured values will fall within these limits.

Applying the analysis described above, the relative uncertainty in determining  $\eta$  is

$$\frac{\omega_\eta}{\eta} = \pm \left[ \left( \frac{\omega P_H}{P_h} \right)^2 + \left( \frac{\omega G_i}{G_i} \right)^2 + \left( \frac{\omega \Delta T_h}{\Delta T_h} \right)^2 + \left( \frac{\omega \Delta T_c}{\Delta T_c} \right)^2 \right]^{1/2}$$

Substituting typical test conditions, e.g.,  $\Delta T_c \approx 3.5^\circ \text{C}$  and  $\Delta T_h \approx 6^\circ \text{C}$ , then

$$\frac{\omega_{\eta}}{\eta} = \pm[(0.02)^2 + (0.10)^2 + (0.1/6)^2 + (0.1/3.5)^2]^{0.5} = \pm 0.10$$

Thus, the relative uncertainty in the measured value of solar collector efficiency would be  $\pm 0.10$ , or  $\pm 10\%$  (based on 20-to-1 odds).

It is apparent from this analysis that the uncertainty in the solar radiation measurement,  $\omega G_i$ , dominates the uncertainty in the result; that is, decreasing the uncertainty in the other measurements has little effect on the result. However, if the value of  $\omega G_i$  is reduced to  $\pm 4\%$ , then  $\omega_{\eta}/\eta$  would be reduced to  $\pm 5\%$ . Further analysis shows that additional reductions in  $\omega G_i$  are less significant, as the uncertainties in the other measurements dominate. It is also worth noting that as  $\Delta T_c$  becomes smaller (corresponding to a reduction in  $\eta$ ), the overall uncertainty in the result increases. This effect could be minimized by reducing the uncertainty in the  $\Delta T_c$  measurement.

## Defining Terms

**Accuracy:** The closeness of agreement of a measured value and the true value.

**Bias error:** The tendency of an estimate to deviate in one direction from a true value.

**Frequency distribution:** A description of the frequency of occurrence of the values of a variable.

**Gaussian or normal distribution curve:** The theoretical distribution function for the conceptual infinite population of measured values. This distribution is characterized by a limiting mean and standard deviation. The interval of the limiting mean plus or minus two times the standard deviation will include approximately 95% of the total scatter of measurements.

**Measurement error:** The difference between true and observed values.

**Multi-sample experiments:** Experiments in which uncertainties are evaluated by many repetitions and many diverse instruments. Such experiments can be analyzed by classical statistical means.

**Precision:** The repeatability of measurements of the same quantity under the same conditions.

**Random errors:** Errors that cause readings to take random values on either side of some mean value. They may be due to the observer or the instrument and are revealed by repeated observations.

**Standard deviation:** A measure of dispersion of a population. It is calculated as the square root of the average of the squares of the deviations from the mean (root mean square) deviation.

**Systematic errors:** Errors which persist and cannot be considered as due to chance. Systematic errors may be due to instrument manufacture or incorrect calibration and relate to instrument accuracy (the ability of the instrument to indicate the true value).

**Uncertainty:** The estimated error limit of a measurement or result for given odds.

## References

- ANSI/ASME. 1985. *Instruments and Apparatus. Part I: Measurement Uncertainty*. PTC 19.1. American Society of Mechanical Engineers, New York. [Note: This standard is dated originally in 1985 but was reaffirmed in 1990.]
- Beckwith, T. G., Marangoni, R. D., and Lienhard, V. J. H. 1993. *Mechanical Measurements*, 5th ed. Addison-Wesley, Reading, MA.
- Bevington, R. P. 1969. *Data Reduction and Error analysis for the Physical Sciences*. McGraw-Hill, New York.
- Bolz, R. E. and Tuve, G. L. 1987. *CRC Handbook of Tables for Applied Engineering Science*, 2nd ed. CRC Press, Inc., Boca Raton, FL.
- Kline, S. J. and McClintock, F. A. 1953. Describing uncertainties in single-sample experiments. *Mech. Eng.* 75:3–8.
- Moffat, R. J. 1985. Using uncertainty analysis in the planning of an experiment. *J. Fluids Eng. Trans. ASME*. 107:173–178.
- Moffat, R. J. 1988. Describing the uncertainties in experimental results. *Exp. Therm. Fluid Sci.* 1:3–17.

## Further Information

- Extensive discussion of measurement instrumentation and an introduction to measurement errors and *et al.* [1993] and J. P. Holman, *Experimental Methods for Engineers*, 6th ed., 1994, McGraw-Hill, New York.
- Detailed discussions of the treatment of systematic and random errors for measurement systems are given in *ASHRAE Handbook of Fundamentals*, Chapter 13, available from ASHRAE, New York.
- The use of uncertainty analysis in the planning of experiments is outlined with examples by Moffat [1985].

Dyer, S. A. "Signal Conditioning"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

### 139.1 Linear Operations

Amplitude Scaling • Impedance Transformation • Linear Filtering

### 139.2 Nonlinear Operations

**Stephen A. Dyer**

*Kansas State University*

Kelvin's first rule of instrumentation states, in essence, that the measuring instrument must not alter the event being measured. For the present purposes, we can consider the instrument to consist of an input transducer followed by a signal-conditioning section, which in turn drives the data-processing and display section (the remainder of the instrument). We are using the term *instrument* in the broad sense, with the understanding that it may actually be a measurement subsystem within virtually any type of system.

Certain requirements are imposed upon the transducer if it is to reproduce an event faithfully: It must exhibit amplitude linearity, phase linearity, and adequate frequency response. But it is the task of the signal conditioner to accept the output signal from the transducer and from it produce a signal in the form appropriate for introduction to the remainder of the instrument.

Analog signal conditioning can involve strictly *linear* operations, strictly *nonlinear* operations, or some combination of the two. In addition, the signal conditioner may be called upon to provide auxiliary services, such as introducing electrical isolation, providing a reference of some sort for the transducer, or producing an excitation signal for the transducer.

Important examples of linear operations include *amplitude scaling*, *impedance transformation*, *linear filtering*, and *modulation*.

A few examples of nonlinear operations include obtaining the *root-mean-square (rms) value*, *square root*, *absolute value*, or *logarithm* of the input signal.

There is a wide variety of building blocks available in either modular or integrated-circuit (IC) form for accomplishing analog signal conditioning. Such building blocks include operational amplifiers, instrumentation amplifiers, isolation amplifiers, and a plethora of nonlinear processing circuits such as comparators, analog multiplier/dividers, log/antilog amplifiers, rms-to-DC converters, and trigonometric function generators.

Also available are complete signal-conditioning subsystems consisting of various plug-in input and output modules that can be interconnected via universal backplanes that can be either chassis- or rack-mounted.

## 139.1 Linear Operations

---

Three categories of linear operations important to signal conditioning are amplitude scaling, impedance transformation, and linear filtering.

### Amplitude Scaling

The amplitude of the signal output from a transducer must typically be scaled—either amplified or attenuated—before the signal can be processed.

#### Amplification

Amplification is generally accomplished by an *operational amplifier*, an *instrumentation amplifier*, or an *isolation amplifier*.

**Operational Amplifiers.** A conventional operational amplifier (op amp) has a differential input and a single-ended output. An *ideal* op amp, used often as a first approximation to model a real op amp, has infinite gain, infinite bandwidth, infinite differential input impedance, infinite slew rate, and infinite **common-mode rejection ratio (CMRR)**. It also has zero output impedance, zero noise, zero bias currents, and zero input offset voltage. Real op amps, of course, fall short of the ideal in all regards.

Important parameters to consider when selecting an op amp include:

1. DC voltage gain  $K_0$ .
2. Small-signal **gain-bandwidth product (GBWP)**  $f_T$ , which for most op amps is  $f_T \approx K_0 f_1$ , where  $f_1$  is the lower break frequency in the op amp's transfer function. The GBWP characterizes the closed-loop, high-frequency response of an op-amp circuit.
3. **Slew rate**, which governs the large-signal behavior of an op amp. Slew rates range from less than 1 V/ $\mu$ s to several thousand V/ $\mu$ s.

Other parameters, such as input and output impedances, DC offset voltage, DC bias current, drift voltages and currents, noise characteristics, and so forth, must be considered when selecting an op amp for a particular application.

There are several categories of operational amplifiers. In addition to "garden-variety" op amps there are many op amps whose characteristics are optimized for one or more classes of use. Some categories of op amps include:

1. *Low-noise* op amps, which are useful in the portions of signal conditioners required to amplify very-low-level signals.
2. *Chopper-stabilized* op amps, which are useful in applications requiring extreme DC stability.
3. *Fast* op amps, which are useful when large slew rates and large GBWPs are required.
4. *Power* op amps, which are useful when currents of greater than a few mA must be provided to the op amp's load.

5. *Electrometer* op amps, which are used when very high ( $>10^{13} \Omega$ ) input resistances and very low ( $<1$  pA) input bias currents are required.

An introduction to op amps and basic circuit configurations occurs in essentially any modern text on circuit theory or electronics, and the reader can find detailed theoretical developments and many useful configurations and applications in Roberge [1975], Graeme *et al.* [1971], Graeme [1973, 1977], Horowitz and Hill [1989], and Stout and Kaufman [1976].

**Instrumentation Amplifiers.** Instrumentation amplifiers (IAs) are gain blocks optimized to provide high input impedance, low output impedance, stable gain, relatively high **common-mode rejection (CMR)**, and relatively low offset and drift. They are well suited for amplification of outputs from various types of transducers such as strain gages, for amplification of low-level signals occurring in the presence of high-level common-mode voltages, and for situations in which some degree of isolation is needed between the transducer and the remainder of the instrument.

Although instrumentation amplifiers can be constructed from conventional op amps [a three-op-amp configuration is typically discussed; see, for example, Stout and Kaufman (1976)], they are readily available and relatively inexpensive in IC form. Some IAs have digitally programmable gains, whereas others are programmable by interconnecting resistors internal to the IA via external pins. More basic IAs have their gains set by connecting external resistors.

**Isolation Amplifiers.** Isolation amplifiers are useful in applications in which a voltage or current occurring in the presence of a high common-mode voltage must be measured safely, accurately, and with a high CMR. They are also useful when safety from DC and line-frequency leakage currents must be ensured, such as in biomedical instrumentation.

The isolation amplifier can be thought of as consisting of three sections: an input stage, an output stage, and a power circuit. All isolation amplifiers have their input stages galvanically isolated from their output stages. Communication between the input and output stages is accomplished by modulation/demodulation.

An isolation amplifier is said to provide two-port isolation if there is a DC connection between its power circuit and its output stage. If its power circuit is isolated from its output stage as well as its input stage, then the amplifier is said to provide three-port isolation. Isolation impedances on the order of  $10^{10} -$  are not atypical.

Isolation amplifiers are available in modular form with either two-port or three-port isolation. Both single-channel and multichannel modules are offered.

## Attenuation

Whereas the majority of transducers are low-level devices such as thermocouples, thermistors, resistance temperature detectors (RTDs), strain gages, and so forth, whose outputs require amplification, there are many measurement situations in which the input signal must be attenuated before introducing it to the remainder of the system.

**Voltage Scaling.** Most typically, the signals to be attenuated take the form of voltages. Broadly, the attenuation is accomplished by either a *voltage divider* or a *voltage transformer*.

*Voltage Dividers.* In many cases a simple chain divider proves adequate. The transfer function

of a two-element chain of impedances  $Z_1(s)$  and  $Z_2(s)$  is

$$\frac{V_o(s)}{V_{in}(s)} = \frac{Z_1(s)}{Z_1(s) + Z_2(s)}$$

where the output voltage  $V_o(s)$  is the voltage across  $Z_1(s)$  and the input voltage  $V_{in}$  is the voltage across the two-element combination.

Of course, the impedances of the source (transducer) and the load (the remainder of the system) must be taken into account when designing the divider network.

*Resistive dividers.* If the elements in the chain are resistors, then the divider is useful from DC up through the frequencies for which the impedances of the resistors have no significant reactive components. For  $Z_1(s) = R_1$  and  $Z_2(s) = R_2$ ,

$$\frac{V_o(s)}{V_{in}(s)} = \frac{R_1}{R_1 + R_2}$$

Other configurations are available for resistive dividers. One example is the Kelvin-Varley divider, which has several advantages that make it useful in situations requiring high accuracy. For a detailed description, see Gregory [1973].

*Capacitive dividers.* If the elements in the chain divider are capacitors, then the divider has as its transfer function

$$\frac{V_o(s)}{V_{in}(s)} = \frac{C_2}{C_1 + C_2}$$

This form of divider is useful from low frequencies up through frequencies of several megahertz. A common application is in the scaling of large voltages. *Inductive dividers.* If the elements in the chain divider are inductors, then an autotransformer results. Inductive dividers are useful over frequencies from a few hertz to several hundred kilohertz. Errors in the parts-per-billion range are achievable.

*Voltage Transformers.* Voltage transformers constitute one of the most common means of accomplishing voltage scaling at line frequencies. Standard double-wound configurations are useful unless voltages above about 200 kV are to be monitored. For very high voltages, alternative configurations such as the *capacitor voltage transformer* and the *cascade voltage transformer* are employed [Gregory, 1973].

**Current Scaling.** Current scaling is typically accomplished via either a current shunt or a current transformer.

A *current shunt* is essentially an accurately known resistance through which the current to be measured is passed. The voltage developed across the shunt as a result of the current is the quantity measured. Shunts are useful at DC and frequencies through the audio range. Two disadvantages are (1) that the shunt consumes power, and (2) that the measurement circuitry must be operated at



the same potential as the shunt.

The *current transformer* overcomes the mentioned disadvantages of the current shunt. Typically, the current transformer consists of a specially constructed toroidal core upon which the secondary (sense) winding is wrapped and through which the primary winding is passed. A single-turn primary is commonly used, although multiturn primaries are available.

**Other Attenuators.** In addition to the aforementioned means of voltage and current scaling are attenuator pads, which provide, in addition to voltage or power reduction, the ability to be matched in impedance to the source and load circuits between which it is connected. The common pads include the T, L, and  $\Pi$  types, either balanced or unbalanced. Resistive attenuator pads are discussed in most textbooks on circuit design [e.g., Cuthbert, 1983]. They are useful from DC through several hundred megahertz.

## Impedance Transformation

Oftentimes the impedance of the transducer must be transformed to a value more acceptable to the remainder of the measurement system. In many cases maximum power must be transferred from the transducer's output signal to the remaining circuitry. In other cases it is sufficient to provide buffering that presents a very high impedance to the transducer, a very low impedance to the rest of the system, and a voltage gain of unity.

Matching transformers, passive matching networks such as attenuator pads, and unity-gain buffers are standard means of accomplishing impedance transformation. Unity-gain buffers are available in IC form.

## Linear Filtering

Although, in general, digital signal processing offers many advantages over analog techniques for filtering signals, there are many relatively simple applications for which *frequency-selective analog filtering* is well suited.

Filters are used within signal conditioners (1) to reduce the effects of noise that corrupts the input signal, (2) as part of a demodulator, (3) to limit signal bandwidth, or (4) if the signal is to be sampled, to limit its bandwidth in order to prevent aliasing. These filters can be built either entirely of passive components or based on active devices such as op amps.

There are many good references that discuss methods of characterizing, specifying, and implementing frequency-selective analog filters. See Van Valkenburg [1960] for design of passive filters; for the design of active-RC filters, see Sedra and Brackett [1978] and Stephenson [1985].

## 139.2 Nonlinear Operations

---

There is a wide variety of nonlinear operations useful to signal-conditioning tasks. Listed below are some typical nonlinear blocks along with brief descriptions. Most of the blocks are available as ICs.

1. *Comparator.* A comparator is a two-input device whose output voltage,  $V_o$ , takes on one of

two stable values,  $V_{o0}$  and  $V_{o1}$ , as follows:

$$V_o = \begin{cases} V_{o0} & \text{if } V_2 < V_1 \\ V_{o1} & \text{otherwise} \end{cases}$$

where  $V_1$  and  $V_2$  are the voltages at the two inputs.

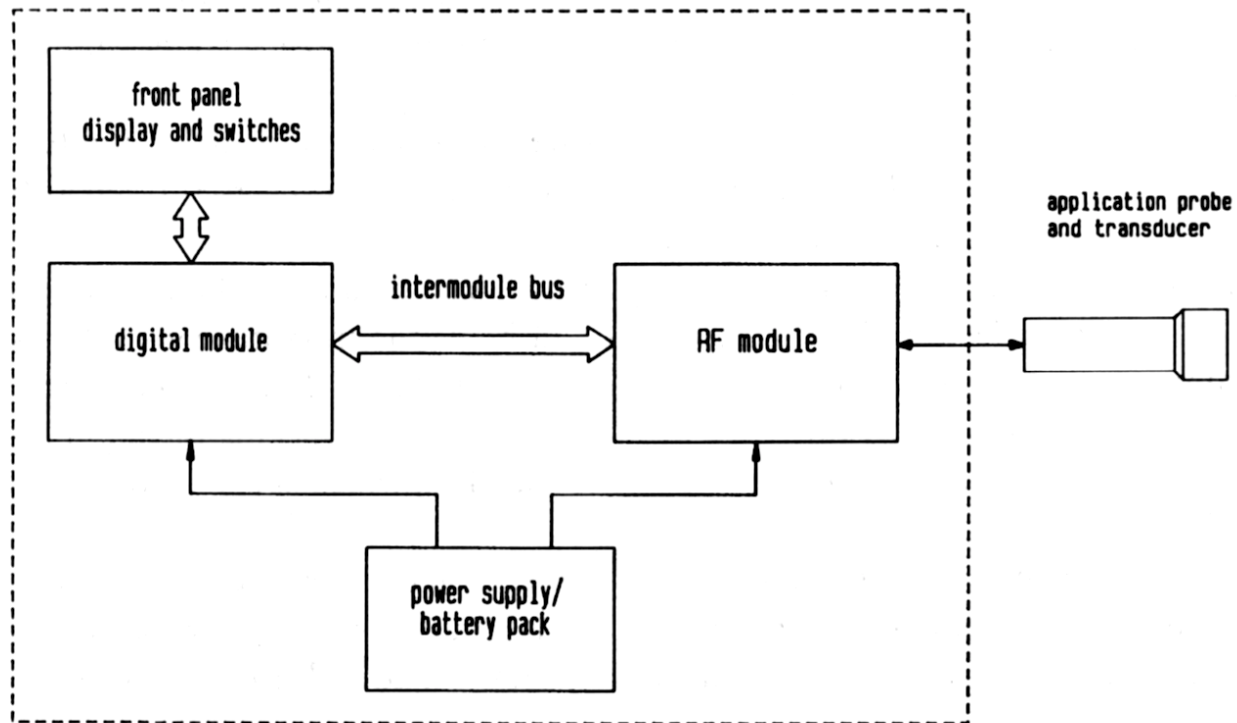
2. *Schmitt trigger*. A Schmitt trigger is a comparator with hysteresis. It can be constructed from a comparator by applying positive feedback.
3. *Multiplier*. A two-input multiplier supplies an output voltage that is proportional to the product of its input voltages.
4. *Divider*. A two-input divider has as its output a voltage proportional to the ratio of its input voltages. The functions of multiplication and division are usually combined within a single device.
5. *Squarer*. A squarer has as its output a voltage proportional to the square of its input. Squarers can be constructed by a number of means: from multipliers, based on diode-resistor networks, based on FETs, and so forth.
6. *Square-rooter*. A square-rooter has as its output a voltage proportional to the square root of its input. A square-rooter can be built most easily from either a divider or a log/antilog amplifier.
7. *Logarithmic/antilogarithmic amplifier*. A log/antilog amplifier produces an output voltage proportional to the logarithm or the antilogarithm of its input voltage.
8. *True RMS-to-DC converter*. A true RMS-to-DC converter computes the square root of the average, over some interval of time, of the instantaneous square of the input signal. The averaging operation is generally accomplished via a simple low-pass filter whose capacitor is selected to give the desired interval.
9. *Trigonometric function generator*. Generators are available in IC form that produce as their outputs any of the standard trigonometric functions or their inverses, taken as functions of the differential voltage at the generator's inputs.
10. *Sample-and-hold and track-and-hold amplifiers*. A sample-and-hold amplifier (SHA) is a device that samples the signal at its input and holds the instantaneous value whenever commanded by a logic control signal. A track-and-hold amplifier is identical to an SHA but is used in applications where it spends most of its time tracking the input signal (i.e., in "sample" or "track" mode), in contrast to the SHA, which spends most of its time in "hold" mode.
11. *Precision diode-based circuits*. Circuits such as precision half-wave rectifiers, absolute-value circuits, precision peak detectors, and precision limiters are relatively easy to design and implement based on diodes and op amps. See Horowitz and Hill [1989], Stout and Kaufman [1976], and Graeme [1977].

A detailed description of these and other nonlinear circuit blocks can be found in Sheingold [1976].

**Example.** We provide briefly an example of a device that has embedded within it several

signal-conditioning circuits. [Figure 139.1](#) shows the basic block diagram of a therapeutic ultrasound unit, which finds widespread use in physical medicine.

**Figure 139.1** Basic block diagram of the therapeutic ultrasound unit discussed as an example.



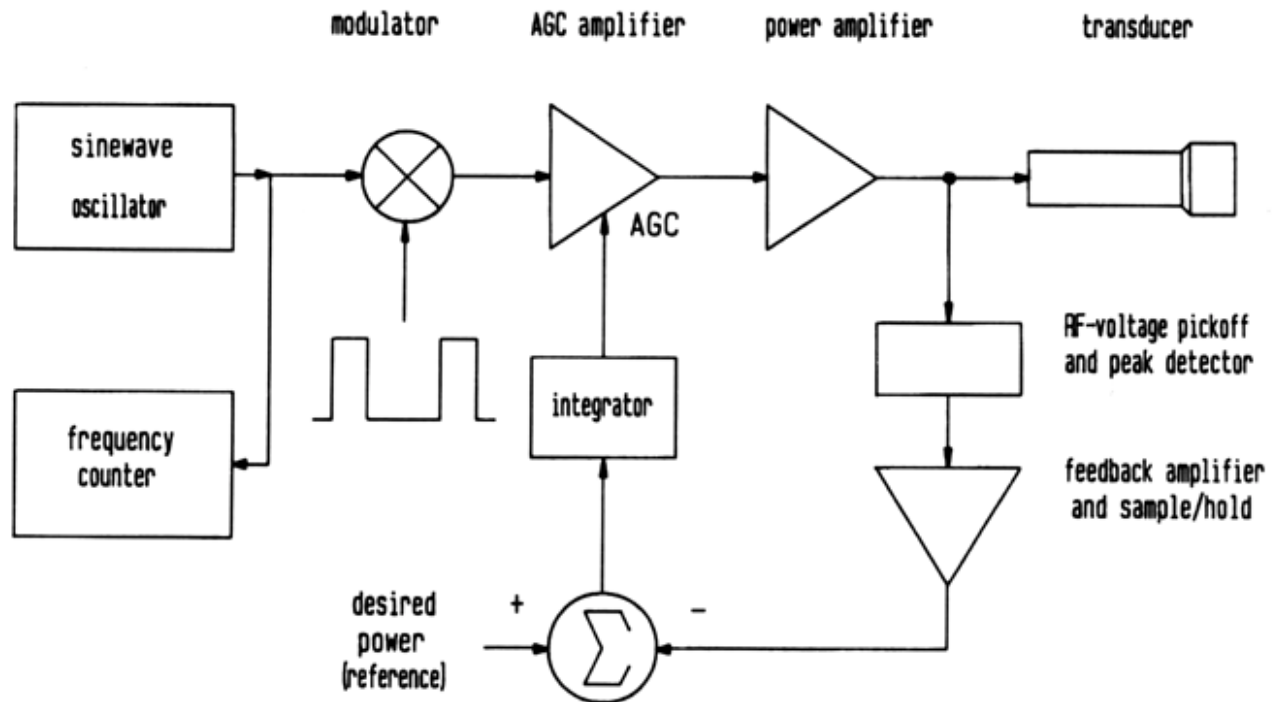
The particular unit being discussed consists of five principal subsystems:

1. An application probe and ultrasound transducer, which imparts ultrasonic energy to the tissue being treated. *Note that this transducer is NOT an input transducer such as has been discussed in relation to signal conditioners.*
2. A radio-frequency (RF) module, which provides electrical excitation to the ultrasound transducer.
3. Front-panel display and switches, which allow communication between the unit and its operator.
4. A microprocessor-based digital module, which orchestrates the overall control of the ultrasound unit.
5. A power supply/battery pack, which provides operating power to the unit.

We focus now on the RF module, whose basic block diagram is shown in [Fig. 139.2](#). The module consists of a sine-wave oscillator that produces a signal at the resonant frequency of the transducer, a modulator that allows that signal to be pulse-modulated, and an amplifier with RF-voltage feedback. Incorporated in the amplifier are a power amplifier capable of driving the transducer and automatic-gain-control (AGC) circuitry required to adjust the output power to coincide with that selected by the operator. The AGC uses a standard feedback-control loop to

maintain a constant-voltage envelope on the RF signal output from the power amplifier.

**Figure 139.2** Simplified block diagram of the RF module used in the ultrasound unit of Fig. 139.1.



Some of the signal conditioners employed within the RF module include the following:

1. The RF-voltage pickoff at the output of the power amplifier. The pickoff employs a half-wave rectifier, followed by a simple capacitive chain divider for voltage scaling.
2. A precision peak detector, which obtains the peak value of the output from the voltage divider during a modulation cycle and presents that value to the feedback loop.
3. An amplifier, having digitally selectable gain, which amplifies the output of the peak detector.
4. A sample-and-hold amplifier, used to hold the amplified output from the peak detector during the "off-time" of the modulator. The SHA is needed since the time constant of the peak detector is not sufficient to prevent significant "droop" during the off-time of the modulator.
5. An integrator (an example of frequency-selective filtering), which develops the control voltage for the AGC loop from the output of the differencer.
6. A current shunt, not shown in Fig. 139.2, that is used to monitor the DC current supplied to the power amplifier.

As can be seen from this simple example, several signal-conditioning functions may be

employed within a single system, and the system itself might not even be an instrument!

## Defining Terms

**Common-mode rejection (CMR):** CMRR given in dB.  $CMR = 20 \log |CMRR|$ . CMR is a nonlinear function of common-mode voltage and depends on other factors such as temperature.

**Common-mode rejection ratio (CMRR):** The ratio of the differential gain to the common-mode gain of an amplifier.

**Gain-bandwidth product (GBWP):** The product of an amplifier's highest gain and its corresponding bandwidth. Used as a rough figure of merit for bandwidth.

**Slew rate:** The maximum attainable time rate of change of an amplifier's output voltage in response to a large step change in input voltage.

## References

- Cuthbert, T. R. 1983. *Circuit Design Using Personal Computers*. John Wiley & Sons, New York.
- Graeme, J. G. 1973. *Applications of Operational Amplifiers*. McGraw-Hill, New York.
- Graeme, J. G. 1977. *Designing with Operational Amplifiers*. McGraw-Hill, New York.
- Graeme, J. G., Tobey, G. E., and Huelsman, L. P. (Ed.) 1971. *Operational Amplifiers*. McGraw-Hill, New York.
- Gregory, B. A. 1973. *An Introduction to Electrical Instrumentation*. Macmillan, London.
- Horowitz, P. and Hill, W. 1989. *The Art of Electronics*, 2nd ed. Cambridge University Press, New York.
- Roberge, J. K. 1975. *Operational Amplifiers*. John Wiley & Sons, New York.
- Sedra, A. S. and Brackett, P. O. 1978. *Filter Theory and Design: Active and Passive*. Matrix, Beaverton, OR.
- Sheingold, D. H. (Ed.) 1976. *Nonlinear Circuits Handbook*. Analog Devices, Norwood, MA.
- Stephenson, F. W. 1985. *RC Active Filter Design Handbook*. John Wiley & Sons, New York.
- Stout, D. F. and Kaufman, M. (Ed.) 1976. *Handbook of Operational Amplifier Circuit Design*. McGraw-Hill, New York.
- Van Valkenburg, M. E. 1960. *Introduction to Modern Network Synthesis*. John Wiley & Sons, New York.

## Further Information

*IEEE Transactions on Instrumentation and Measurement*. Published bimonthly by the Institute of Electrical and Electronics Engineers.

*IEEE Transactions on Circuits and Systems* 34II: *Analog and Digital Signal Processing*. Published monthly by the Institute of Electrical and Electronics Engineers.

*The Best of Analog Dialogue*, 1967–1991. 1991. Analog Devices, Norwood, MA. A collection of practical articles covering circuits, systems, and software for signal processing.

*Analog Devices Special Linear Reference Manual* and *Analog Devices Amplifier Reference Manual*. Present an extensive selection of ICs, modules, and subsystems for signal

conditioning.

Pallás-Areny, R. and Webster, J. G. 1991. *Sensors and Signal Conditioning*. John Wiley & Sons, New York. Provides an excellent introduction to sensors and signal-conditioning circuits required by them.

Sheingold, D. H. (Ed.) 1980. *Transducer Interfacing Handbook*. Analog Devices, Norwood, MA. Covers signal-conditioning techniques applicable to temperature, pressure, force, level, and flow transducers.

Horan, S. "Telemetry"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 140

## Telemetry

---

[140.1 Telemetry Systems](#)

[140.2 Frame Telemetry](#)

[140.3 Packet Telemetry](#)

**Stephen Horan**

*New Mexico State University*

Telemetry systems are found in a variety of applications—from automobiles, to hospitals, to interplanetary spacecraft. Although these examples represent a broad range of applications, they all have many characteristics in common: a natural parameter is measured by a sensor system, the measurement is converted to numbers or data, the data are transported to an analysis point, and an end user makes use of the data gathered—after all, the implication of telemetry is to "measure at a distance."

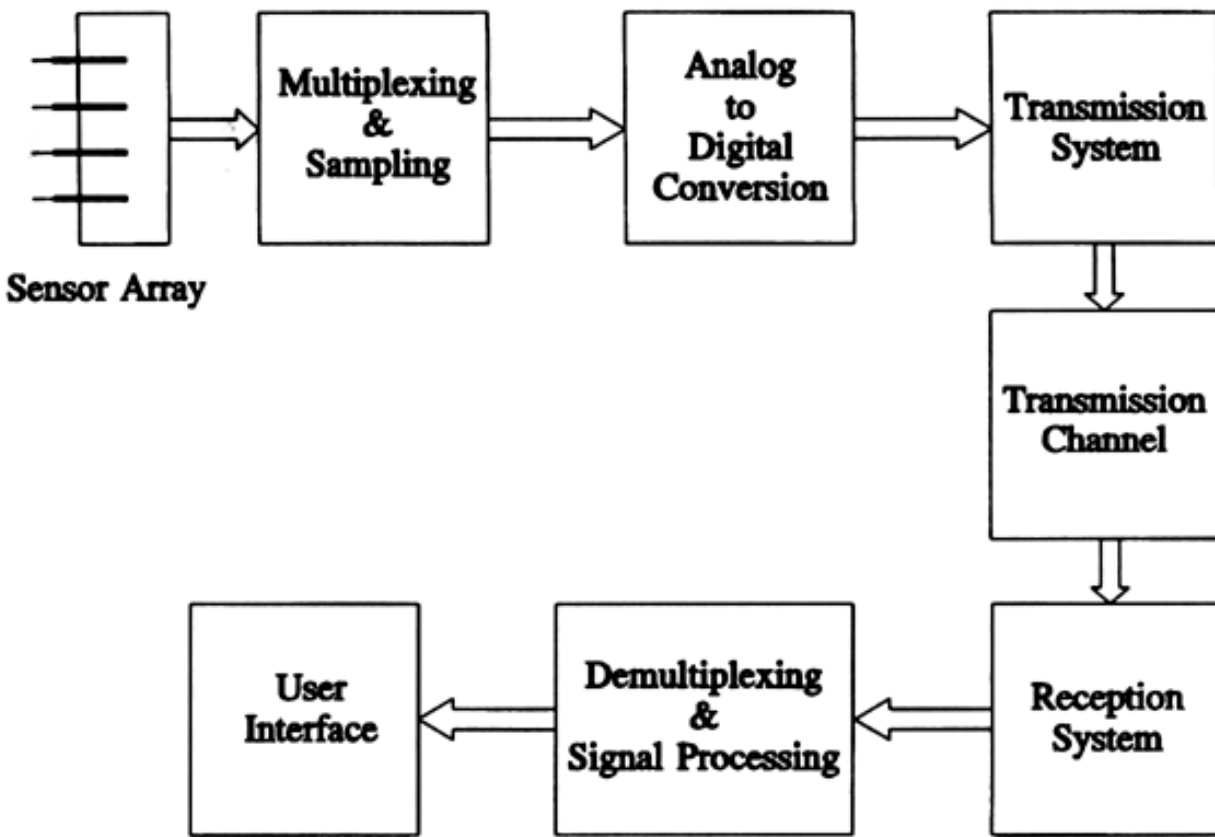
### 140.1 Telemetry Systems

---

The basic telemetry system components are illustrated in [Fig. 140.1](#). The system starts with the sensor array for measuring the natural world. These sensors are specific to the measurement task being performed, for example, atmospheric parameters, flow rates, position measures, accelerations, and so forth. The output of the sensor array is usually an analog voltage level proportional to the signal the sensor is measuring. Some sensors make their output in terms of a current or a time difference instead of a voltage. The sensor output is then sampled at a rate appropriate for the bandwidth of the signal. The sampled signals are then frequently digitized by using an analog-to-digital converter. This produces a **pulse-coded modulation (PCM)** data stream. The channel bandwidth is a limiting factor to the system because it restricts the volume of data that can be reliably sent. Consequently, the digitized signals are then time-multiplexed in a repeating pattern for transmission. When the signals for the sensors are received, the processing stage will demultiplex the data streams into sensor channels. The PCM signal will either be converted back to an analog waveform or left as a series of discrete measurements for use and analysis. The user interface may be a display screen, a set of gages, or a chart recorder. The user interface often contains data-logging capabilities to provide a permanent record of the measurements.



**Figure 140.1** Overall telemetry system components.



The sampling rate for each sensor is determined by the signal's **Nyquist rate**. If  $W$  is the signal bandwidth in hertz, the Nyquist sampling rate,  $f_N$ , in samples per second, is given by

$$f_N = 2W \quad (140.1)$$

In practice, a minimal sampling rate of five times the signal bandwidth is necessary to accurately reconstruct the signal. When the telemetry system is designed, a master clock is established to determine the base rate of signal sampling or the system commutation rate. For most signals this forms the highest sampling rate necessary. Signals having a lower Nyquist rate will be sampled at a rate that is an integer multiple of the sampling period for the higher rate signals. The data are transmitted from the sensors to the users over some type of channel. This channel may be a radio link, a fiber cable, an electrical cable, or a computer network. The choice of which medium is used depends upon the application, the signal bandwidth, the distance to be covered, and the transmission energy available. While the data are being transmitted, the samples may be packaged into frames or packets to give structure to the data and to allow the receiving end to efficiently synchronize to the data stream for processing.

Many systems contain a return command link from the user to the sensor system to allow the sensor system to be controlled. This type of return link is referred to as a *telecommand link*.

## 140.2 Frame Telemetry

Frame telemetry is the traditional method for time-multiplexing data from the source to the



$$P_{FL} = \frac{\sum_{i=0}^k \binom{N}{i}}{2^N} \quad (140.2)$$

where  $N$  is the length of the synchronization code in bits and  $k$  is the number of differences allowed between the received code and the exact code value.

The synchronization code can also be missed if the data become corrupted in the channel. The probability of a missed synchronization code,  $P_M$ , due to channel errors is given by

$$P_M = \sum_{i=k+1}^N \binom{N}{i} p^i (1-p)^{N-i} \quad (140.3)$$

where  $p$  is the channel bit-error rate and  $N$  and  $k$  are as before.

The synchronization process for a telemetry frame follows these steps:

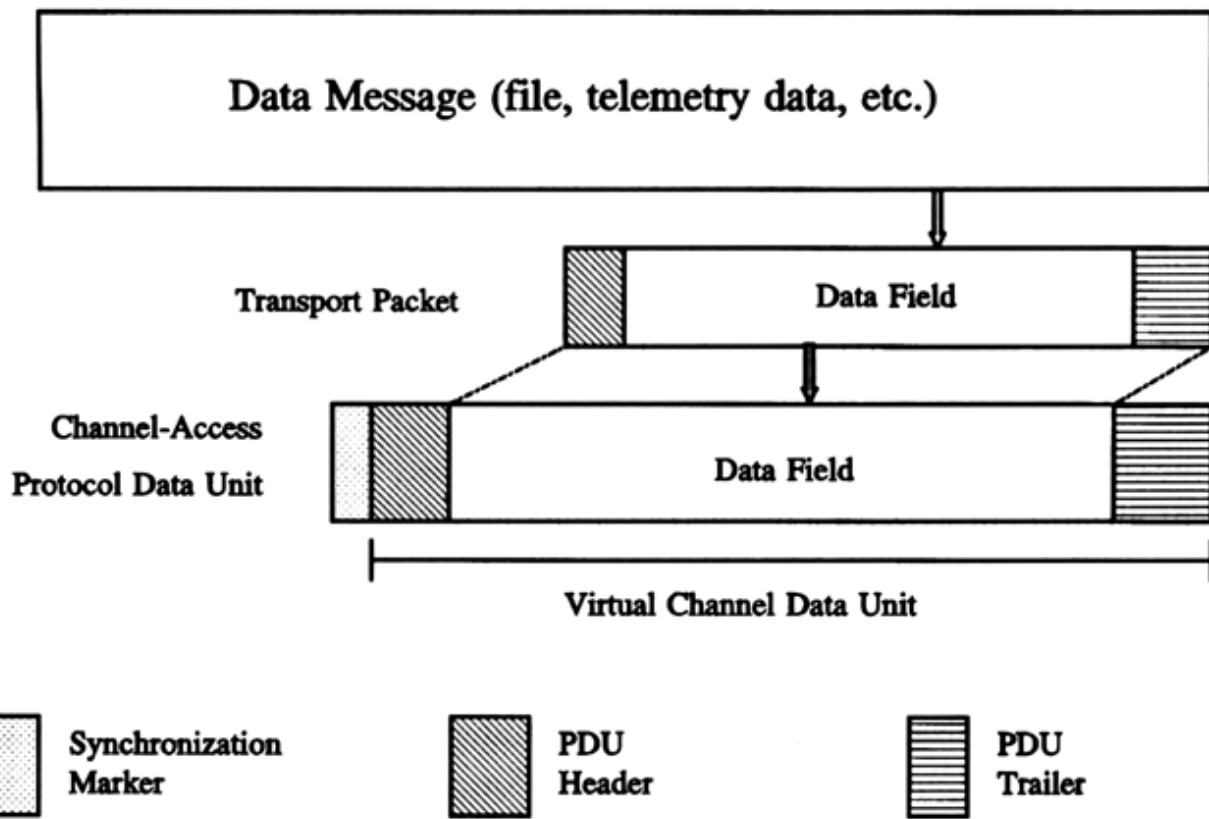
1. Find the individual bits in the data stream using a bit synchronizer circuit.
2. Find the occurrences of the frame synchronization word by using a correlator comparing the data with the desired stored pattern.
3. Once the synchronization marker is reliably found, ensure that it repeats at the minor frame rate and that any management information in the frame is intact by using a frame synchronizer.
4. Once the frame structure is fully identified, begin processing the data.

## 140.3 Packet Telemetry

---

Packet telemetry systems are becoming more common, especially in systems where the data acquisition and data reception subsystems have high degrees of computational capability and the link between them can be viewed as a reliable link. Packet systems have several advantages over frame systems, with the main advantage being flexibility. With a packet system, instead of having a master commutation rate, the sampling rate for each sensor or sensor system can be individualized to the natural signal bandwidth. For example, battery voltages being measured may not change significantly over five minutes. With frame telemetry they may be sampled more frequently than once per minute. With packets they may be sampled only as needed. Packets also have the advantage of allowing the data to be more easily routed over a computer network for analysis and distribution to end users. An example of a packet system is the one developed by the members of the Consultative Committee for Space Data Standards [CCSDS, 1993]. Figure 140.3 shows how the packet system is organized for transmission into individual protocol data units (PDU). The telemetry data are normally organized as a large block of data or a data file. This file is normally too large to be sent in a single packet; therefore, the overall data set is broken into manageable transport packets. These provide end-to-end accounting and reconstruction of the data set. For actual sending across the transmission channel, a channel packet is used. This packet may multiplex the transport packets from several subsystems together for efficient transport. It is common for the channel packets to be sent at regular intervals to maintain transmission synchronization. When this is done, fill packets are used to keep the channel active if there are no actual data to be sent. The packet header will then have a special code to indicate that the packet is a fill packet and should not be processed.

**Figure 140.3** Packet telemetry transmission.



The general packet format is composed of a header, containing accounting and addressing information, followed by the actual data. The packet may end with a trailer composed of error-checking codes or other administrative information. The addressing information in the packet identifies the sensor system originating the packet and the destination process for analysis. Other information included in the header might be counters to identify the sequence number or a time stamp to show when the packet was created. The header will often contain a size parameter to specify the length of the data field.

The packet usually begins with a synchronization marker, just as frame telemetry does. The same synchronization codes can be used in packet systems as in frame systems. After synchronization the header is analyzed to identify the type of processing to be performed based on the source of the data.

## Defining Terms

**Commuted data:** Data that are sampled once per main sampling interval. This main sampling interval is called the *commutation rate*. Minor frames, major frames, and subframes are tied to this commutation rate.

**Major frame:** The set of an integer number of minor frames where each sensor value is sampled at least once.

**Minor frame:** The set of sensor values, synchronization markers, and other management data between successive synchronization words.

**Nyquist rate:** The minimum sampling rate for signal recovery. If a signal of limited bandwidth is sampled at twice this rate, then the signal can be reconstructed in principle. Most signals are sampled at a higher rate than the Nyquist rate to give better reconstruction.

**Pulse-coded-modulation (PCM):** Modulation in which each analog sensor value is converted to a digital number once per sampling interval. The number of bits used in the representation is typically 8 to 16 bits.

**Subcommutated data:** Data whose sampling interval is less frequent than the commutation rate.

**Subframe:** A group of sensors that are subcommutated together. Usually, the data are tied to a single physical subsystem.

**Supercommutated data:** Data whose sampling interval is more frequent than the commutation rate.

## References

Consultative Committee for Space Data Standards. 1993. *Packet Telemetry*. CCSDS 102.1-B-3. NASA, Washington, DC.

Telemetry Group, Range Commanders Council. 1993. *Telemetry Standards, IRIG Standard 106-93*. Secretariat, Range Commanders Council, U.S. Army White Sands Missile Range, NM.

## Further Information

An overview of many aspects of telemetry systems is given by S. Horan in *Introduction to PCM Telemetry Systems*.

The *Proceedings of the International Telemetry Conference* are published yearly by the Instrument Society of America. These proceedings contain theoretical developments as well as system, individual subsystem, and component developments.

There is no single comprehensive journal on telemetry. Various aspects related to data sampling, transmission, and processing are published in IEEE transactions and journals and the *Journal of the International Test and Evaluation Association*.

Owens, W. "Recording Instruments"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Recording Instruments

---

### 141.1 Types of Recording Instruments

Analog vs. Hybrid Recorders • Strip Chart Recorders • Writing Technology

### 141.2 Methods of Data Recording

Chart • Circular Chart • Paperless • Pen Recorders • Multipoint Recorders • Process Recorders • Test Recorders

### 141.3 Future Directions: Distributed Data Acquisition and Recording

#### William Owens

*Johnson Yokogawa Corporation*

A dozen years ago, industrial **recorders** were often considered large, inflexible analog instruments that occupied huge volumes of panel space and offered relatively few modern advantages in return. Slow to scan and easy to wear because of many mechanical parts, chart recorders required frequent maintenance and visual chart interpretation.

Although recorders were a fixture in many process applications, automation professionals declared that they had become outdated. They had no built-in capability for communicating or computing, so plant personnel were required to read and analyze data off-line. In the early 1980s, leading recorder manufacturers began redesigning the instruments to meet the more advanced needs of modern, digital plants.

Today's chart recorders offer speed and high performance in a smaller, more reliable, and more versatile package. Cost of ownership has actually decreased due to the longer life cycle, reduced maintenance, and on-board computing/communication capability of modern recording instruments.

### 141.1 Types of Recording Instruments

---

The engineer is faced with a myriad of selections and options when choosing a recording instrument. Recording and data acquisition instruments are typically classified as strip chart, circular chart, or paperless.

Modern-day recording instruments have added power and expanded capabilities to meet changing market demands. High-resolution data displays, personal computer links, surface-mount technology, application-specific integrated circuits, and microprocessors have changed the instrument from a simple trending on paper to a sophisticated and diverse computing device.

## Analog vs. Hybrid Recorders

The term *hybrid* defines a fundamental separation from conventional, older analog recorders which operate electromechanically and do not incorporate microprocessors.

## Strip Chart Recorders

Strip chart recorders print data on a chart in either a continuous trace or a multipoint arrangement. These continuous-paper-feed instruments are used to record data for virtually all market segments, covering a multitude of applications. In order to select the correct instrument for the application, users must consider sample rate, recording time, number of measuring points, and planned use of recorded data. There are basically two types of strip chart recorders: process and test devices. For the purpose of this chapter, the distinction between process and test classifications for recorders will be based on the method of installation. Process will denote those recorders mounted in a panel or some other fixed piece of equipment, as opposed to test recorders, which require portability and flexibility.

## Writing Technology

Continuous ink pen, print head dot matrix or raster scan with a color ribbon, and thermal print on paper represent the different writing technologies available for recording data on the chart. Colors assist the user in distinguishing the multiple-trace data on the paper. With the increased color selection offered by some manufacturers, ten or more colors in multipoint units, shorter ribbon life may result in units that double-strike the ribbon to increase color selection.

## 141.2 Methods of Data Recording

---

### Chart

Chart recording of data is the most commonly used form. Typical paper sizes range from 100 mm (4 in.), as in the instrument shown in [Fig. 141.1](#), to 250 mm (10 in.). Chart width considerations include number of points and resolution—the wider the chart, the greater the resolution. The addition of microprocessors to recorders, digital notation of data and digital logging minimize the resolution consideration. Added information to the chart record, such as tags, units, headers, and messages, are available in many units.



**Figure 141.1** MicroR1000 4-in. recorder.

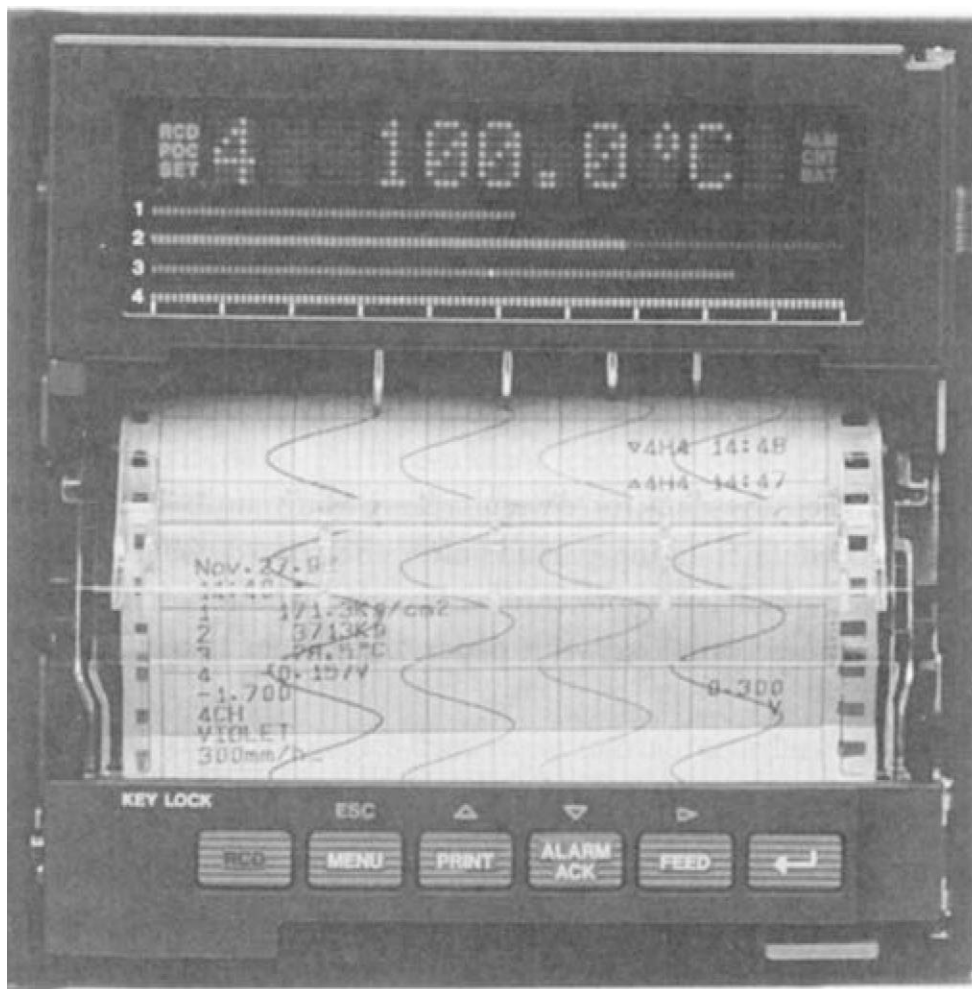


Chart records are required for several industries to comply with regulations such as EPA (environmental monitoring), FDA (food and pharmaceutical), and NERC (power grid interchange). Storage of the chart requires little special handling. Review of the chart is simple for the user. Many recorders offer fanfold paper, which allows the user to view several days or weeks of data with little effort.

## Circular Chart

Similar to the strip chart recorder, the distinctive advantage of the circular chart is the ability to look at 24 hours, 7 days, or 30 days of trace information without operator intervention. Paper sizes range up to 12 inches. Many circular chart models offer math, totalization, and PID control loops.

## Paperless

Paperless is a term referring to the growing variety of recorders that don't use paper. Basically, the

paperless recorder replaces the chart with a CRT, LCD, or TFT display. Some recorders incorporate chart data recording and electronic medium storage simultaneously. A paperless recorder is also a form of data acquisition. Like the traditional chart recorder, the paperless unit has advanced to do more than just record data. Serving as a front end, the recorder performs signal conditioning for I/O to a computer, programmable controller (PLC), or distributed control system (DCS) host. A paperless recorder stores data to an electronic medium such as hard disk, IC memory card, floppy disk, optical disk, or magnetic tape.

The amount of time needed to store data, the sample rate, and the capacity of disk, card, or tape are issues that must be resolved to ensure sufficient storage capacity. The question of what is required to enable the user to view or produce graphs of the data after collection should also be explored. Some store in ASCII comma-delimited files, which require importation to a spreadsheet or database program. Others provide programs that allow the user to select the type of file for data conversion, such as parsed data files with .WKS extensions, which can be read directly in most spreadsheet or database programs.

The initial cost of the paperless unit is higher than that of its counterpart with paper for comparable point capacity. A careful cost analysis must be performed to determine cost of ownership for both options, which includes pens and charts for the paper model and disk and effort to store data for the paperless models. The expected life of paperless displays for continuous operation also should be factored into the evaluation. Most displays have a limited life in continuous operation.

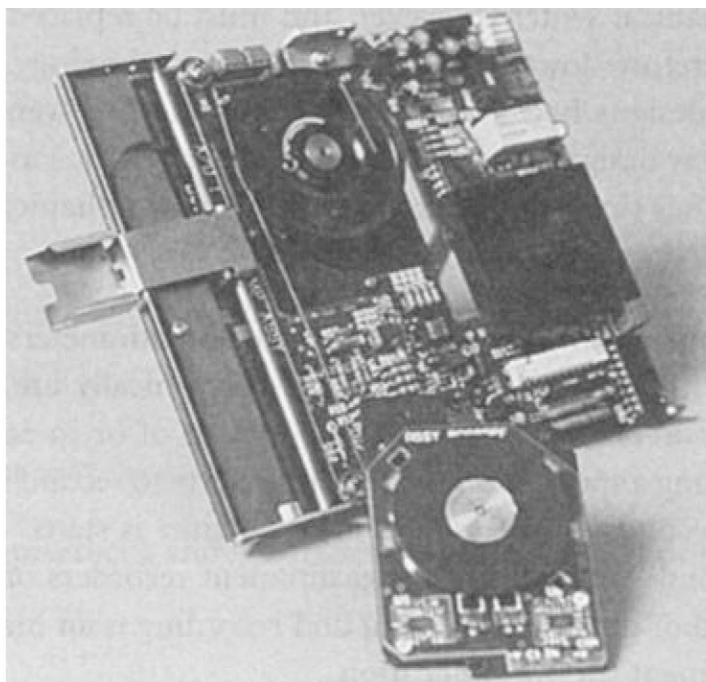
What must one do to keep the data for several years? Many units require additional steps to maintain or archive stored data for long periods after the recorder has collected the electronic information. One example is the IC memory card, although the battery fails in time. Disk and tapes also have limited storage times compared with paper, but longer than those of battery-powered devices.

## Pen Recorders

Pen recorders have one dedicated analog-to-digital converter per input. These are typically used in faster applications such as flow or other rapidly changing inputs. The pen recorder is better able to graph a true representation of the input signal. In some applications, like those found in the NERC regulation for power interchange monitoring, a continuous-writing pen-to-paper chart recorder is specified.

Modern advances of servo pen continuous recorders have dramatically reduced the maintenance associated with old designs incorporating slide wires with contacts, DC brush motors, and wire cables to position the pen. New or improved technologies available in pen recorders are DC brushless servo motors, as shown in [Fig. 141.2](#), and noncontact ultrasonic positioning for the servo pen, which have greatly reduced maintenance by eliminating wiper contacts for pen position and the brushes in the DC motor.

**Figure 141.2 Noncontact servo with brushless DC motor.**



## Multipoint Recorders

Multipoint recorders have several inputs sharing an analog-to-digital converter. The advantage is the recorder's ability to accept more inputs in a given space. Due to the slower scan speed in most models, the multipoint recorder is best suited for more slowly changing signals, such as most temperature measurements. It should be noted that some manufacturers have made significant improvements in scan speed. Modern-day instruments meet or exceed a rate of 1 second for all points.

Recording times also have improved and must be evaluated to determine the suitability of the chart for the data application. Some manufacturers use the latest technology to increase reliability and quality of measurement data. Surface-mount circuit board design and other advanced technologies increase the mean time between failure (MTBF) for electronic components.

Inputs are switched using high-breakdown-voltage, solid state switches to give higher quality and consistency of measurements. In units that use solid state relays, no contact resistance change as a result of make-and-break contact wear or corrosion is possible, nor are mechanical input switch failures.

Following are typical switch technologies used in industrial recorders:

*Electro-mechanical relay:* This traditional technology is subject to wear and regular replacement. Changes in contact surfaces alter resistance, which affects measurement consistency. Improved contact material extends life only marginally. This type of relay can withstand high electrical noise spikes and offers low contact resistance.

*Reed relay:* This type of mechanical relay technology can withstand normal industrial noise

voltages. Low switch resistance allows measurements of low-level voltage signals (such as RTDs). The reed relay is a mechanical switch, however, and must be replaced. The reed relay requires "settling time" and is therefore slower than solid state relay technology.

*Solid state relay:* Initial designs had 30–50 V voltage limits and were subject to failure with high-frequency noise. New designs have overcome these limitations to provide the best of both worlds: noise characteristics similar to hard-contact, electro-mechanical relays and the long life of an integrated circuit.

From a maintenance point of view, units that eliminate potentiometers for adjustments in favor of ADC calibration, with all adjustments accomplished electronically, are preferred.

Some recorder manufacturers state speed in hertz instead of or in addition to scan rate. The simplest method of translating a speed specification from hertz to second is to apply the conversion 1 hertz equals 1 scan per second. Therefore, if a recorder time is stated as 30 hertz, the effective scan rate is 30 times a second. Some test and measurement recorders offer speeds as fast as 100 kHz. The scan rate or speed of data measurement and recording is an important consideration in choosing the correct instrument for the application.

## Process Recorders

Process recorders consist of signals that monitor either continuous or batch process parameters such as temperature, pressure, flow, pH, conductivity, and/or the 4–20 mA inputs found in power generation, refineries, chemical plants, cement plants, pharmaceutical and biotech applications, and others. Most of the signals are slower-changing, requiring a slower scan or sample time.

## Test Recorders

Test recorders are found primarily in laboratory applications and applications that require faster sample collection of data. Test recorders are also found in a diverse market encompassing inspection, R&D, and quality control. Commonly measured inputs are strain gage, accelerometers, pressure, AC transients, power line monitoring, and transducers.

## 141.3 Future Directions: Distributed Data Acquisition and Recording

---

Recorders have added features that greatly expand their role in conditioning data as front-end devices in addition to performing basic recording. Offering universal inputs, recorders are involved in signal conditioning, providing alarm annunciators, relay outputs, digital display capability, providing chart records with digital notation of alarms and messages, and front-end data to PC, PLC, and DCS systems.

With increased power, speed, and communication capability, the recording instrument is

becoming a distributed data acquisition and recording key component. The differences in performance will narrow between process and test recorders.

The present trend of enhanced functionality will continue into the foreseeable future. Although paperless recorders are beginning to appear, the need for hard-copy chart records for regulatory agencies will keep traditional paper recorders on the market for a long time to come.

Recorders will grow in scope as increased microprocessor power yields benefits in speed, function, and application versatility. Recording instruments will continue to move into more sophisticated applications and will assume nontraditional roles as smart network devices and data nodes. With the move toward fieldbus interoperable communications, recording instruments may incorporate elements of control and offer entirely new presentations of data formats.

## **Defining Term**

**Recorder:** A device that provides historical data records to a specific medium and with a given accuracy from sensor inputs.

## **Further Information**

For buyer's guide and market analysis with application information:

Measurements & Control  
2994 West Liberty Ave.  
Pittsburgh, PA 15216-9923

For applications, new products, and new technology:

Intech  
ISA Services, Inc.  
P.O. Box 5272  
Pittsfield, MA 01203-9878  
Control  
Creative Data Center  
650 S. Clark St.  
Chicago, IL 60605-9914

von Maltzahn, W. W., Meyer-Waarden, K. "Bioinstrumentation"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

[142.1 Basic Bioinstrumentation Systems](#)[142.2 Applications and Examples](#)[142.3 Summary](#)**Wolf W. von Maltzahn***University of Karlsruhe***Karsten Meyer-Waarden***University of Karlsruhe*

Biomedical instruments measure, amplify, process, record, and store physiological quantities to help physicians diagnose illnesses, set up treatment plans, or restore lost functions. The scope of such instruments is enormous, both in complexity and in the range of applications. Only a few can be covered in this chapter. Since a choice among topics must be made, this chapter focuses primarily on instruments that measure physiological signals; it does not cover devices that treat diseases or restore compromised physiological functions. For further information, the reader should consult the references.

In contrast to technical systems, physiological systems comprise living cells and organs. This requires that measurements do not adversely affect the living system by introducing toxic substances; by destroying delicate tissues; or by otherwise interfering with the chemical, electrical, or mechanical balance of the living cell or organ. Furthermore, many physiological measurement sites are either inaccessible for direct measurements or only accessible invasively—that is, by puncturing the skin. Noninvasive biomedical instruments often use indirect measurement methods, even though these methods are usually less accurate, slower, and less informative than their direct counterparts.

Physiological measurements seldom depend on one variable alone. The electrical impedance of tissue, for instance, depends not only on resistivity, cross-sectional area, and length, but also on electrolytes, enzymes, temperature, and other factors related to life processes. To extract meaningful information from impedance measurements, the system needs to focus on one of these factors and suppress the others. Physiological events in living systems are unstationary, stochastic, time-dependent, and noisy. It takes a lot of skill and ingenuity to measure them.

Biological cells pump primarily sodium and potassium ions across the cell membrane to maintain specific concentration differences. The resulting membrane potential of about  $-90$  mV

can be calculated from a modified Nernst equation. Nerve and muscle cells are excitable cells and generate action potentials by first changing the membrane permeability of sodium and then that of potassium. Action potentials influence the electric fields in the tissues surrounding the excitable cells all the way to the skin, from where they can be detected and recorded as surface biopotentials. These biopotentials are characterized by high source impedances, small signal voltages, significant interference voltages, and a modest frequency range. Special biopotential amplifiers convert these biopotentials into high-quality signals by amplifying them, suppressing interferences, and preparing them for further signal processing, display, or recording.

Biomedical instruments operate in hospital rooms, where they interact with other devices, work in the vicinity of other devices, or connect to patients. Such interconnections not only cause unpredictable interferences, but also provide current pathways that may endanger the lives of patients and/or operators. To ensure high-quality measurements and electrical safety, requirements must be imposed on biomedical instruments, uncommon in typical industrial measurement systems. Biomedical measurements must do the following:

- Amplify the biosignal and suppress interferences
- Protect input stages against high voltages generated by other devices
- Be electrically safe for operator and subject
- Avoid adversely affecting the living system

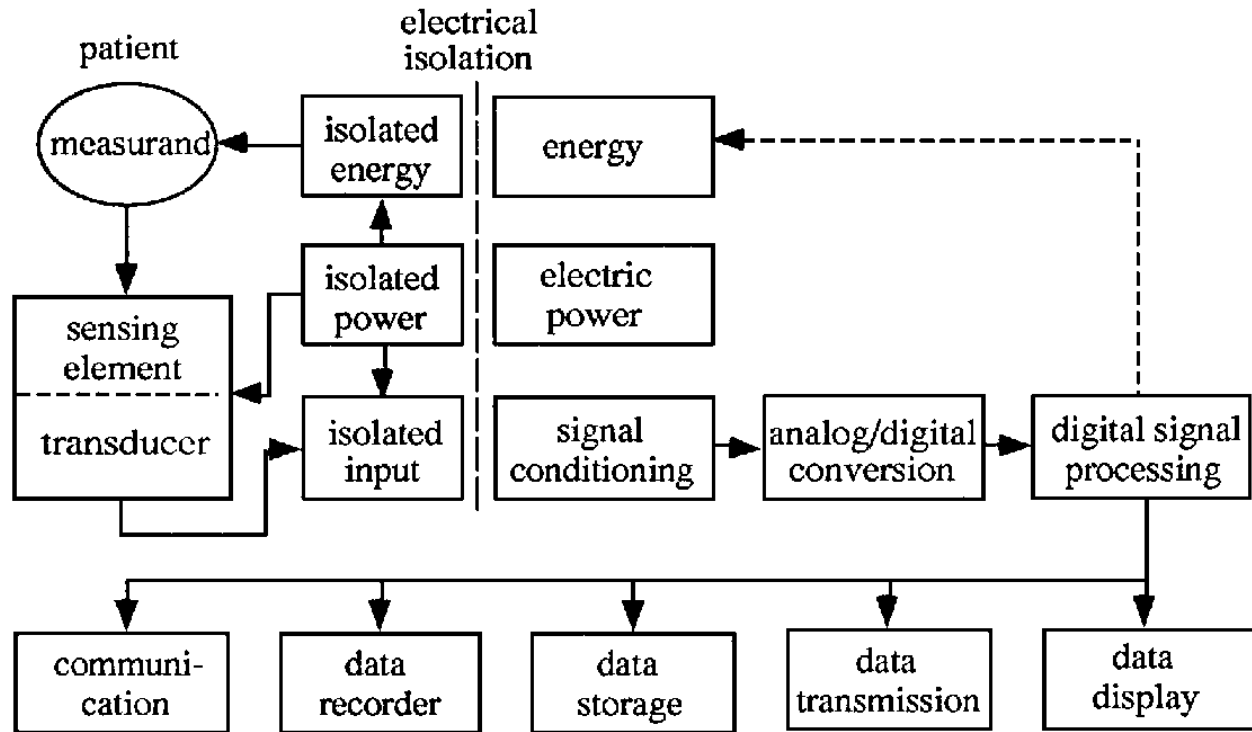
## 142.1 Basic Bioinstrumentation Systems

---

[Figure 142.1](#) shows the block diagram of a basic bioinstrumentation system. The physiological variable to be measured, the measurand, may be the result of a molecular, cellular, or systemic event and be mechanical, electrical, or chemical in nature. Blood pressure and blood flow, for instance, are important mechanical variables of the circulatory system. The electrocardiogram, electromyogram, and electroencephalogram provide information on the electrical activity of the heart, skeletal muscle, and brain, respectively. Partial pressures of oxygen or carbon dioxide reflect the status of the chemical balance in the blood. The sensing element in [Fig. 142.1](#) interacts with the measurand, directly or indirectly, and is designed with three main goals in mind: to minimize the unavoidable disturbance of the measurand and its environment, to avoid interference with life processes, and to provide an output signal sensitive primarily to the measurand and insensitive to other parameters.



**Figure 142.1** Schematic block diagram of a biomedical instrument.



The transducer takes the output of the sensing element and transforms it into an electrical signal. The most common analog signal-conditioning elements are amplifiers, filters, rectifiers, triggers, comparators, and wave shapers, just like in other instrumentation systems. Appropriate amplification and filtering prepare the analog signal for analog-to-digital conversion (ADC). Once in digital form, a microcomputer further processes the signal and displays it on a CRT screen or on an LCD panel, generates a hard copy on a recorder, sends it to another device, or stores it in a mass storage device. Some measurements require external energy or stimulation applied to the subject across an isolation barrier.

Medical instruments contain special circuits to safeguard patients and operators against electrical shock, to suppress interferences from the noisy hospital environment, and to protect sensitive input stages. The most important of these is the isolation barrier, made of transformers, capacitors, optocouplers, or a combination of these. The isolation barrier keeps current densities in the body below the safe level of  $100 \mu\text{A}/\text{cm}^2$ . Several national and international standards give limits for safe voltages and currents, describe methods to avoid potentially dangerous connections, and give detailed test procedures to ensure electrical safety [NFPA, 1990; AAMI, 1990; IEC, 1982].

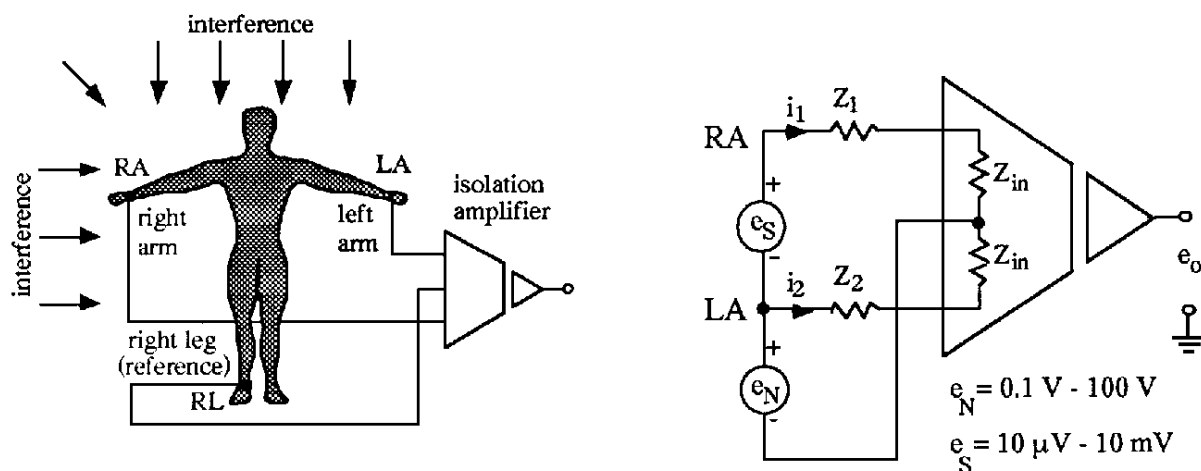
A single hospital room may contain electrocardiographs, blood pressure monitors, X-ray machines, cardiac output monitors, respirators, electrosurgery devices, defibrillators, and other devices. The useful and beneficial output of one device may severely interfere with another. To function in the same environment, these devices must contain special hardware circuits or software algorithms that suppress undesired outputs from other devices.

## 142.2 Applications and Examples

Action potentials of nerve and muscle cells inside the body manifest themselves on the surface of the skin as small voltages between  $10\ \mu\text{V}$  and  $100\ \text{mV}$  with high source impedances and high levels of interfering noise signals. These surface biopotentials are detected with surface electrodes that function as transducers between the ionic currents inside the body and electronic currents in wires and amplifiers. Electrodes represent complicated electrochemical systems, as described in many bioinstrumentation books [Aston, 1990; Bronzino, 1986; Geddes and Baker, 1989; Norman, 1988; Profio, 1993; Webster, 1992]. Most electrodes for recording surface biopotentials are made of silver or silver chloride because they provide stable and relatively noise-free recordings.

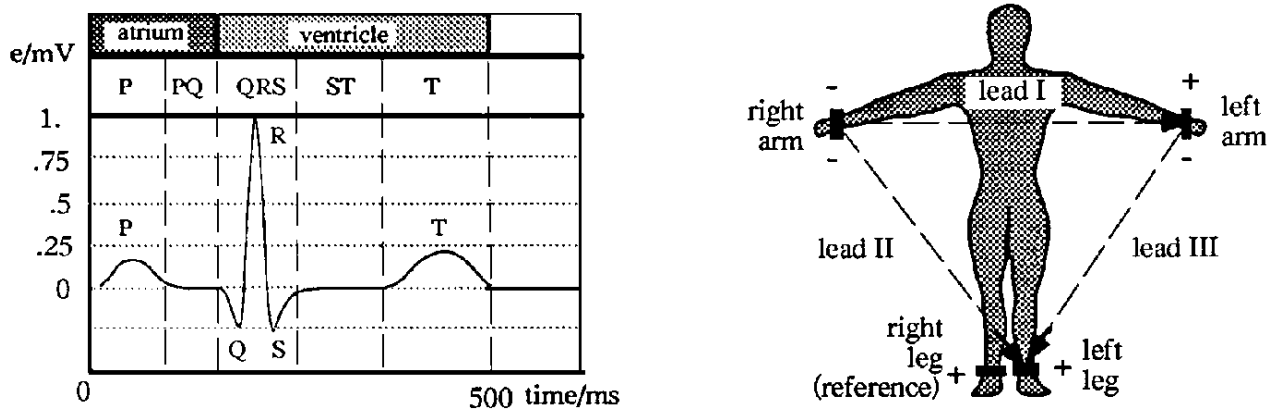
The diagram in Fig. 142.2 for obtaining a lead I electrocardiogram serves as an example of how to obtain a high-quality biopotential recording in general. Three electrodes attach to the subject—two for recording the biopotential and one for providing a reference potential. The electrodes on the left and right arms connect directly to the differential input of an isolated instrumentation amplifier; the reference electrode on the right leg connects to the floating ground terminal. This configuration separates the desired differential-mode biosignal  $e_S$  from the undesired common-mode noise  $e_N$ , as shown in the electric equivalent circuit of Fig. 142.2. The instrumentation amplifier provides a high differential-mode gain and a low common-mode gain to suppress power line interferences and electrode potentials. The ratio between these two gains is called common-mode rejection ratio (CMRR). Although the CMRR of a good instrumentation amplifier may be as high as 130 dB, the CMRR of the overall circuit is seldom greater than 80 dB or 10000:1. The primary cause for this large reduction in CMRR lies in the differences in electrode and skin impedances, represented by  $Z_1$  and  $Z_2$  in Fig. 142.2. Some bioinstrumentation systems use additional amplifiers that set the voltage on the right leg equal to the mean voltage between the left and right arm ("driven right leg"), thereby decreasing common-mode voltages.

**Figure 142.2** Left panel: schematic diagram for measuring biopotentials; right panel: electrical equivalent circuit.



The most important and most investigated biopotential waveform is the *electrocardiogram* (ECG) [Webster, 1992]. Since the electrocardiogram represents the heart's electrical activity associated with cardiac contraction, it provides diagnostic insight into many heart functions. The curve in the left panel of Fig. 142.3 represents a typical ECG waveform with the standard P, QRS, and T labels. The P wave and the QRS complex represent depolarization of cardiac muscle cells—the first one of atrial cells and the second one of ventricular cells. Although atrial repolarization cannot be seen in the ECG, the T wave represents ventricular repolarization. These P waves and QRS complexes precede cardiac contraction and the pumping of blood, first of the atria then of the ventricles.

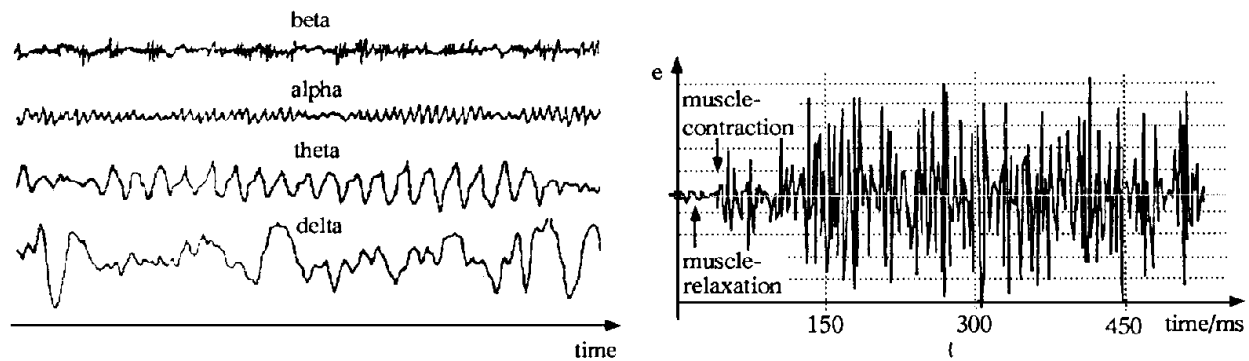
**Figure 142.3** Left panel: typical lead II ECG recording; right panel: Einthoven triangle.



To enable comparisons of waveforms between subjects, certain standardized recording techniques have evolved. Einthoven, the pioneer of electrocardiography, developed an extremity lead system still in use today. As shown in the right panel of Fig. 142.3, the three recording electrodes of the Einthoven triangle are placed on the left arm, the right arm, and the left leg, whereas the reference electrode is placed on the right leg. The voltage differences between the three recording electrodes are known as lead I, lead II, and lead III recordings. The lead II recording from the right arm to the left leg runs almost in parallel to the main axis of the heart and is therefore the preferred recording of a single-channel ECG. In addition to Einthoven's recordings, physicians use augmented limb leads and unipolar chest leads.

Recordings of the electrical activities of the brain, called *electroencephalograms* (EEGs), are more difficult to obtain than recordings of the ECG. The skull is a poor electrical conductor and EEG voltages on the scalp are in the  $\mu\text{V}$  range, as opposed to ECG voltages in the mV range. Furthermore, it is difficult to relate EEG voltages to specific neuronal activities; they are the net result of many different neurons firing seemingly independently of each other. With a subject in a relaxed state with closed eyes, alpha waves are easily recorded between two electrodes on the scalp. As shown in Fig. 142.4, alpha waves are simple periodic waveforms in the frequency band of 8 to 13 Hz with slightly varying amplitudes. Alpha waves disappear when the subject's eyes are opened. The other waveforms shown in Fig. 142.4 have clinical significance, particularly in diagnosing epilepsy and sleep disorders or in providing feedback during general anesthesia. The low-frequency theta waves between 4 and 8 Hz indicate sleep, whereas the high-frequency beta waves between 14 and 22 Hz appear during high states of alertness.

**Figure 142.4** Left panel: typical EEG waveforms; right panel: typical EMG waveform.

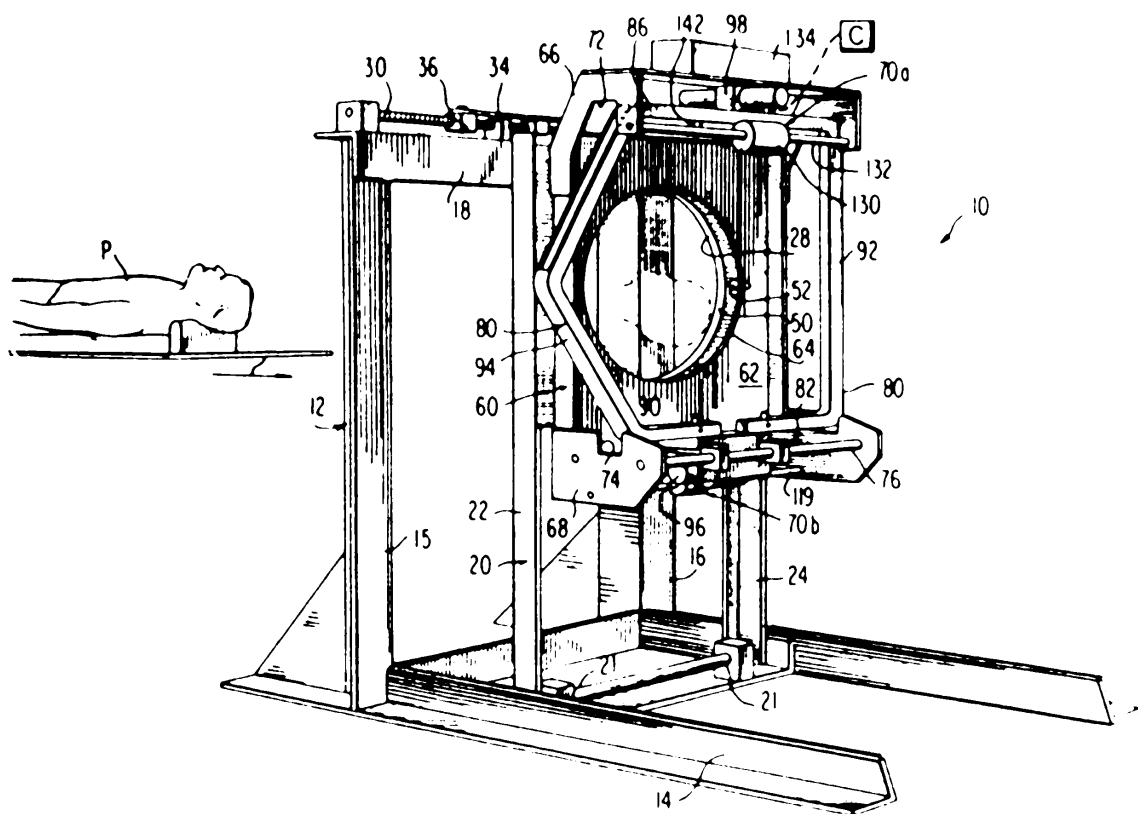


Recordings of the electrical activities of skeletal muscles are called *electromyograms* (EMGs). They can be obtained with surface electrodes similar to ECG electrodes or needle electrodes. The latter puncture the skin and are positioned close to the muscle group, from which they record. The mean amplitude and mean frequency of the EMG power spectrum reveal both muscle strength and muscle fatigue, which can provide feedback information in rehabilitation engineering and functional electrical stimulation [Phillips, 1991]. Table 142.1 summarizes amplitudes and frequency ranges of ECG, EEG, and EMG. Similarly, other biopotentials originating from other nerve or muscle cells can be recorded from the surface of the body.

**Table 142.1** Amplitude and Frequency Ranges of ECG, EEG, and EMG

Origin of Electrical Activity	Name	Amplitude (mV)	Frequency Band(Hz)
Cardiac muscle	ECG	1–10	0.01–150
Skeletal muscle	EMG	5–100	10–5000
Brain	EEG	<0.2	1–30
	$\beta$	0.01	14–30
	$\alpha$	0.03	8–13
	$\theta$	0.05–0.1	4–7
	$\delta$	0.1–0.15	1–4

Bioelectric impedance measurements provide qualitative and quantitative information about volume changes in the heart and in peripheral arteries and determine body characteristics such as percent body fat, total body fluid volume, and cell volume. They are also used in sleep apnea monitoring, especially in infants, and in the detection of venous thrombus. Multiple bioelectric impedance measurements lead to computed cross-sectional images of the body, so-called *computed impedance tomograms*.



## DIAGNOSTIC X-RAY SYSTEMS

*Robert S. Ledley*

*Patented November 25, 1975*

*#3,922,522*

An excerpt:

It is the principal objective of this invention to improve the state of the prior art in contributing a tomographic scanning instrument having a degree of efficiency, practicality, and flexibility heretofore unknown.

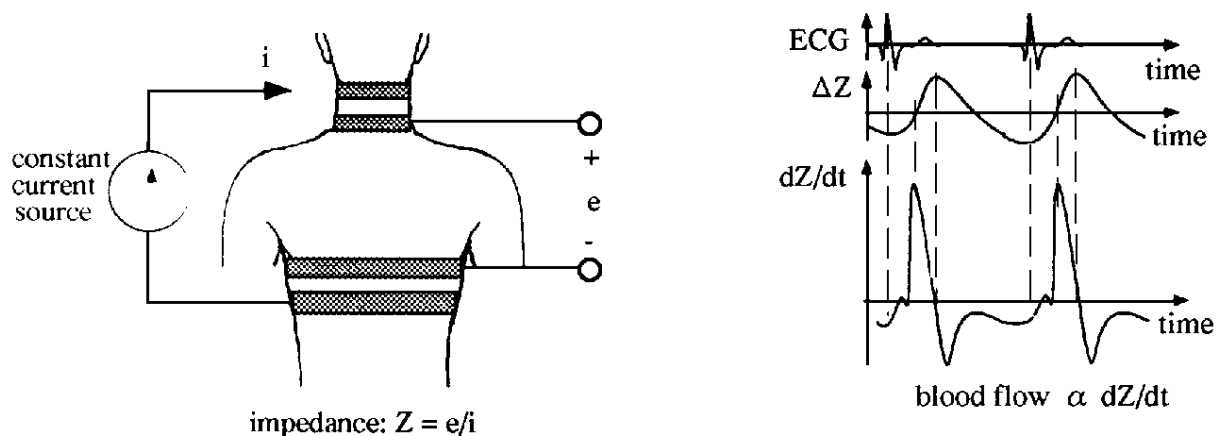
Another principal objective of this invention is to provide a tomographic scanner having a capability of scanning any portion of the human body with equal efficiency and accuracy.

Another important object of the invention is to provide a medical diagnostic apparatus of the type described in which the plane of the scanner can be tilted over a relatively wide range of attitudes.

Taking X-ray images of organs in slices (tomography) had been known as a valuable (and cumbersome) diagnostic tool for some time, but Ledley developed in one neat package a machine for capturing these whole-body image slices and reconstructing them for interpretation by physicians. This technique is now known as Computer Aided Tomography or CAT scanning. (©1992, DewRay Products, Inc. Used with permission.)

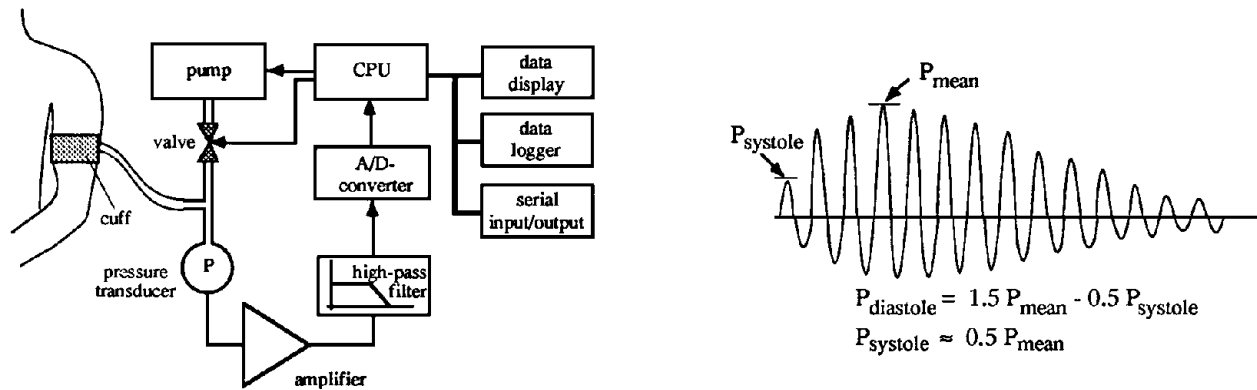
Figure 142.5 shows four band electrodes for the measurement of respiration, stroke volume, and cardiac output. A constant sinusoidal current with an amplitude in the range of 0.5 to 4 mA and a frequency between 50 and 100 kHz flows from the top electrode through the thorax to the bottom electrode. This current is maintained independent of skin or tissue impedances. The voltage drop between the two middle electrodes is amplified by a voltage amplifier with a high input impedance. The output voltage  $e_o$  is proportional to impedance changes  $\Delta Z$  that arise from thoracic volume changes due to respiration and from blood volume changes due to the pumping heart. Figure 142.4 shows the curves for  $\Delta Z$  and  $dZ/dt$ —the former being proportional to respiratory volume changes and the latter to blood flow.

**Figure 142.5** Electrode positions and typical waveforms of electrical impedance of thorax measurements.



The noninvasive measurement of blood pressure is important in the diagnosis of many cardiovascular problems. Blood pressure varies—often within seconds—to meet the physiological demands of organs and muscles. Although single blood pressure readings are valuable for entry-level screening, they do not provide as much information as monitoring over a specified length of time, usually 24 hours. The most common body-worn instrument for monitoring blood pressure noninvasively is a battery-powered oscillometric blood pressure monitor. In this monitor an inflatable cuff fits snugly around the upper or lower arm and connects to a pump and pressure transducer. The microcontroller (CPU) triggers the pump periodically to inflate the cuff up to a pressure slightly greater than systolic arterial pressure. It then controls the valve to release the cuff pressure at a rate of about 4 mmHg/s and measures the cuff pressure. The resulting cuff pressure curve consists of a descending curve superimposed by small periodic oscillations. These oscillations start just before the cuff pressure equals systolic pressure, increase rapidly, reach the maximum amplitude at mean arterial pressure, and then gradually decrease. A special algorithm suppresses motion artifacts, extracts systolic and mean pressure ( $P_s$  and  $P_m$ ) from the pressure curve, and calculates the diastolic arterial pressure. The oscillations also permit the calculation of heart rate. Blood pressure data, heart rate, date, and time are displayed on an LCD panel and stored in a data logger. The serial output permits up-loading data to another computer for further evaluations. Figure 142.6 shows the block diagram of this device and an annotated typical oscillometric waveform.

**Figure 142.6** Block diagram and waveforms of noninvasive oscillometric blood pressure monitor.



## 142.3 Summary

This chapter has focused primarily on electrical measurement techniques and examples thereof. Many interesting subjects and instruments could not be covered, such as blood flow and blood volume, partial pressures of oxygen and carbon dioxide, oxygen saturation, glucose, and enzyme concentrations, to mention only a few. The emerging technologies related to home care could not even be mentioned. For further reading the interested reader is referred to the texts listed at the end of the chapter.

## Defining Terms

**Action potential:** The reversible depolarization of the membrane potential of an excitable cell in response to a mechanical, electrical, or chemical stimulus. The peak action potential of a single cell is about 70 mV.

**Common-mode rejection ratio (CMRR):** The ratio of difference-mode gain over common-mode gain in a difference amplifier. It is a measure of the degree to which common-mode signals are suppressed in relation to difference-mode signals.

## References

- AAMI. 1990. *Design of Clinical Engineering Quality Assurance and Risk Management Programs*. Association for the Advancement of Medical Instrumentation, 3330 Washington Blvd., Suite 400, Arlington, VA 22201-4598. Phone (800)332-2264.
- Aston, R. 1990. *Principles of Biomedical Instrumentation and Measurement*. Merrill, Columbus, OH.
- Bronzino, J. D. 1986. *Biomedical Engineering and Instrumentation: Basic Concepts and Applications*. PWS Engineering, Boston, MA.
- Carr, J. J. and Brown, J. M. 1993. *Introduction to Biomedical Equipment Technology*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.
- Geddes, L. A. and Baker, L. A. 1989. *Principles of Applied Biomedical Instrumentation*, 3rd ed.

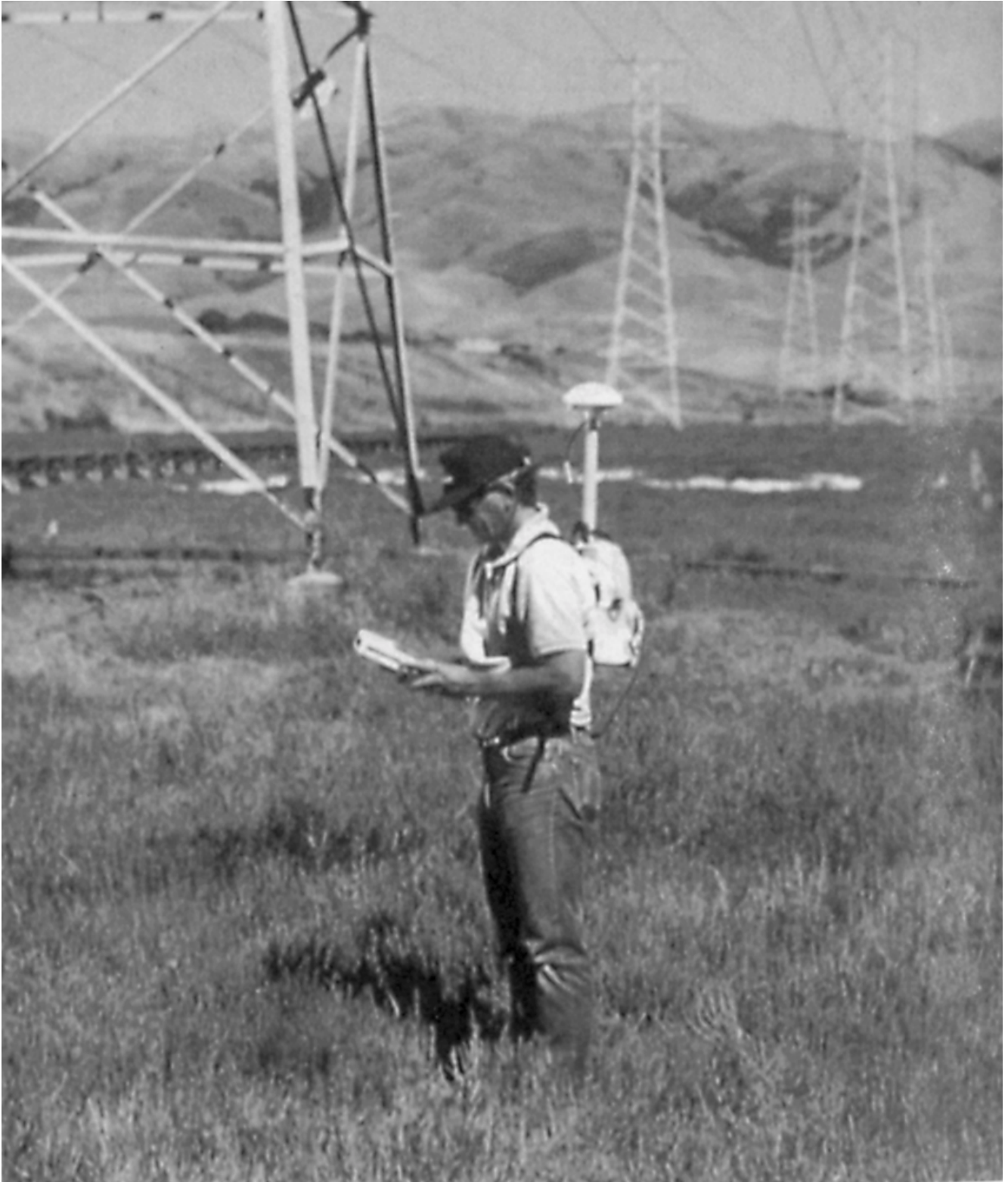
- John Wiley & Sons, New York.
- IEC. 1982. *Regulations for Electro-Medical Devices*. IEC-601. International Electrical Commission, Beuth Verlag GmbH, Berlin and Cologne, Germany.
- NFPA. 1990. *Standard for Health Care Facilities*. NFPA-99. National Fire Protection Association, Publication Sales Department, Batterymarch Park, Quincy, MA 02269. Phone (800)344-3555.
- Norman, R. A. 1988. *Principles of Bioinstrumentation*. John Wiley & Sons, New York.
- Phillips, C. A. 1991. *Functional Electrical Rehabilitation*. Springer Verlag, New York.
- Profio, E. A. 1993. *Biomedical Engineering*. John Wiley & Sons, New York.
- Webster, J. G. (Ed.) 1992. *Medical Instrumentation: Application and Design*, 2nd ed. Houghton Mifflin, Boston, MA.

## **Further Information**

- The monthly journal *IEEE Transactions on Biomedical Engineering* publishes research articles on recent advances in biomedical instrumentation. For subscription contact: IEEE Service Center, 445 Hoes Lane, P.O. 1331, Piscataway, NJ 08855-1331. Phone (800)678-IEEE.
- The monthly journal *Annals of Biomedical Engineering* also publishes research articles on recent advances in biomedical instrumentation. For subscription contact: Biomedical Engineering Society, P.O. Box 2399, Culver City, CA 90230. Phone (310) 618-9322
- ECRI evaluates medical devices, collects information about medical devices and publishes periodic reports. Contact Emergency Care Research Institute, 5200 Butler Pike, Plymouth Meeting, PA 19462-1298. Phone (215)825-6000. Fax (215)834-1276.



McCormac, J. "Surveying"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



Global positioning system (GPS) technology is revolutionizing the way people survey and collect geographic data. Originally developed as a navigation and timing system for military applications, GPS has

emerged as a leading technology for geographic information system (GIS) data collection and general mapping. GPS is quickly becoming one of the preferred ways to collect and update information for GIS, CAD-based mapping applications, automated mapping and facilities management (AM/FM) systems, and land information systems (LIS).

The Trimble® GPS Pathfinder Pro XL is a global positioning system that can compute a location in less than a second. The process called differential GPS (DGPS) provides the capability of obtaining accuracies within decimeters. Shown above, a utility crew member uses the Pathfinder Pro XL to record the position of a transmission tower. The GPS receiver automatically records the position data while the crew member records attributes about the tower. With this system, the operators could log the position and attributes of one pole every two minutes, including the travel time between poles. For additional information on GIS basics and how GPS works, see page 1598. (Photo courtesy of Trimble Navigation.)

# XXIII

## Surveying

---

**Jack McCormac**

*Clemson University*

**143 Quality Control** *B. H. W. van Gelder*

Errors • Precision • Law of Propagation of Errors • Statistical Testing • Accuracy • Reliability • Actuality

**144 Elevation** *S. D. Johnson*

Measures of Elevation and Height • Deflection of the Vertical • Vertical Datums • Elevation Measurement • Systematic Errors

**145 Distance Measurements** *R. B. Buckner*

Fundamentals of Distance Measurement • Applications and Calculations

**146 Directions** *B. A. Dewitt*

Angles • Meridians • Direction • Back Bearing and Back Azimuth • Applications

**147 Photogrammetry and Topographic Mapping** *J. Bethel*

Basic Concepts • Orientation and Model Setup • Data Collection for Topography • Data Processing for Topography • Data Presentation

**148 Surveying Computations** *B. H. W. van Gelder*

Principles of Multivariate Calculus • Principles of Linear Algebra • Model of Two Sets of Variables, Observations, and Parameters: The Mixed Model • Observations as a Function of Parameters Only: The Model of Observation Equations • All Parameters Eliminated: The Model of Condition Equations • An Example: Traversing • Dynamical Systems

**149 Satellite Surveying** *B. H. W. van Gelder*

A Satellite Orbiting the Earth • The Orbital Ellipse • Relationship between Cartesian and Keplerian Orbital Elements • Orbit of a Satellite in a Noncentral Force Field • The Global Positioning System (GPS) • Gravity Field and Related Issues

**150 Surveying Applications for Geographic Information Systems** *J. F. Thompson*

GIS Fundamentals • Monumentation or Control Surveying • Topographic Surveying • Future GIS Surveying Applications

**151 Remote Sensing** *R. W. Kiefer and T. M. Lillesand*

Electromagnetic Energy • Atmospheric Effects • Remote Sensing Systems • Remote Sensing from Earth Orbit • Digital Image Processing

DURING MOST OF THE 20TH CENTURY there has been a gradual improvement in the quality of surveying equipment. Not only have better steel tapes been manufactured, but also better and better instruments have been made for determining elevation differences and measuring angles or directions. In the last few decades there has been an amazing acceleration in the improvement of instruments.

Electronic distance measuring instruments (EDMIs) were developed with which the surveyor

can precisely measure distances of a few feet or many miles almost instantaneously. Devices called total stations were developed with which distances, directions, and elevations can be determined with equal facility and the values read digitally and recorded electronically.

In the last few years another tremendous change has occurred in the surveying field: the development of the global positioning system (GPS). With this system points on the earth's surface can be located in three dimensions with speed and accuracy by means of radio signals sent out from artificial satellites.

The chapters to follow describe the recent, almost unbelievable changes in the surveying profession and its equipment. Two topics closely related to surveying are also described: remote sensing and geographic information systems (GISs).

Since the mid-1960s the term *remote sensing* has been used to describe images of the earth's surface produced by electronic sensing devices or aerial photographs made from aircraft or satellites. Their purpose is to inventory and monitor the earth's resources. Geographic information systems are computer-based systems which enable users to store, recall, and display data relating to natural and human-made features of the earth's surface such as elevations, property boundaries, streams, forests, agricultural uses, zoning regulations, soil types, utilities, and so on.

van Gelder, B. H. W. "Quality Control"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 143

## Quality Control

---

143.1 Errors

143.2 Precision

143.3 Law of Propagation of Errors

143.4 Statistical Testing

Normalized Residual Test • Tau-test • Testing of Variances

143.5 Accuracy

143.6 Reliability

Internal Reliability • External Reliability

143.7 Actuality

**Boudewijn H. W. van Gelder**

*Purdue University*

Surveyors collect measurements that are often crucial in supporting the needs of others. They serve the public by providing information about property lines, and they serve the civil engineer by providing a proper stakeout of a subdivision or the distances between pillars of a bridge. They may provide a state department with an accurate location for a global positioning system (GPS) base station, to which the department ties all its geographic information system (GIS) surveys. A surveyor may provide the scientific community with the proper distance between two monuments on each side of a fault zone, in order to monitor movement along that fault zone, or may provide the Federal Aviation Agency with proper coordinates of GPS beacons, which aid automated landings of commercial aircraft. In addition, a surveyor provides the Department of Natural Resources with proper heights for the mapping of flood plains.

The importance of the positions, heights, or distances resulting from such measurements makes the surveyor very careful about the quality of the measurement data, let alone the quality of the results computed from those measurements.

### 143.1 Errors

---

Even the most careful surveyor, like anybody else, makes errors—errors in judgment and, not necessarily more important, measurement errors. Some errors cannot be controlled or known about at the moment of measurement collection. Some errors cannot be avoided but should be foreseen and a measurement plan designed accordingly. To distinguish among a variety of classes (and this is not a game of semantics), one speaks of the precision of a measurement, the accuracy of a measurement, the reliability of a measurement, and, last but not least, the "up-to-dateness" of a

measurement. In this respect we have to view a measurement in a very loose manner: we may want to address the accuracy of the measurement itself or the accuracy of the variable directly or indirectly related to the actual measurement. In subsequent sections, each type will be addressed.

A classification which will be addressed in passing is often found in the statistical literature [see Papoulis, (1985), among many others]:

- Stochastic or random errors
- Systematic errors, including periodic errors
- Blunders

Stochastic errors and blunders first come to mind when talking about precision. Systematic errors are often viewed as an issue of accuracy. Reliability addresses all three types of errors. There are cases where a measurement is very precise but not accurate. In other cases the measurement or related outcome may be very precise and accurate, but not reliable. In still other cases a measurement or the resulting parameter may be very precise, very accurate, and very reliable, but not up to date, so that the overall quality is still low.

## 143.2 Precision

---

Precision in the classical sense deals with the repeatability of an experiment. A surveyor measures a distance over and over again, and the outcome differs each time. These are errors of a stochastic nature, which the surveyor does not have any control over. If readings give the same outcome, the surveyor can make them a little more precise, for instance by reading the tape not to the nearest centimeter but to the nearest millimeter, to make the errors (deviations) look random. The surveyor may compute the arithmetic mean of ten distance measurements, according to

$$\hat{\mu} = \frac{\sum_{i=1}^n l_i}{n} \quad (143.1)$$

where  $\hat{\mu}$  is the estimate of the average  $\mu$  based on the sample of  $n$  observations  $l_i$ .

Based on this sample of ten distance measurements  $l_i$ , with  $i = 1, \dots, 10$  ( $n = 10$ ), the surveyor gets an estimate of the average  $\hat{l}$ . Collecting an additional ten distance measurements would undoubtedly give a different average. This is why we say that the calculated average  $\hat{l}$  is just an estimate of the average based on that sample. So, for this example, we have

$$\hat{l} = \frac{\sum_{i=1}^{10} l_i}{10} = \frac{l_1 + l_2 + l_3 + \dots + l_{10}}{10} \quad (143.2)$$

In statistics this average is known as the first moment.

A second interesting quantity to look at is the spread of the  $n$  observations. If the  $n$  observations are close together, we may call our measurement procedure precise. If the  $n$  values vary over a large range, we tend to think of them as the result of a less precise measurement scheme. To

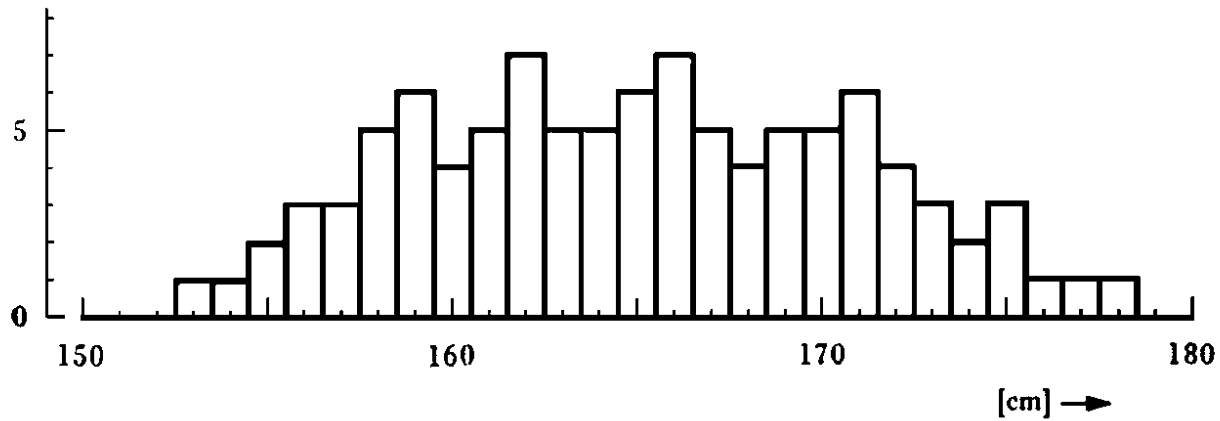


compute the spread, we compute the variations with respect to the estimated mean  $\hat{l}$ . We subtract the estimated average from each individual observation to get a residual  $v_i$ :

$$v_i = l_i - \hat{l} \quad (143.3)$$

If we counted how many observations happen to fall within a specified range, we would obtain an histogram, as in [Fig. 143.1](#).

**Figure 143.1** Histogram of  $n$  observations  $l_i$ .



If we add all the squared residuals and divide the sum by  $n - 1$ , we obtain an estimate of the standard deviation  $\hat{\sigma}$ . [We do not divide by  $n$  since we already extracted one piece of information from the sample of  $n$  observations: the estimate of the mean (and we used this estimate to compute the spread).] The sample variance  $\hat{\sigma}^2$  is the square of the sample standard deviation. We have for the sample standard deviation

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n v_i^2}{n - 1}} \quad (143.4)$$

and for the sample variance

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n v_i^2}{n - 1} \quad (143.5)$$

In statistics the variance is also known as the second moment.

### 143.3 Law of Propagation of Errors

Often one is interested not so much in the standard deviation of the individual observations (say, those ten distance measurements), but in the standard deviation of the variable  $x$  computed from

those  $n$  observations. Differentially small errors will propagate according to the rules of total and partial derivatives of multivariate calculus. An error  $dl_i$  in an observation  $l_i$  will propagate into the vector of variables  $x_j$  as an error  $dx_j$  according to

$$dx_j = \sum_{i=1}^n \frac{\partial x_j}{\partial l_i} dl_i \quad (143.6)$$

For all  $u$  variables  $x_j$  ( $j = 1, \dots, u$ ), we have

$$\begin{aligned} dx_1 &= \frac{\partial x_1}{\partial l_1} dl_1 + \dots + \frac{\partial x_1}{\partial l_n} dl_n \\ dx_2 &= \frac{\partial x_2}{\partial l_1} dl_1 + \dots + \frac{\partial x_2}{\partial l_n} dl_n \\ &\vdots \\ dx_u &= \frac{\partial x_u}{\partial l_1} dl_1 + \dots + \frac{\partial x_u}{\partial l_n} dl_n \end{aligned} \quad (143.7)$$

or, in matrix form,

$$\vec{dx} = \mathbf{J}_{ji} \vec{dl} \quad (143.8)$$

where the matrix  $\mathbf{J}_{ji}$  is the so-called Jacobian.

If we collect all standard deviations of the observations into a matrix  $\Sigma$ , the standard deviations of the variables  $x_j$  are computed from

$$\begin{aligned} \Sigma_x &= \vec{dx} \vec{dx}^T \\ &= \mathbf{J}_{ji} \vec{dl} (\mathbf{J}_{ji} \vec{dl})^T \\ &= \mathbf{J}_{ji} \vec{dl} \vec{dl}^T \mathbf{J}_{ij} \\ &= \mathbf{J} \Sigma_l \mathbf{J}^T \end{aligned} \quad (143.9)$$

where  $\mathbf{J}^T$  is the transpose of the Jacobian in Eq. (143.8).

$\Sigma$  represents the variance/covariance matrix of the variables. Consequently, for the observations  $l_i$ , the matrix  $\Sigma_l$  is

$$\Sigma_l = \begin{bmatrix} \sigma_{l_1 l_1} & \sigma_{l_1 l_2} & \cdots & \sigma_{l_1 l_n} \\ \sigma_{l_2 l_1} & \sigma_{l_2 l_2} & \cdots & \sigma_{l_2 l_n} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{l_n l_1} & \sigma_{l_n l_2} & \cdots & \sigma_{l_n l_n} \end{bmatrix} \quad (143.10)$$

In Eq. (143.10)  $\sigma_{l_i l_i}$  is the variance, as in Eq. (143.5). So we have

$$\sigma_{l_i l_i} = \sigma_{l_i}^2 \quad (143.11)$$

In Eq. (143.11)  $\sigma_{l_i l_j}$  is the covariance. From the covariance, the so-called correlation coefficient  $\rho_{l_i l_j}$  can be computed:

$$\rho_{l_i l_j} = \frac{\sigma_{l_i l_j}}{\sqrt{\sigma_{l_i}^2 \sigma_{l_j}^2}} \quad (143.12)$$

The values of the correlation coefficient will range between  $-1$  and  $1$ . Values close to  $1$  indicate high positive correlation, or that an error in quantity  $l_i$  will cause an error of equal sign and size in quantity  $l_j$ . Likewise, values close to  $-1$  indicate that an error in  $l_i$  will cause an error with opposite sign in quantity  $l_j$ . In general, we have

$$-1 \leq \rho_{l_i l_j} \leq +1 \quad (143.13)$$

In surveying we strive for uncorrelated observations, meaning that  $\sigma_{l_i l_j}$  is equal to zero. This results in a diagonal variance/covariance matrix  $\Sigma_l$  for the observations.

**Example: Standard Deviation of the Sample Mean.** To compute the standard deviation of the sample mean  $\hat{\mu}$ , we have to apply the law of propagation of errors. For the error in the sample mean, we have from Eq. (143.1)

$$d\hat{\mu} = \sum_{i=1}^n \frac{\partial \hat{\mu}}{\partial x_i} dx_i = \frac{1}{n} \sum_{i=1}^n dx_i \quad (143.14)$$

This leads to the variance of the sample mean

$$\hat{\sigma}_{\hat{\mu}}^2 = n \frac{1}{n^2} \hat{\sigma}_{x_i}^2 = \frac{1}{n} \hat{\sigma}_{x_i}^2 \quad (143.15)$$

or, for the standard deviation of the sample mean,

$$\hat{\sigma}_{\hat{\mu}} = \sqrt{\frac{\sum_{i=1}^n v_i^2}{n(n-1)}} \quad (143.16)$$

## 143.4 Statistical Testing

---

We can test the individual observation against an arbitrary number or against the sample mean. Since the possible outcomes of an experiment, including the values of their sample means and sample standard deviations, are limitless, we work with normalized observations  $l_i^n$ :

$$l_i^n = \frac{l_i - \hat{\mu}}{\hat{\sigma}} \quad (143.17)$$

The advantage of the normalized observations is that the expected outcome of the mean is equal to zero, and the standard deviation of the normalized observations is equal to 1. Tables are available in the statistical literature to test any normalized observation against any value.

The probability that an observation lies in the interval  $[a, b]$  is readily computable:

$$\begin{aligned} P(a \leq l_i \leq b) &= P\left(\frac{a - \hat{l}}{\hat{\sigma}} \leq \frac{l_i - \hat{l}}{\hat{\sigma}} \leq \frac{b - \hat{l}}{\hat{\sigma}}\right) \\ &= P\left(\frac{a - \hat{l}}{\hat{\sigma}} \leq l_i^n \leq \frac{b - \hat{l}}{\hat{\sigma}}\right) \\ &= P\left(l_i^n \leq \frac{b - \hat{l}}{\hat{\sigma}}\right) - P\left(l_i^n \leq \frac{a - \hat{l}}{\hat{\sigma}}\right) \end{aligned} \quad (143.18)$$

Assuming that the standard deviation of a distance measurement is 5 cm and the sample mean is 38.95 m, the probability that an individual observation will fall between 38.90 and 39.10 m is

$$\begin{aligned} P(38.90 \leq l_i \leq 39.05) &= P\left(\frac{38.90 - 38.95}{0.05} \leq \frac{l_i - 38.95}{0.05} \leq \frac{39.05 - 38.95}{0.05}\right) \\ &= P(-1 \leq l_i^n \leq +2) \\ &= P(l_i^n \leq +2) - P(l_i^n \leq -1) \\ &= 0.9772 - 0.1587 = 0.8185 \\ &= 81.85\% \end{aligned} \quad (143.19)$$

Similarly, the probability that the sample mean lies in an interval  $[a, b]$  is computed according to Eqs. (143.16) and (143.18):

$$\begin{aligned}
 P(a \leq \mu_i \leq b) &= P\left(\frac{a - \hat{\mu}}{\sigma_{\hat{\mu}}} \leq \frac{\mu_i - \hat{\mu}}{\sigma_{\hat{\mu}}} \leq \frac{b - \hat{\mu}}{\sigma_{\hat{\mu}}}\right) \\
 &= P\left(\frac{a - \hat{\mu}}{\sigma_{\hat{\mu}}} \leq \mu_i^n \leq \frac{b - \hat{\mu}}{\sigma_{\hat{\mu}}}\right) \\
 &= P\left(\mu_i^n \leq \frac{b - \hat{\mu}}{\sigma_{\hat{\mu}}}\right) - P\left(\mu_i^n \leq \frac{a - \hat{\mu}}{\sigma_{\hat{\mu}}}\right)
 \end{aligned} \tag{143.20}$$

Given the example of ten distance measurements, we first realize that the standard deviation of the mean is, using Eq. (143.15),

$$\sigma_{\hat{\mu}} = \frac{\sigma_{l_i}}{\sqrt{n}} = \frac{5 \text{ cm}}{\sqrt{10}} = 1.6 \text{ cm} \tag{143.21}$$

Knowing that the standard deviation of the mean of the ten distance measurements is 1.6 cm and the sample mean is 38.95 m, the probability that the mean will fall between 38.94 and 38.97 m is

$$\begin{aligned}
 P(38.94 \leq \mu_i \leq 38.97) &= P\left(\frac{38.94 - 38.95}{0.016} \leq \frac{\mu_i - 38.95}{0.016} \leq \frac{38.97 - 38.95}{0.016}\right) \\
 &= P(-0.625 \leq \mu_i^n \leq +1.250) \\
 &= P(\mu_i^n \leq +1.250) - P(\mu_i^n \leq -0.625) \\
 &= 0.8944 - 0.2659 = 0.6285 \\
 &= 62.85\%
 \end{aligned} \tag{143.22}$$

Actually, various tests have to be distinguished, depending on the type of variable investigated and whether its standard deviation is considered to be known or unknown [Hamilton, 1964]. When the standard deviation is known, we invoke a normalized test, as in the example above. For tests of variables with unknown sample standard deviation, we use the Student's  $t$ -test. Under these circumstances the mean would fall within the specified limits with a probability of

$$\begin{aligned}
P(38.94 \leq \mu_i \leq 38.97) &= P\left(\frac{38.94 - 38.95}{0.016} \leq \frac{\mu_i - 38.95}{0.016} \leq \frac{38.97 - 38.95}{0.016}\right) \\
&= P(-0.625 \leq \mu_i^t \leq +1.250) \\
&= P(\mu_i^t \leq +1.250) - P(\mu_i^t \leq -0.625) \\
&= 0.877 - 0.274 = 0.603 \\
&= 60.3\%
\end{aligned} \tag{143.23}$$

To test individual observations and errors attached to them, two tests are available: the normalized residual test, as part of the B-method of statistical testing [Baarda, 1968], and the  $\tau$ -test, as proposed by Pope [1976].

## Normalized Residual Test

The first test computes the normalized residual according to

$$w_i = -\frac{v_i}{\sigma_{v_i}} \tag{143.24}$$

At a specified significance level  $\alpha_0$ , the residuals  $v_i$  of the observations  $l_i$  are tested in the following standardized way:

$$|w_i| < \sqrt{F_{1-\alpha_0;1,\infty}} \tag{143.25}$$

with  $F$  the related critical value of the  $F$ -statistic for one degree of freedom. It should be noted that Eq. (143.24) only holds for uncorrelated observations. For the more general case of correlated observations, one is referred to Baarda [1968]. The multidimensional test on the sum of the squared residuals is

$$\frac{\hat{\sigma}_0}{\sigma_0} < \sqrt{F_{1-\alpha_0;r,\infty}} \tag{143.26}$$

where  $\sigma_0$  and  $\hat{\sigma}_0$  are the a priori and a posteriori variances of unit weight, respectively, and where

- $r$  = the number of degrees of freedom (generally the difference between the number of observations and the number of parameters to be estimated from the observations)
- $\alpha$  = the significance level of testing for  $r$  degrees of freedom

For the estimation of the mean (one parameter) from ten distance measurements, we would have  $r = n - 1 = 10 - 1 = 9$  .

## Tau-test

The  $\tau$ -test avoids the cumbersome computations of the variance/covariance matrix of the residuals [Pope, 1976] needed in the denominator in Eq. (143.24). The standard deviation of the residual  $\sigma_{v_i}$  is approximated by

$$\sigma_{v_i} \cong \sqrt{\frac{n-u}{n}} \sigma_{l_i} = \sqrt{\frac{r}{n}} \sigma_{l_i} \quad (143.27)$$

## Testing of Variances

For tests of variables of a quadratic nature, such as the variance, we use the so-called  $\chi^2$ -test. The probability that the standard deviation lies in the interval  $[a, b]$  is readily computable from

$$\begin{aligned} P(a \leq \sigma_{l_i}^2 \leq b) &= P\left(\frac{1}{a} \geq \frac{1}{\sigma_{l_i}^2} \geq \frac{1}{b}\right) \\ &= P\left(\frac{r\hat{\sigma}^2}{a} \geq \frac{r\hat{\sigma}^2}{\sigma_{l_i}^2} \geq \frac{r\hat{\sigma}^2}{b}\right) \\ &= P\left(\chi_{r;\alpha_0/2}^2 \geq \chi^2 \geq \chi_{r;1-\alpha_0/2}^2\right) \\ &= 1 - \alpha_0 \end{aligned} \quad (143.28)$$

In our example, we compute that our standard deviation falls, with a probability of 95%, in the interval

$$\left[ \sqrt{\frac{r\hat{\sigma}^2}{\chi_{r;\alpha_0/2}^2}} \leq \sigma \leq \sqrt{\frac{r\hat{\sigma}^2}{\chi_{r;1-\alpha_0/2}^2}} \right] \quad (143.29)$$

with the substituted values

$$\left[ \sqrt{\frac{9 \cdot 25}{19.02}} \leq \sigma \leq \sqrt{\frac{9 \cdot 25}{2.70}} \right] \quad (143.30)$$

or

$$[3.44 \text{ cm} \leq \sigma \leq 9.13 \text{ cm}] \quad (143.31)$$

For testing ratios of two squared estimates of the same (squared) variable we use the  $F$ -test.

Consult **Chapter 148** and the statistical literature for more details.

From statistical tables for normally distributed data (see the statistical literature for the definition of a normal distribution), it can be seen that, for instance, 95% of the observations fall in the region  $[-1.96\sigma, +1.96\sigma]$ . However, in 2.5% of the cases we may expect to have an observation which is more than  $1.96\sigma$  away from its mean. This is a good observation (it belongs to the distribution just described) but will likely be removed from the data set. If we make such an "error," we speak of an *error of the first kind*, or the so-called producer's error (e.g., a TV manufacturer erroneously rejects a good although deviating specimen). However, if this data set contained a systematic error, such an outlier may be an observation belonging to a set of observations for which the systematic error had been removed. The surveyor is, in this sense, a producer as well: a producer of survey results who does not want to erroneously reject good data. On the other hand, the erroneous acceptance of false data is an *error of the second kind*, or the consumer's error (e.g., a person purchasing a TV accepts a faulty specimen that erroneously slipped through the manufacturer's testing procedures). The client of the surveyor should be most worried about this type of error. However, the surveyor generally serving the public or any other client has to design the measurements in such a way as to minimize these types of errors, or at least the effect of errors of the second kind. The latter type of error brings us to the notions of accuracy and reliability.

## 143.5 Accuracy

---

Assume that a surveyor measures a distance very diligently (the proper tension is applied, temperature corrections are being applied, and so forth), but because of an earlier breakage the tape was repaired, resulting in an even 50 millimeters missing. All distances (assuming that all distances measured are shorter than one tape length) will have outcomes which are highly repeatable (the  $\hat{\sigma}_{\hat{\mu}}$  is very small); however, all distances will be 50 millimeters too long. The surveyor got inaccurate results despite a very precise measurement procedure. Inaccuracy is the deviation of the measurement outcomes from the "real" or "true" value of the quantity the surveyor tried to measure. (The adjectives *real* and *true* are in quotes because it can be easily argued that the true value of, say, a distance will never be known, since any measurement process will only provide us with a sample mean, never the true value.)

A dangerous but widespread practice is to increase the standard deviations of the measurements (remember, the standard deviations reflect precision or repeatability, not accuracy) by a factor of 2 to 10 in order to catch any inaccuracies caused by mismodeling, uncalibrated instrumental errors, setup errors, warming-up effects on a tripod that cause it to torque, personal errors, and so on. So we have

$$\sigma_{\text{accuracy}} = c \cdot \sigma_{\text{precision}} \quad (143.32)$$

with

$$2 \leq c \leq 10 \text{ or larger} \quad (143.33)$$

The preferred procedure is to apply any correction to the observations, such as the personal



equation of the surveyor who experiences a consistent delay in the triggering of a chronometer between the moment a star crosses a cross wire and the moment of an audible time signal. Similarly, it is better to strive for the correct atmospheric setting on a total station, rather than increasing the standard deviation of the directional/angular measurements.

Despite all the precautions, the surveyor wants, on behalf of the client, to have insight into the size of any inaccuracies that may be avoided by following certain measurement procedures as part of an overall network of measurements. This leads us to the notion of reliability.

## 143.6 Reliability

---

Having calibrated the survey equipment, having applied temperature corrections, and so on, the surveyor is able to assure himself and the client about the size of the error (inaccuracies) that can be detected with a certain probability under the given circumstances. This is often subject to a cost-versus-effect analysis.

### Internal Reliability

To be able to detect distance measurement errors of 10 cm with a probability of 80%, one has to evaluate the distance measurements as an intricate part of the overall network design. Increasing the probability from 80% to 90% will increase the cost of the survey. Similarly, the ability to detect errors smaller than 10 cm with a probability of 80% will also considerably increase the cost of the survey. This is an issue of *internal reliability*: what size measurement error is detectable with what probability? Such errors are referred to as *marginally detectable errors* (MDE). The marginally detectable error is denoted by a nabla ( $\nabla$ ).

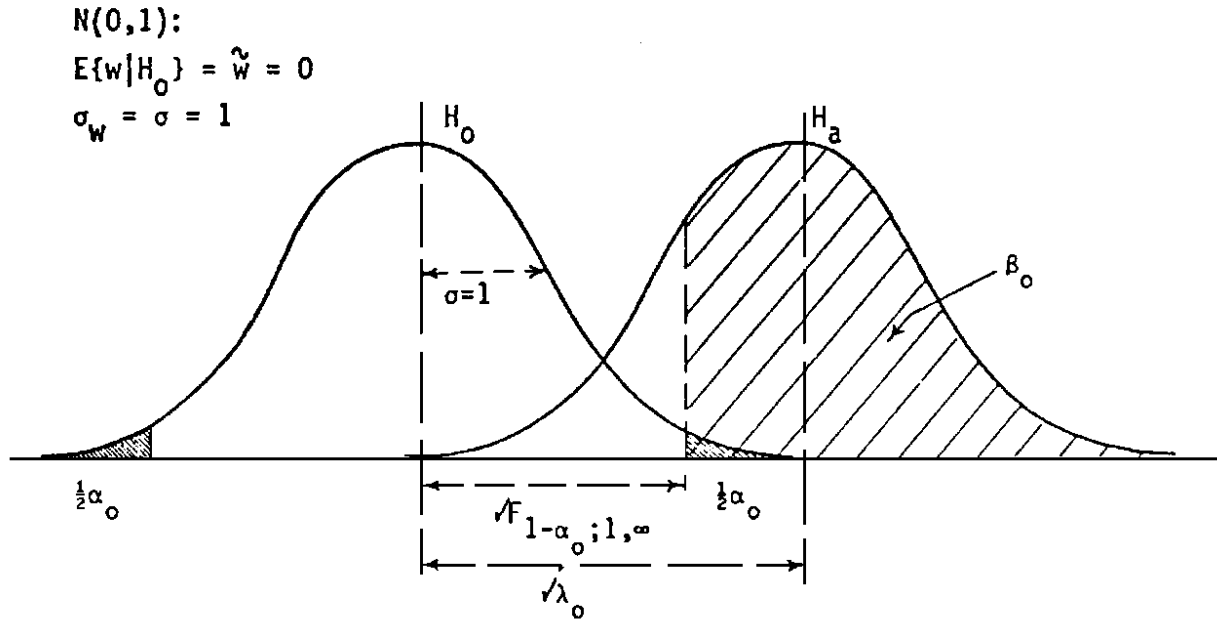
$\lambda_0$  denotes the distance between the normal distribution under the hypothesis ( $H_0$ ) that the observation belongs to this distribution, and the distribution corresponding to the alternative hypothesis ( $H_a$ ) (see Fig. 143.2). In the statistical literature one can find tables or graphs depicting  $\lambda_0$  as a function of the level of significance  $\alpha_0$  and the power of the test  $\beta_0$  [Baarda, 1968, p. 21–26]. Common values for  $\alpha_0$  and  $\beta_0$  are 5% and 80%, respectively. It can be shown that for uncorrelated observations the marginally detectable error is

$$\nabla_o l_i = \frac{\sigma_{l_i}^2}{\sigma_{v_i}} \sqrt{\lambda_0(\alpha_0, \beta_0)} \quad (143.34)$$

Following are three sets of examples for values of  $\alpha_0$  and  $\beta_0$ :

1.  $\lambda_0 = 11.70[(2.58 + 0.84)^2]$  for  $\alpha_0 = 0.01$  (1%) and  $\beta_0 = 0.80$  (80%). The alternative distribution is shifted by  $\sqrt{11.70} = 3.42$  units.
2.  $\lambda_0 = 7.84[(1.96 + 0.84)^2]$  for  $\alpha_0 = 0.05$  (5%) and  $\beta_0 = 0.80$  (80%). The alternative distribution is shifted by  $\sqrt{7.84} = 2.80$  units.
3.  $\lambda_0 = 10.50[(1.96 + 1.28)^2]$  for  $\alpha_0 = 0.05$  (5%) and  $\beta_0 = 0.90$  (90%). The alternative distribution is shifted by  $\sqrt{10.50} = 3.24$  units.

**Figure 143.2** The one-dimensional test of the normal standardized variable  $w$ , shown for hypothesis  $H_0$  and alternative hypothesis  $H_a$ .



If we apply Eq. (143.34) to the example of section 143.2 (ten uncorrelated distance measurements), we may be able to detect an error of 17.5 cm ( $= 25 \cdot 2.8/4$ ) with a probability of 80% (and level of significance 5%) if  $\sigma_l = 5$  cm and the standard deviation of the residuals is  $\sigma_v = 4$  cm.

## External Reliability

Rather than the size of the marginally detectable error, we are more interested in the effect of the MDE on the actual results computed from the measurements. Repeating Eq. (143.8), we assume that variables  $x_j$  depend on the measurements  $l_i$  according to

$$d\vec{x} = J_{ji} d\vec{l} \quad (143.35)$$

The influence of a marginally detectable error of the size of  $\nabla l_{(i)}$  on the variables  $x_j$  can be computed from

$$\nabla \vec{x} = J_{ji} \nabla \vec{l}_{(i)} \quad (143.36)$$

It should be noted that the parentheses around the subscript  $i$  denote that the vector  $\nabla l_{(i)}$  in Eqs.

(143.34) and (143.36) contains all zeros except for the  $i$ th observation for which the influence on all variables  $x_j$  is computed.

Surveyors design networks to minimize the influence of MDEs in the observations. For classical terrestrial networks this leads to designs avoiding triangles with small angles. Survey resection problems with small top angles lead to large MDEs with large negative effects on the positional accuracy for the point to be resected. In modern surveying using the global positioning system (GPS), bad DOPs (see **Chapter 149** for an explanation) lead similarly to large MDEs in the GPS observables. In these cases precise and accurate measurements may lead to highly unreliable results because of the system's incapability to flag errors of small magnitudes. Large errors in the GPS network positions may go undetected as a result [see (FGCC, 1984) and (FGCC, 1989)].

## 143.7 Actuality

---

Precise, accurate, and reliable measurements and the resulting maps may be worthless if they tend to represent locations of a network in a highly dynamic environment. Leveled heights referring to some vertical datum in an area of large annual subsidence may be worthless. Maps showing property boundaries or a GIS land use database for an area where the land use rapidly changes soon become worthless also.

The vector  $\vec{x}_j$  of results should be time-tagged and reflect the epoch  $t_0$  at which the situation is reflected:

$$\vec{x}_j = \vec{x}_j(t_0) \quad (143.37)$$

Models may be available to represent the time history of the vector  $\vec{x}_j$ . Transformation parameters are capable of expressing the current result vector  $\vec{x}_j(t)$  as a function of the result vector at reference epoch  $t_0$ . In this case, a Jacobian matrix  $\Phi$  will relate the state of  $x_j(t)$  at  $t$  to the state of  $x_j(t_0)$  at  $t_0$ :

$$\begin{aligned} d\vec{x}(t) &= \frac{\partial \vec{x}(t)}{\partial \vec{x}(t_0)} d\vec{x}(t_0) \\ &= \Phi_{t,t_0} d\vec{x}(t_0) \end{aligned} \quad (143.38)$$

with the state transition matrix being

$$\Phi(t, t_0) = \begin{pmatrix} \frac{\partial x_1(t)}{\partial x_1(t_0)} & \frac{\partial x_1(t)}{\partial x_2(t_0)} & \cdots & \frac{\partial x_1(t)}{\partial x_u(t_0)} \\ \frac{\partial x_2(t)}{\partial x_1(t_0)} & \frac{\partial x_2(t)}{\partial x_2(t_0)} & \cdots & \frac{\partial x_2(t)}{\partial x_u(t_0)} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial x_u(t)}{\partial x_1(t_0)} & \frac{\partial x_u(t)}{\partial x_2(t_0)} & \cdots & \frac{\partial x_u(t)}{\partial x_u(t_0)} \end{pmatrix} \quad (143.39)$$

Analytic expressions for the behavior of a set of coordinates as a function of time may be available. These functions may have the character of polynomials or may reflect more realistic (geo)physically or geometrically oriented time behavior.

## References

- Baarda, W. 1968. A testing procedure for use in geodetic networks. *Publ. Geodesy*. N.S. 2(5).  
Hamilton, W. C. 1964. *Statistics in Physical Science: Estimation, Hypothesis Testing, and Least Squares*. Ronald Press, New York.  
FGCC. 1984. *Standards and Specifications for Geodetic Control Networks*. Federal Geodetic Control Committee, Rockville, MD.  
FGCC. 1989. *Geometric Geodetic Accuracy Standards and Specifications for Using GPS Relative Positioning Techniques*. Version 5.0. Federal Geodetic Control Committee, Rockville, MD.  
Papoulis, A. 1985. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill Publishing Co., New York.  
Pope, A. J. 1976. *The Statistics of Residuals and the Detection of Outliers*. NOAA Technical Report NOS 65 NGS 1. National Oceanic and Atmospheric Administration, Rockville, MD.

## Further Information

### Textbooks and Reference Books

For additional reading and more background, from the very basic to the advanced level, one may consult specific chapters in a variety of textbooks on geodesy, satellite geodesy, physical geodesy, surveying, photogrammetry, or statistics itself. The reader is referred to the following textbooks (in English):

- Bjerhammer, E. A. 1973. *Theory of Errors and Generalized Matrix Inverses*. Elsevier Scientific Publishing, New York.  
Bomford, G. 1980. *Geodesy*. Clarendon Press, Oxford.  
Ch. 1: Triangulation, Traverse, and Trilateration (Field Work)  
Ch. 2: Computation of Triangulation, Traverse, and Trilateration  
App. D: Theory of Errors  
Burnside, C. D. 1991. *Electromagnetic Distance Measurement*. BSP Professional Books, Oxford,

- England.
- Ch. 6: The Measurement Process and Calibration of Instruments
- Davis, R. E., Foote, F. S., Anderson, J. M., and Mikhail, E. M. 1981. *Surveying: Theory and Practice*. McGraw-Hill Publishing Co. New York.
- Ch. 2: Survey Measurements and Adjustments
- App. B: Least-Squares Adjustment
- Escobal, P. R. 1976. *Methods of Orbit Determination*. John Wiley & Sons, New York.
- App. IV: Minimum Variance Orbital Parameter Estimation
- Heiskanen, W. A. and Moritz, H. 1967. *Physical Geodesy*. W.H. Freeman & Co., New York.
- Ch. 7: Statistical Methods in Physical Geodesy
- Hirvonen, R. A. 1965. *Adjustment by Least Squares in Geodesy and Photogrammetry*. Frederick Ungar Publishing Co., New York.
- Hofmann-Wellenhof, B., Lichtenegger, H., and Collins, J. 1995. *GPS: Theory and Practice*. Springer-Verlag, New York.
- Ch. 9: Data Processing
- Kaula, W. M. 1966. *Theory of Satellite Geodesy: Applications of Satellites to Geodesy*. Blaisdell Publishing Co., New York.
- Ch. 4: Geometry of Satellite Observations
- Ch. 5: Statistical Implications
- Ch. 6: Data Analysis
- Koch, K. R. 1988. *Parameter Estimation and Hypothesis Testing in Linear Models*. Springer-Verlag, New York.
- Kok, J. 1984. *On Data Snooping and Multiple Outlier Testing*. NOAA Technical Report NOS NGS 30. National Oceanic and Atmospheric Administration, Rockville, MD.
- Kraus, K. 1993. *Photogrammetry*. Ferd. Dümmlers Verlag, Bonn.
- App. 4.2-1: Adjustment by the Method of Least Squares
- Leick, A. 1995. *GPS: Satellite Surveying*. John Wiley & Sons, New York.
- Ch. 4: Adjustment Computations
- Ch. 5: Least-Squares Adjustment Examples
- App. B: Linearization
- App. C: One-Dimensional Distributions
- McCormac, J. C. 1995. *Surveying*. Prentice Hall, Englewood Cliffs, NJ.
- Ch. 2: Introduction to Measurements
- Mikhail, E. M. 1976. *Observations and Least Squares*. IEP-A Dun-Donnelley, New York.
- Mikhail, E. M. and Gracie, G. 1981. *Analysis and Adjustment of Survey Measurements*. Van Nostrand Reinhold, New York.
- Moffitt, F. H. and Bouchard, H. 1992. *Surveying*. HarperCollins Publishers, New York.
- Ch. 1-10: Errors and Mistakes
- Ch. 1-11: Accuracy and Precision
- Ch. 5: Random Errors
- App. A: Adjustment of Elementary Surveying Measurements by the Method of Least Squares
- App. B: The Adjustment of Instruments
- Mueller, I. I. and Ramsayer, K. H. 1979. *Introduction to Surveying*. Frederick Ungar Publishing

- Co., New York.  
 Ch. 2: Nature of Errors and Measurements  
 Ch. 5: Adjustment Computation by Least Squares
- Roberts, J. 1995. *Construction Surveying*. Delmar Publishers, New York.  
 Ch. 6-2: Establishing Dimension Control Program Standards (ISO 6643)
- Tienstra, J. M. 1966. *Theory of Adjustment of Normally Distributed Observations*. Argus Publishing Co., Amsterdam.
- Uotila, U. A. 1985. *Adjustment Computations Notes*. Department of Geodetic Science and Surveying, Ohio State University, Columbus.
- Vaníček, P. and Krakiwsky, E. J. 1982. *Geodesy: The Concepts*. North-Holland Publishing Co., Amsterdam.  
 Part I: Introduction  
 Ch. 3: Mathematics and Geodesy  
 Part III: Methodology  
 Ch. 10: Elements of Geodetic Methodology  
 Ch. 11: Classes of Mathematical Models  
 Ch. 12: Least-Squares Solution of Overdetermined Models  
 Ch. 13: Assessment of Results  
 Ch. 14: Formulation and Solving of Problems
- Wolf, P. R. 1983. *Elements of Photogrammetry*. McGraw-Hill Publishing Co., New York.  
 App. A: Random Errors and Least Squares Adjustment
- Wolf, P. R. 1987. *Adjustment Computations: Practical Least Squares for Surveyors*. Landmark Enterprises, Rancho Cordova, CA.
- Wolf, P. R., and Brinker, R. C. 1994. *Elementary Surveying*. HarperCollins College Publishers, New York.  
 Ch. 2: Theory of Measurements and Errors  
 App. A: Instrument Testing and Adjusting  
 App. C: Propagation of Random Errors and Least-Squares Adjustment

## Journals and Organizations

The latest results from research of statistical applications in geodesy, surveying, mapping, and photogrammetry are published in a variety of journals.

The following are two international magazines under the auspices of the International Association of Geodesy, both published by Springer-Verlag (Berlin/Heidelberg/New York):

*Bulletin Géodésique*

*Manuscripta Geodetica*

Geodesy- and geophysics-related articles can be found in:

American Geophysical Union, Washington, DC: *EOS* and *Journal of Geophysical Research*

Royal Astronomical Society, London: *Geophysical Journal International*

Statistical articles related to kinematic GPS can be found in:

Institute of Navigation: *Navigation*

Many national mapping organizations publish journals in which recent statistical applications in

geodesy/surveying/mapping/photogrammetry are documented:

American Congress of Surveying and Mapping:

*Surveying and Land Information Systems*

Cartography and Geographic Information Systems

American Society of Photogrammetry and Remote Sensing: *Photogrammetric Engineering & Remote Sensing*

American Society of Civil Engineers: *Journal of Surveying Engineering*

Deutscher Verein für Vermessungswesen: *Zeitschrift für Vermessungswesen*, Konrad Wittwer Verlag, Stuttgart

The Canadian Institute of Geomatics: *Geomatica*

The Royal Society of Chartered Surveyors: *Survey Review*

Institute of Surveyors of Australia: *Australian Surveyor*

Worth special mention are the following trade magazines:

*GPS World*, published by Advanstar Communications, Eugene, OR

*P.O.B. (Point of Beginning)*, published by P.O.B. Publishing Company, Canton, MI

*Professional Surveyor*, published by American Surveyors Publishing Company, Arlington, VA

*Geodetical Info Magazine*, published by Geodetical Information & Trading Centre bv., Lemmer, the Netherlands

National mapping organizations such as the U.S. National Geodetic Survey (NGS) regularly make software available (free and at cost). Information can be obtained from:

National Geodetic Survey

Geodetic Services Branch

National Ocean Service, NOAA

1315 East-West Highway, Station 8620

Silver Spring, MD 20910-3282

Johnson, S. D. "Elevation"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



# 144

## Elevation

---

### 144.1 Measures of Elevation and Height

### 144.2 Deflection of the Vertical

### 144.3 Vertical Datums

### 144.4 Elevation Measurement

Ordinary Differential Leveling • Precise Leveling • Trigonometric Leveling • Instruments

### 144.5 Systematic Errors

Earth Curvature • Atmospheric Refraction • Instrument Adjust-ment • Orthometric Correction

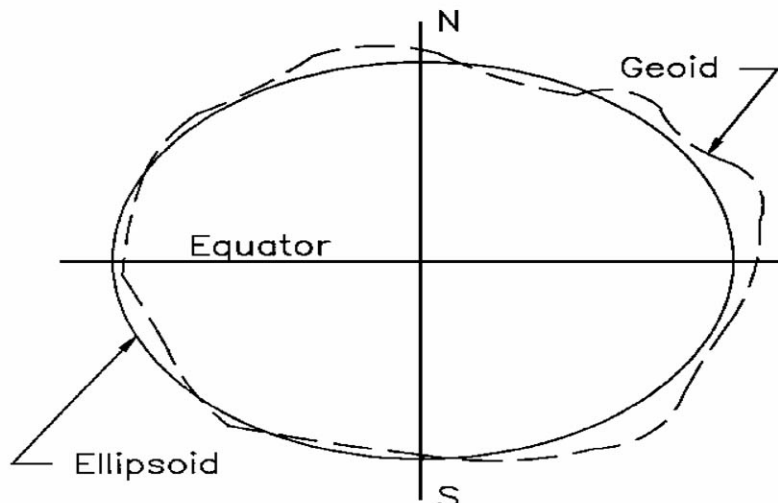
## Steven D. Johnson

*Purdue University*

**Elevation** is the distance above or below a specified reference surface. In engineering and surveying the surface most commonly used to reference elevation is the geoid. The geoid is defined as an equipotential surface that closely approximates **mean sea level**. However, the geoid and mean sea level surfaces are not coincident. They may be separated by a meter or more at any specific location.

By definition, the potential due to gravity is equal at all points on the geoid, and the force of gravity is perpendicular to the geoid at all points. The geoid is an undulating, irregular surface that is affected by density variations within the earth. The geoid is not readily defined by mathematical equations. The mathematical ellipsoid used as a **datum** surface for geodetic position can approximate the geoid, but the ellipsoid and geoid will be separated by up to 100 meters or more for a mean global fit. [Figure 144.1](#) illustrates the general relationship between the geoid and some approximating geodetic ellipsoid.

**Figure 144.1** Relationship between geoid and approximating geodetic ellipsoid.



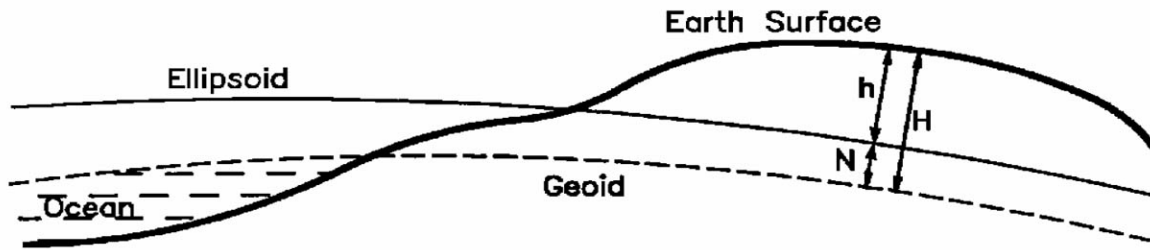
## 144.1 Measures of Elevation and Height

The vertical distance referenced to the geoid is called an orthometric **height** (elevation),  $H$ . Orthometric height is measured along the plumb line. A height referenced to the ellipsoid is called an *ellipsoidal height*,  $h$ . Ellipsoidal height is measured along the normal to the ellipsoid. Geoid height,  $N$ , is the distance between the geoid and the ellipsoid measured along the normal to the ellipsoid. Neglecting the deviation between the plumb line and the ellipsoidal normal, the geoid height is related to the orthometric and ellipsoidal heights by the equation

$$h = H + N$$

Thus, Fig. 144.2 shows a negative geoid height, which is typical in the U.S.

**Figure 144.2** Illustration of negative geoid height.



Equipotential surfaces (level surfaces) are not parallel to another. The distance between the surfaces decreases toward the earth's north pole. Each level surface has a different gravity potential relative to the geoid that can be expressed as a geopotential number measured in terms of geopotential units: 1 geopotential unit (gpu) = 1000 gal-meters. As an expression of height, the geopotential number (a potential) is constant for a given level surface, whereas the orthometric height decreases for a given level surface proceeding toward the pole. The change in orthometric height can be calculated using the formula found at the end of this chapter. Geopotential numbers can be converted to distance units of height. One possible type of height is defined by dynamic height,  $H_d$ , given as

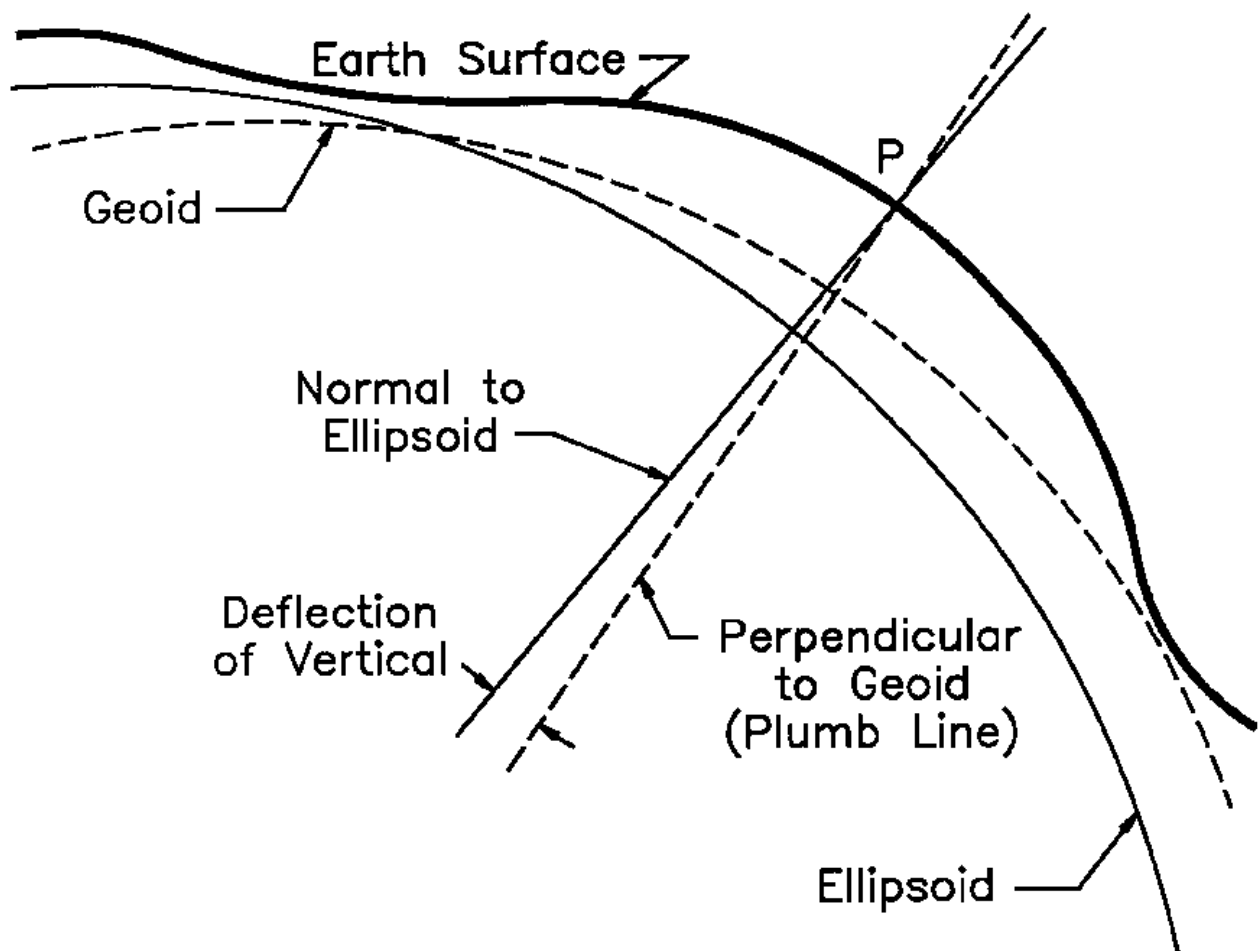
$$H_d = \frac{\text{GPN}}{g_n}$$

where GPN is the geopotential number of the point and  $g_n$  is the value of gravity calculated on the ellipsoid at  $45^\circ$  latitude using a standard gravity formula.

## 144.2 Deflection of the Vertical

A **vertical** or plumb line is perpendicular to the geoid and therefore parallel to the direction of gravity. A normal line is perpendicular to the ellipsoid. Because the geoid and ellipsoid are generally not parallel, a vertical line and a normal line are not coincident for most points on the earth's surface. The deflection of the vertical is the angle between these two lines at a point on the earth's surface. Figure 144.3 illustrates the definition of the deflection of the vertical. The deflection of the vertical is generally resolved into two components:  $\xi$  in the N-S direction and  $\eta$  in the E-W direction.

**Figure 144.3** Definition of the deflection of the vertical.



The geoid has been modeled regionally for the U.S. by the National Geodetic Survey (NGS). The model is known as GEOID93, and it is distributed as a computer program by NGS. The program estimates geoid height differences with a reported accuracy of 10 cm (one standard deviation) over lengths of approximately 100 km.

### 144.3 Vertical Datums

Vertical datums are defined for the purpose of specifying the elevation of points with respect to the geoid. There are two relevant continental vertical datums within the U.S.: the National Geodetic Vertical Datum of 1929 (NGVD 29) and the North American Vertical Datum of 1988 (NAVD 88).

The National Geodetic Vertical Datum of 1929 (NGVD 29) is a long-used reference for mean sea level in the U.S. NGVD 29 was the product of a 1929 general adjustment of the U.S. and Canadian vertical control networks. The 1929 adjustment was based, in part, upon the assumption

that the local mean sea level at the tide stations used in the adjustment was equal (same equipotential surface). This is not a valid assumption since the elevation of mean sea level varies from the Atlantic coast to the Pacific coast of the U.S. This distortion of the vertical datum caused the official name of the 1929 datum to change from "Sea Level Datum of 1929" to "National Geodetic Vertical Datum of 1929" in 1976. Other distortions, including those from upheaval and subsidence of the earth's crust, are present in the NGVD 29; however, it remains a datum of reference for the U.S.

A new adjustment of the vertical datum for North America has been completed recently. This project, known as the North American Vertical Datum of 1988 (NAVD 88), is a least-squares readjustment of over 600 000 **benchmarks** across the North American continent, resulting in a better approximation of the geoid. The project includes re-leveling of approximately 83 000 km of first-order vertical control within the U.S. Leveling data from Canada, Mexico, and Central America is included in the adjustment. The result of the NAVD 88 adjustment is a computer database of vertical control stations and elevations across the U.S. This improved model of the geoid is beneficial for the determination of orthometric heights using Global Positioning System (GPS)–derived heights above the ellipsoid.

The change in elevations from NGVD 29 to NAVD 88 varies, depending upon the area of the country involved. The relative elevation between existing benchmarks will change only by a few millimeters. The absolute elevation of benchmarks may change by as much as a few decimeters. An elevation correction constant between the two datums will suffice for most project areas.

## 144.4 Elevation Measurement

---

Elevation may be measured by several methods. Some of these methods measure elevation directly, for example, GPS satellite ranging, photogrammetric aerotriangulation, inertial surveying methods, and barometric altimetry. Some of these methods measure the difference in elevation from a reference benchmark to the point to be determined, for example, ordinary and precise differential leveling and trigonometric leveling. It should be noted that the direct methods typically must also be referenced to benchmarks so that translations and rotations can be performed to establish the proper relationship to the elevation datum. Thus the difference in elevation is, fundamentally, the important value to be measured.

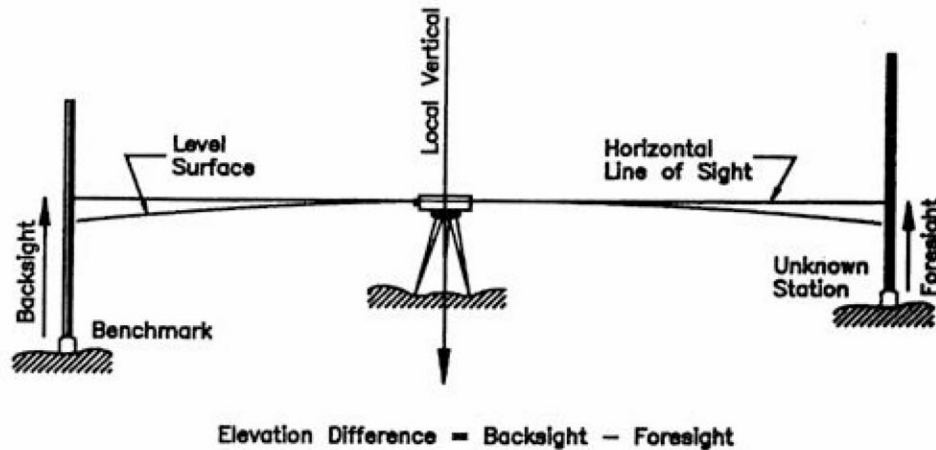
### Ordinary Differential Leveling

Differential leveling is a very simple process based on the measurement of vertical distances from a horizontal line. Elevations are transferred from one point to another through the process of using a leveling instrument to read a rod held vertically on, first, a point of known elevation and, then, on the point of unknown elevation.

A single-level setup is illustrated in [Fig. 144.4](#). The known elevation of the backsight point is transferred vertically to the line of sight by adding the known elevation and the backsight rod reading. The elevation of the line of sight is the height of instrument, HI. By definition, the line of sight generates a horizontal plane at the instrument location when the telescope is rotated on the

vertical axis. The line-of-sight elevation is transferred down to the unknown elevation point by turning the telescope to the foresight, subtracting the rod reading from the height of instrument. Note that the difference in elevation from the backsight station to the foresight station is determined by subtracting the foresight rod reading from the backsight rod reading.

**Figure 144.4** Using differential leveling to measure elevation.



A level route consists of several level setups, each one carrying the elevation forward to the next foresight using the differential leveling method. A level route is typically checked by closing on a second known benchmark or by looping back to the starting benchmark. At the closing benchmark, closure = computed elevation – known elevation. Since differential leveling is usually performed with approximately equal setup distances between turning points, the level route is adjusted by distributing the closure equally to each setup:

$$\text{Adjustment} = \left( -\frac{\text{Closure}}{n} \right) \text{ per setup}$$

where  $n$  is the number of setups in the route. When a network of interconnected routes is surveyed, a least-squares adjustment is warranted.

## Precise Leveling

Precise leveling methods are used when the highest accuracy is required for engineering and surveying work. Instruments used in precise leveling are specifically designed to obtain a high degree of accuracy in leveling. Improved optics in the telescope, improved level sensitivity, and carefully calibrated rod scales are all incorporated into the differential leveling process.

Typically when performing precise leveling, a method of leveling called *three-wire leveling* is used. This involves reading the center cross hair as well as the upper and lower "stadia" cross hairs.

The basic process of leveling is the same as ordinary differential leveling, except that the three cross hair readings are averaged to improve the precision of each backsight and foresight value.

Another method that can be used to improve the precision of the level rod reading is to use an optical micrometer on the telescope. The optical micrometer is a rotating parallel-plate prism attached in front of the objective lens of the level. The prism enables the observer to displace the line of sight parallel with itself and set the horizontal cross hair exactly on the nearest rod graduation. The observer adds the middle cross hair rod reading and the displacement reading on the micrometer to obtain a precise rod reading to the nearest 0.1 millimeter.

## Trigonometric Leveling

Trigonometric leveling is a method usually applied when a total station is used to measure the slope distance,  $S$ , and the vertical angle,  $\alpha$ , to a point. Assuming the total station is set up on a station of known elevation and the height of instrument, HI, and reflector, HR, are measured, the elevation of the unknown station is

$$V = S \sin \alpha$$
$$P_{\text{Elev}} = A_{\text{Elev}} + \text{HI} + V - \text{HR}$$

The precision of trigonometric elevations is determined by the uncertainty in the vertical angle measurement and the uncertainty caused by atmospheric refraction effects. For long lines the effects of earth curvature and atmospheric refraction must be included.

## Instruments

### Altimeters

Surveying altimeters are precise aneroid barometers that are graduated in feet or meters. As the altimeter is raised in elevation, the barometer senses the atmospheric pressure drop. The elevation is read directly on the face of the instrument. Although the surveying altimeter may be considered to measure elevation directly, best results are obtained if a difference in elevation is observed by subtracting readings between a base altimeter kept at a point of known elevation and a roving altimeter read at unknown points in the area to be surveyed. The difference in altimeter readings is a better estimate of the difference in elevation, since the effects of local weather changes, temperature, and humidity that affect altimeter readings are canceled in the subtraction process. By limiting the distance between base and roving altimeters, accuracies of 3 to 5 feet are possible. Other survey configurations utilizing low and high base stations or leap-frogging roving altimeters can yield good results over large areas.

### Level Bubble Instruments

Level bubble instruments, or spirit levels, contain a level vial with a bubble that must be centered to define a horizontal plane. Field instruments consist of three main components: a telescope to

define a line of sight and magnify the object sighted, a level vial attached to the telescope to define the orientation of the instrument with respect to gravity, and a leveling head to tilt and orient the instrument.

All level bubble instruments are designed around the same fundamental relationships. These relationships are as follows:

- The axis of the level bubble (or compensator) should be perpendicular to the vertical axis.
- The line of sight should be parallel to the axis of the level bubble.

When these instrument adjustment relationships are true and the instrument is properly set up, the line of sight will sweep out a horizontal plane that is perpendicular to gravity at the instrument location.

Instruments that use a level bubble to orient the axes to the direction of gravity depend on the bubble's sensitivity for accuracy. Level bubble sensitivity is defined as the central angle subtended by an arc of one division on the bubble tube. The smaller the angle subtended is, the more sensitive the bubble is to dislevelment. A bubble division is typically 2 millimeters long, and bubble sensitivity typically ranges from 60 seconds to 1 second.

### **Builder's Level**

The builder's level typically is less precise than other instruments in this category, but it is one of the most inexpensive and versatile instruments that is used by field engineers for construction layout. In addition to being able to perform leveling operations, it can be used to turn angles, and the scope can be tilted for inclined sights.

### **Transit**

Although the primary functions of the transit are for angle measurement and layout, it can also be used for leveling because it has a bubble attached to the telescope. However, the field engineer should be aware that the transit may not be as sensitive and stable as a quality level.

### **Dumpy Level**

The engineer's dumpy-type level has been the workhorse of leveling instruments for more than 150 years. Even with advancements in other leveling instruments, such as the automatic level and the laser, the dumpy may still be the instrument of choice in a construction environment because of its stability.

### **Automatic Compensator Levels**

Compensators were developed about 50 years ago and incorporated into field levels. The compensator is a free-swinging pendulum arrangement in the optical path that maintains a fixed relationship between the line of sight and the direction of gravity. If the instrument is in adjustment, the line of sight will be maintained as a horizontal line. Compensator instruments are extremely fast to set up and level.

## **Laser Levels**

A laser level uses a laser beam directed at a spinning optical reflector. The reflector is oriented so that the rotating laser beam sweeps out a horizontal reference plane. The level rod is equipped with a sensor to detect the rotating beam. By sliding the detector on the rod, a vertical reading can be obtained at the rod point. Laser levels are especially useful on construction sites. The spinning optics can also be oriented to produce a vertical reference plane.

## **Digital Levels**

Digital levels are electronic levels that can be used to more quickly obtain a rod reading and make the reading process more reliable. The length scale on the level rod is replaced by a bar code. The digital level senses the bar code pattern and compares it to a copy of the code held in its internal memory. By matching the bar code pattern, a rod reading length can be obtained. Digital levels are available for ordinary and precise leveling applications.

## **Level Rods**

In addition to the chosen leveling instrument, a level rod is required to be able to transfer elevations from one point to another. The level rod is a graduated length scale affixed to a rod and held vertically on a turning point or benchmark. The scale is read to obtain the vertical distance from the point to the line of sight.

Level rods are graduated in feet, inches, and fractions; feet, tenths, and hundredths; or meters and centimeters. Rods used in ordinary leveling may be multipiece extendable rods with graduations marked directly on the rod material or on a metal strip affixed to the rod for support. Rods used in precise leveling are one-piece rods with a stable invar metal graduated scale supported under constant tension by the rod. A precise rod can be calibrated for changes in length caused by temperature.

Accurate field leveling work is also aided by the use of rod targets, rod levels, and stable turning point pins when required.

# **144.5 Systematic Errors**

---

## **Earth Curvature**

The curved shape of the earth results in the equipotential surface through the telescope departing from the horizontal plane through the telescope as the line of sight proceeds to the horizon. This effect makes actual level rod readings too large by the following approximate relation,

$$C = 0.0239D^2$$

where  $D$  is the sight distance in thousands of feet.

## **Atmospheric Refraction**

The atmosphere refracts the horizontal line of sight downward, making the level rod reading



smaller. The typical effect of refraction is equal to about 14% of the effect of earth curvature. Thus, the combined effect of curvature and refraction is approximately

$$(C - r) = 0.0206D^2$$

## Instrument Adjustment

If the geometric relationships defined in the preceding discussion are not correct in the leveling instrument, the line of sight will slope upward or downward with respect to the horizontal plane through the telescope. The test of the line of sight of the level to ensure that it is horizontal is called the "two-peg test." If the line of sight is inclined, the difference in elevation obtained from the two setups will not be equal. Either the instrument must be adjusted, or the slope of the line of sight must be calculated. The slope is expressed as a collimation factor,  $C$ , in terms of rod reading correction per unit of sight distance. It may be applied to each sight by the following:

$$\text{Corrected rod reading} = \text{Rod reading} + (C_{\text{Factor}} \cdot D_{\text{Sight}})$$

In ordinary differential leveling, these effects are canceled in the field procedure by always setting up so that the backsight distance and foresight distance are equal. The errors are canceled in the subtraction process. If long unequal sight distances are used, the rod readings should be corrected for curvature and refraction and for collimation error.

## Orthometric Correction

When long, precise level routes are surveyed, it is necessary to account for the fact that the equipotential surfaces converge as the survey proceeds north. The correction to be applied for convergence of equipotential surfaces at different elevations can be calculated by

$$\text{Correction} = -0.0053 \sin 2\phi H \Delta\phi_{\text{rad}}$$

where  $\phi$  is the latitude at the beginning point,  $H$  is the elevation at the beginning point, and  $\Delta\phi$  is the change in latitude from the southerly station to the northerly station expressed in radians.

## Defining Terms

**Benchmark (BM):** A benchmark is a permanent object having a mark of known elevation, for example, a cross chiseled on a boulder or a concrete monument with an embedded brass disk.

**Datum:** Any quantity or set of such quantities that may serve as a reference or basis for calculation of other quantities.

**Datum sea level:** An equipotential surface passing through a specified point at mean sea level that is used as a reference for elevations; a surface passing through mean sea level at certain specified points to which elevations determined by leveling are referred. Note that, in general, the latter surface is not an equipotential surface.

**Elevation:** The distance, measured along the direction of gravity (plumb line), between a point and a reference equipotential surface, usually the geoid.

**Height:** The distance, measured along a perpendicular, between a point and a reference surface, for example, the ellipsoidal height; the distance, measured along the direction of gravity, between a point and a reference surface of constant geopotential, for example, the orthometric height. Note that the term *elevation* is preferred when the geoid is used as the reference surface.

**Mean sea level:** The arithmetic mean of elevations (heights) of the water's surface observed hourly over a specific 19-year cycle.

**Vertical:** The direction in which gravity acts.

## References

Bomford, G. 1980. *Geodesy*, 4th ed. Clarendon Press, Oxford, England.

Leick, A. 1990. *GPS Satellite Surveying*. John Wiley & Sons, New York.

Moffitt, F. H. and Bouchard, H. 1992. *Surveying*, 9th ed. HarperCollins, New York.

NGS. 1984. *Standards and Specifications for Geodetic Control Networks*, Federal Geodetic Control Committee, National Geodetic Survey, U.S. Department of Commerce, Silver Spring, MD.

NGS. 1986. *Geodetic Glossary*, National Geodetic Survey, U.S. Department of Commerce, Silver Spring, MD.

Schwarz, C. R. (Ed.). 1989. *North American Datum of 1983*. NOAA Professional Paper NOS 2, January.

Torge, W. 1991. *Geodesy*, 2nd ed. de Gruyter, Berlin.

Vaniček, P. and Krakiwsky, E. J. 1986. *Geodesy: The Concepts*, 2nd ed. Elsevier Science, New York.

Wolf, P. R. and Brinker, R. C. 1994. *Elementary Surveying*, 9th ed. HarperCollins College, New York.

## Further Information

The material in this chapter is intended only as an overview of elevation reference systems and basic surveying methods. There are many textbooks dedicated completely to the various aspects of surveying. For a more complete presentation of surveying theory, consult *Surveying*, 9th edition, by Moffitt and Bouchard, HarperCollins Publishing, 1992, or *Elementary Surveying*, 9th edition, by Wolf and Brinker, HarperCollins College Publishers, 1994. For more detailed information on the capabilities of various instruments and software, Business News Publishing prepares the trade magazine *P.O.B.* (Business News Publishing Company, 755 W. Big Beaver Rd., Suite 1000, Troy, MI 48084), and American Surveyors Publishing Company prepares the trade magazine *Professional Surveyor* (American Surveyors Publishing Company, Inc., Suite 501, 2300 Ninth

Street South, Arlington, VA 22204). Each of these publications conducts annual reviews of surveying instruments and software. These listings allow the reader to keep up to date and compare "apples to apples" when analyzing equipment.

Survey control information, software, and many useful technical publications are available from the National Geodetic Survey (NGS). The address is

National Geodetic Survey Division  
National Geodetic Information Branch, N/CG17  
1315 East-West Highway, Room 9218  
Silver Spring, MD 20910-3282

Buckner, R. B. "Distance Measurements"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Distance Measurements

---

### 145.1 Fundamentals of Distance Measurement

The Need for Corrections • Taped Measurements • Tacheometric Measurements • Electronic Distance Measurements • Indirect Computational Methods • Approximate Methods

### 145.2 Applications and Calculations

Taping • Stadia • Subtense Bar • Electronic Distance Measurements

#### **R. Ben Buckner**

*Surveying Education Consultant*

When the word **distance** is used in surveying without qualification, it is usually construed to mean *horizontal distance*<sup>3/4</sup>that is, a distance either measured directly along a horizontal line or measured along a slope and then mathematically projected to the horizontal. If linear measurements along other alignments are being discussed, they are given specific designations, such as *slope distance* or *vertical distance* (elevation difference).

Since ancient times, direct distance measurements have been made using lines, cords, ropes, rods, chains, tapes, and other such devices. Distances have also been measured indirectly, employing stadia and other tacheometric methods, as well as trigonometrically, using a combination of distance and angle measurements and calculations. Technological advances during the second half of the 20th century have created changes in the methods used to measure many distances. During the 1960s and 1970s most surveyors made a gradual shift from the chain and tape to electronic distance systems that measure distances with light waves or microwaves. During the 1980s this trend continued into using *total station* systems, which measure both distances and angles electronically and process the combined measurements into a Cartesian coordinate form for the survey points. Thus, many distances are determined by computations from the coordinate positions rather than by direct measurements. Most recently, there has been a trend toward using the global positioning system (GPS) for precise determination of positions of points, and these positions are then used to indirectly determine the distances between survey points. Eventually, GPS may replace both the tape and electronic distance instruments (including the total station) for most surveying measurements, including angles, elevations, and distances. GPS is covered in **Chapter 149**.

An overview of the concepts of direct and indirect distance measurements will be given here. The geometric, electronic, and other physical principles of the measurement systems will not be discussed—only the basic measuring procedures and the appropriate corrections needed to achieve optimum accuracy.

## 145.1 Fundamentals of Distance Measurement

---

### The Need for Corrections

Measurements, whether of distance or any other quantity, are but estimates of the magnitude of the quantity. Few initial readings (observations) in surveying contain the accuracy desired without some correction for systematic errors. Much of the consideration regarding measurements is not so much of how to operate the instruments and perform the calculations to reduce the data, but of how to identify and evaluate the magnitude of the corrections that must be applied to the readings. The "true value," theoretically, is always equal to the reading plus the corrections. That is,  $T = R + \sum C$ . If the true value is known and the reading is needed to achieve this,  $R = T - \sum C$ . Application of this basic concept will be illustrated subsequently. The following sections discuss the typical methods of determining survey distances.

### Taped Measurements

Most **taped** measurements are made using a steel tape, usually of 100 ft length. The tape is laid along a line between points and stretched tight. If the distance desired is longer than the length of the tape, more than one tape length is employed, the ends of the tape being marked with taping pins (sometimes called "arrows"). Alignment of the ends of the tape, to keep the tape on a straight line, is achieved either by hand signals from a rear tapeman or by a transit or theodolite centered over one of the two ends of the line.

The tape is either (1) allowed to rest on the ground surface ("fully supported"), or (2) suspended throughout its length ("end supported"), with the distance mark on the tape being transferred to the ground using plumb bobs at one or both ends. In the former case corrections for ground slope must be made, the slope having been measured using other instruments, such as hand levels, clinometers, or theodolites. In the latter case the two ends of the tape are usually held near the same elevation as determined by hand levels or clinometers. End-supported taping requires corrections for the sag of the tape, but not the slope. In either type of taping, corrections must usually be made (depending on the accuracy desired) for calibrated length, changes from calibration temperature to field temperature, and tension (if different from that used when calibrated).

Steel tapes are made in various lengths other than 100 feet, for example, 25 meters, 50 meters, 200 feet, and so forth. The principles in the use and correction of readings is much the same for various tapes, regardless of length.

Tapes are also made of other materials, including woven fiber and invar (a nickel-steel alloy). The fiber ("cloth") tapes are used only for rough measurements, and there is usually no attempt to apply corrections to the readings. In contrast, invar tapes are used for precise surveys, such as the establishment of calibration base lines for electronic distance measurements, and all applicable corrections must be made. The use of these more precise tapes and the less precise "cloth" tapes will not be covered here.

## Tacheometric Measurements

The most common method in the **tacheometry** category uses the stadia. Such measurements are made through the telescope of a transit, theodolite, or level instrument. The measurement is achieved by observing where the two horizontal stadia hairs, viewed through the telescope, strike a graduated rod held vertically on a survey point. The difference between the two readings is a function of the separation of the stadia hairs, the focal length of the telescope, and the distance between the instrument and the rod. When the line of sight is other than horizontal, the vertical or zenith angle must also be considered. Stadia distances are generally accurate to only about one or two feet for normal sight distances, and thus there is little need to be concerned about corrections to the data.

The subtense bar is another tacheometric instrument. The principle employs the measurement of a precise horizontal angle between two distinct targets at the two ends of the bar. The bar is erected on a tripod to lie horizontally and perpendicular to the line of sight, centered over a survey point. The theodolite is centered over the other end of the line. The separation between the two targets defining the bar is a known length (commonly two meters). Using this known length and the measured angle, the horizontal distance is computed by trigonometry. For relatively short distances (up to perhaps 150–200 feet) this method can exceed the accuracy of both taping and electronic measurements, and thus it is very useful for measurements across busy streets or other small, inaccessible places. Because this method lacks error due to slope corrections, it is useful for measuring to high places, such as to tops of buildings. Its application is probably overlooked nowadays because of the ease of using electronic instruments—even though it is more accurate for short measurements, has advantages similar to those of electronic distance measurement, and avoids slope corrections.

## Electronic Distance Measurements

The most common of **electronic distance instruments** (EDMs) employs a visible light beam reflected off a system of reflectors called *retroprisms*. The light beam is reflected onto the instrument for interpretation of the wavelengths and partial wavelengths comprising the double-slope distance between the instrument and the reflector. Older EDMs are individual units, measuring distance only. The most modern version is part of the aforementioned total station system, which also measures angles. Whether an individual unit or part of a more complete survey system, the principle of EDM operation is much the same.

A common fallacy, particularly when using total station systems, is that the measurements are free of errors. As has been mentioned, there is error in any measuring system. When measuring with an EDM, the surveyor must be concerned with calibration, just as with any distance-measuring system. For example, the electronic center of the instrument may not be located precisely along the same vertical line as the geometric center plumbed over the ground station. The instrument will usually have this small, constant instrument error, as well as an error that is proportional to the distance measured (often called the *parts per million* or *PPM correction*). Also, the reflector has a "constant," and the optical plummets in the tribrach mounting systems can be out of adjustment. The magnitude and sign of these errors should be determined by field tests for best accuracy. EDMs are also affected by variations in atmospheric pressure and temperature. Microwave measurements are also affected by humidity.

All measurements made with EDMs are "slope" measurements between the center of the EDM and the reflector, and thus must be corrected to horizontal using measured vertical or zenith angles, and also instrument heights.

## Indirect Computational Methods

Distances are commonly determined by **indirect measurement** using trigonometric principles. The most common of these methods are *intersection* and the coordinate *inverse*. Using intersection, a distance is measured (or computed from other measurements) and angles measured to a common point from the two ends of this *base line*, which forms an oblique triangle. The unknown sides of the triangle are solved using the sine law. Using the inverse, the computation starts with determination of the departure (difference in  $x$  coordinates) and the latitude (difference in  $y$  coordinates) of the line defined by the two coordinate points. The distance is the hypotenuse of a right triangle whose sides are the departure and latitude of the line; it is determined using the Pythagorean theorem. Many variations of these indirect methods occur in practice and usually involve some variation of an oblique triangle solution or the coordinate inverse. Some of these concepts are discussed in section 145.2.

Photogrammetry is another such method for determining distances. After proper orientation of a stereomodel, coordinates can be read from it, from which distances can be determined by the same inverse computation as discussed earlier. The science of photogrammetry is briefly explained in **Chapter 147**.

## Approximate Methods

As mentioned, all measurements are estimates of an unknown quantity. The method used to determine a distance is chosen as dictated by the accuracy needed. The above methods are typical for measurements where accuracies of a few millimeters to a fraction of a meter are desired. If less accuracy is needed, a range finder, a measuring wheel, pacing, the odometer on a bicycle or vehicle, map scaling, digitizing coordinates from maps (with subsequent "inverse" computations), or even visual estimation can be used. Range finders are optical instruments that might be considered under the classification of tachymetric methods. These instruments and the calibrated measuring wheel or an odometer on a bicycle have accuracy comparable to the stadia method. A car odometer can yield an accuracy of perhaps 50 to 100 feet in a mile, if calibrated. The accuracies of the other approximate methods depend on several factors, particularly the map scale and the map accuracy. It must be emphasized that digitizing from maps cannot yield high accuracy of positions or distances. Even for a fairly large-scale map, the accuracy of such methods is seldom better than 10 feet.

## 145.2 Applications and Calculations

---

### Taping

The variables affecting a horizontal distance using a tape, with their correction equations, are as



follows:

1. *Calibration.*  $C_l = l_t - l_r$  . The correction per tape length is the actual ("true") length of the tape minus the nominal length (reading between end marks). The actual length is determined by calibration at some observed temperature, tension, and support condition, comparing the tape's length with an accurate base line.
2. *Temperature.*  $C_t = K_t l (t_f - t_s)$  . Temperature correction is given by the coefficient of thermal expansion multiplied by the nominal length of the tape times the difference between the field temperature and the calibration temperature. The correction is positive when  $t_f > t_s$  , since the tape expands when the temperature is warmer. The value for  $K_t$  is 0.00000645 units per unit per °F (the constant is 0.0000116 if °C is used).
3. *Pull (tension).*  $C_p = (P_f - P_s) l \div AE$  . This is the correction for tension if the field tension is different from the standardization tension. In this equation  $A$  is the cross-sectional area of the tape and  $E$  is the modulus of elasticity, which is 29000000 lb/in.<sup>2</sup>. Care should be taken that the field tension appears first in the equation, subtracting the standardization tension from it.
4. *Sag.*  $C_s = -(w^2 l) / (24 P^2)$  . This is the correction for sag, where  $w$  and  $l$  are the weight and length of the portion of the tape suspended between supports. The pull,  $P$ , is the actual tension on the tape. It has no relationship to the standardization pull.
5. *Slope.*  $C_g = -(v^2) / (2l)$  . This is the correction for grade or slope, which, when added to the slope length, yields the horizontal distance. This correction is applied when the tape ends are not at the same elevation. In this equation  $s$  is the slope distance being corrected and  $v$  is the elevation difference between the two tape ends at this length. This equation is inexact but is accurate for slopes less than about 10%. The Pythagorean theorem can be used in any case, solving the equation  $h^2 + v^2 = s^2$  for the horizontal distance,  $h$ .

Slope corrections can also be made trigonometrically if the vertical angle (slope angle) is measured with a clinometer or theodolite. For a line of any length  $H = S \cos \gamma$  , where  $H$  is the horizontal distance,  $S$  is the slope distance, and  $\gamma$  is the vertical angle. ( $H = S \sin \gamma$  if the zenith angle is used.)

The alignment error is generally not corrected in practice, but instead rendered negligible by the process of careful tape alignment.

For many errors the correction can usually be made for one tape length and then multiplied by the number of tape lengths, as long as the condition causing the error does not vary between tape lengths. This generally always applies to the calibration, temperature, and tension errors, and sometimes to the sag and slope errors.

Taping problems are of two types: (1) calculation of the horizontal distance between two established points, and (2) calculation of the reading to be observed to establish a given distance. The theory of systematic errors is applicable in making taping corrections. Solving for the true value  $T = R + \sum C$  is the first type of problem. If "layout" is required, then the reading  $R = T - \sum C$  is solved from the given value and the corrections.

The following problems will utilize a calibrated 100 ft tape, found to be 99.992 feet long at 70°F, 15 lb tension, fully supported. It has a cross-sectional area of 0.006 in.<sup>2</sup> and weighs 2.2 lb. In the solutions TL is the number of tape lengths.

**Example 145.1.** A reading of 458.97 feet is observed between two points when the field temperature is 40°F, along a 4% slope. Find the correct horizontal distance.

**Solution.** There are three systematic errors to consider: calibration, temperature, and slope.

$$C_L = (l_t - l_r)TL = (99.992 - 100.000)4.59 = -0.037 \text{ ft}$$

$$C_t = K_t l(t_f - t_s)TL = 0.00000645(100)(40 - 70)4.59 = -0.089 \text{ ft}$$

$$C_g = -\frac{v^2}{2l}TL = [4^2 \div (2 \times 100)](4.59) = -0.367 \text{ ft}$$

$$\sum C = -0.493 \text{ ft} , \text{ from which } T = 458.97 - 0.49 = 458.48 \text{ ft} .$$

**Example 145.2.** A distance of 200.00 feet is to be laid out along a horizontal alignment. The tape must be suspended for 60 feet of one of its tape lengths. A tension of 30 lb is used for this portion of the layout; otherwise, 15 lb is used. Temperature is 70°F.

**Solution.** There are three systematic errors to consider: calibration, tension, and sag.

$$C_L = (l_t - l_r)TL = (99.992 - 100.000)2.00 = -0.016 \text{ ft}$$

$$C_p = (P_f - P_s)l \div AE = (30 - 15)60 \div (0.006 \times 29 \times 10^6) = 0.005 \text{ ft}$$

$$C_s = -\frac{w^2 l}{24P^2} = -(2.2 \times 0.6)^2 60 \div (24 \times 30^2) = -0.005 \text{ ft}$$

$$\sum C = -0.016 \text{ ft} \quad \text{and} \quad R = 200.00 - (-0.016) = 200.016 = 200.02 \text{ ft}$$

It is seen in this example that the added tension for the end-supported part of the taping compensated for the sag effect.

## Stadia

Stadia hairs or lines are placed in most telescopes so that the *stadia interval factor*,  $K$ , equals 100. This makes it convenient to measure distance, merely subtracting the two stadia hair readings and multiplying by 100 to get the distance between the rod and the theodolite. Since the stadia hairs are each read with a precision of approximately  $\pm 0.01$  feet at ordinary distances, the precision of a stadia distance is no better than  $\pm 1.0$  feet, with variation according to how clearly the rod is seen and how carefully it is read.

Old *external focusing* telescopes had a *stadia constant*,  $C$ , of about one foot. The instruments of recent generations, however, being *internally focusing*, eliminate this constant.

For a horizontal sighting  $H = KI + C$ , where  $K$  and  $C$  are as defined earlier and  $I$  is the rod intercept (difference between the upper and lower rod readings). When vertical angles are involved,  $H = KI \cos^2 \gamma + C \cos \gamma$ , where  $\gamma$  is the vertical angle.

**Example 145.3.** A theodolite has a stadia interval factor of 100. The reading on the upper stadia hair is 7.54 ft and on the lower hair it is 3.66 ft. The zenith angle to the center crosshair is  $96^{\circ}36'30''$ . What is the horizontal distance, to the nearest foot?

**Solution.** Assuming the stadia constant to be zero and converting the zenith angle to vertical angle, the appropriate values in the preceding equation are as follows:

$$H = 100(7.54 - 3.66) \cos^2(-6^{\circ}36'30'') = 383 \text{ ft}$$

## Subtense Bar

After the horizontal angle is measured between the two end targets on the bar, the horizontal distance is computed from  $H = \frac{1}{2}b \cot \alpha/2$ , where  $b$  is the bar length and  $\alpha$  is the horizontal angle. Using a bar of the usual 2-meter length, the value of  $H$  is in meters, since half the bar length is 1 meter.

Note that the distance is *always* horizontal, since the horizontal angle is the same regardless of the relative elevation of the two points. Thus, no slope corrections are ever required.

In practice, a 1" theodolite is generally used and several angles are measured, in order to achieve adequate precision.

**Example 145.4.** A 1" theodolite is used to measure the angle between the targets on the ends of a 2-meter subtense bar. The mean of six independent readings of the angle is  $0^{\circ}45'46''$ . Compute the horizontal distance between the theodolite and the bar.

$$H = \frac{1}{2}b \cot \alpha/2 = \frac{1}{2}(2 \text{ meters}) \cot (0^{\circ}45'46''/2) = 75.110 \text{ m}$$

## Electronic Distance Measurements

Although the instrument constant is in practice, usually adjusted to zero whenever the instrument is serviced, calibration can discover a small instrument constant. Similarly, the reflector constant is usually keyed into the instrument by the surveyor and thus compensated, but a field test of reflector constants can discover slight discrepancies between what the manufacturer states the constant to be and what it actually is. Likewise, the atmospheric errors are generally keyed into the instrument after reading the temperature and pressure but are sometimes overlooked, and an old setting remains in the instrument. The surveyor should be aware of these possible error sources. The following example assumes that the atmospheric errors and reflector constant have been handled properly.

**Example 145.5.** An EDM has been calibrated using a four-station NGS base line, and errors are found as follows:  $C = +0.003$  meters and  $P = +0.000\,004\,56$ , where  $C$  is the constant correction and  $P$  is the "scale" correction, which may be expressed as 4.56 PPM. The zenith angle along the line is  $88^{\circ}34'42''$ . The observed slope distance is 1789.783 meters. What is the corrected

horizontal distance?

**Solution.** The PPM correction is  $+0.00000456(1789.8 \text{ m}) = +0.0082 \text{ m}$ . The constant correction is  $+0.003 \text{ m}$ . The corrected slope distance is

$$\begin{aligned} 1789.783 + 0.008 + 0.003 &= 1789.794 \text{ m} \\ H &= 1789.794 \sin 88^\circ 34' 42'' = 1789.243 \text{ m} \end{aligned}$$

An additional correction might be necessary if the reflector is not set at the same height as the EDM because, if so, the measured slope angle does not correspond to the slope of a line connecting the two ground points. The correction involves adjusting the measured zenith angle before calculating the horizontal distance as done previously.

## Defining Terms

**Distance:** A linear value, either measured or computed. Unless qualified otherwise, it is understood to lie along the horizontal. The observed distance is considered inaccurate until corrected for systematic errors caused by instruments, nature, or other sources.

**Electronic distance instrument:** An instrument that measures distances using reflected light waves or microwaves.

**Indirect measurement:** A measurement that has been computed from other measurements, usually employing trigonometric principles.

**Tacheometry:** A distance-measuring procedure that involves measuring intervals between cross hairs on a rod or angles subtended between marks on a bar of known length, or a similar indirect method.

**Tape:** A surveying instrument used to measure distance by stretching it, end for end, along a line between points.

## Reference

Buckner, R. B. 1983. *Surveying Measurements and Their Analysis*. Landmark Enterprises, Rancho Cordova, CA.

## Further Information

Davis, R. E., Foote, F. S., Anderson, J. M., and Mikhail, E. M. 1981. *Surveying Theory and Practice*. McGraw-Hill, New York.

Moffitt, F. H. and Bouchard, H. 1992. *Surveying*. Harper Collins, New York.

Wolf, P. R. and Brinker, R. C. 1993. *Elementary Surveying*. Harper & Row, New York.

Bon A. Dewitt. "Directions"  
*The Engineering Handbook.*  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 146

## Directions

---

- 146.1 Angles
- 146.2 Meridians
- 146.3 Direction
- 146.4 Back Bearing and Back Azimuth
- 146.5 Applications

**Bon A. Dewitt**  
*University of Florida*

In surveying, direction is the term used to denote the course or heading of a line. Here, a line is defined by its end points, giving it a magnitude (length) and direction, much like a vector. By convention, direction is specified in terms of angles and is separated into horizontal and vertical components. This chapter deals primarily with the horizontal component of direction within the context of **plane surveys**. In **geodetic surveys**, where the earth's curvature is taken into account, the fundamental concepts of direction still apply, though their use in subsequent calculations is far more complex.

Traditionally, direction has been established through astronomic observations or compass readings. Although these traditional approaches may still be applicable in certain situations, the **Global Positioning System** is now preferred due to its accuracy and convenience.

### 146.1 Angles

---

Angles form the basis for quantification of direction. There are several conventions available for specification of angular units. In the U.S., the sexagesimal system is currently the most commonly used for surveying applications. In this system a full circle is divided into 360 degrees, with further subdivisions into minutes and seconds. Sixty minutes is equivalent to one degree and sixty seconds is equivalent to one minute. Thus, an angle can be expressed in degrees ( $^{\circ}$ ), minutes ( $'$ ), and seconds ( $''$ ), much like time is expressed in terms of hours, minutes, and seconds. As an alternative, angles can be expressed in terms of degrees and a decimal fraction thereof, though this is not conventional surveying notation.

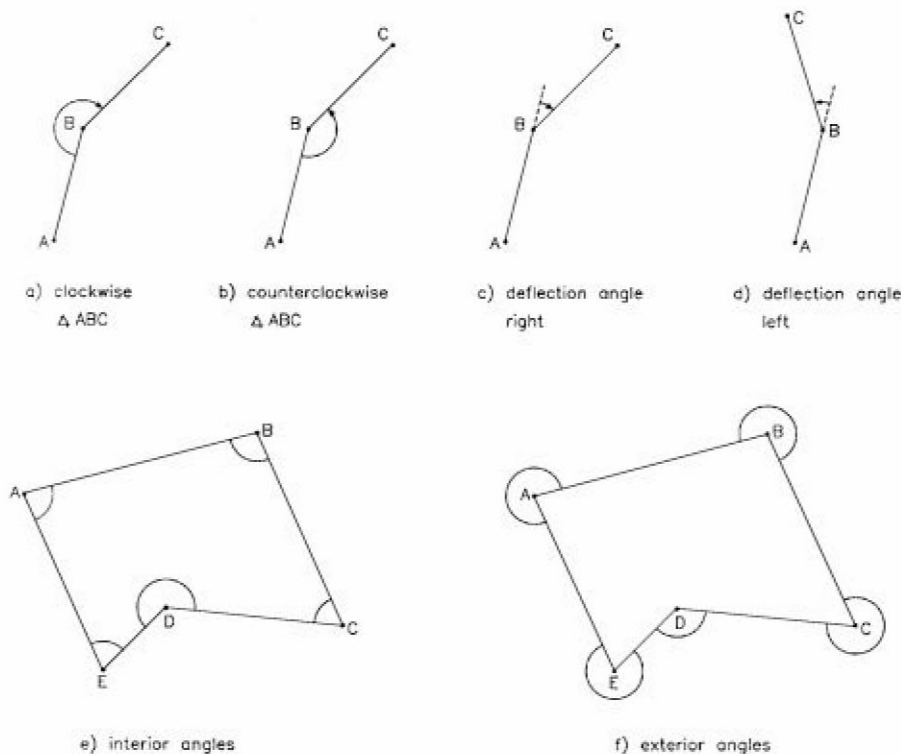
Other angular units are available, such as radians, grads, or mils [see Eq. (146.1) for unit equivalents]. The radian is a dimensionless unit that utilizes the ratio of the length of a circular arc to its radius to express the magnitude of the subtended angle. The grad (or *gon*) is a unit of measure in the centesimal system (widely used in Europe) that corresponds to 1/400 of a full circle. The mil is an angular unit corresponding to 1/6400 of a full circle and is used by the U.S. military,

primarily in artillery applications. These three alternate systems are decimal based, as opposed to the base-sixty approach of the sexagesimal system.

$$\text{Right angle} = 90 \text{ degrees} = 100 \text{ grads} = 1600 \text{ mils} = \pi/2 \text{ radians} \quad (146.1)$$

**Horizontal angles** are measured in a plane perpendicular to the direction of gravity. There are many types of horizontal angles, such as angles to the right (clockwise) or left (counterclockwise), interior angles, exterior angles, and deflection angles. Figure 146.1 shows the various types of angles. Angles to the right—with the backsight, occupied, and foresight points specified—are recommended for modern surveys due to their applicability to electronic data collectors and computer software. An illustration of this notation is given in Fig. 146.1(a), where point A is the backsight, B is occupied, and C is the foresight point.

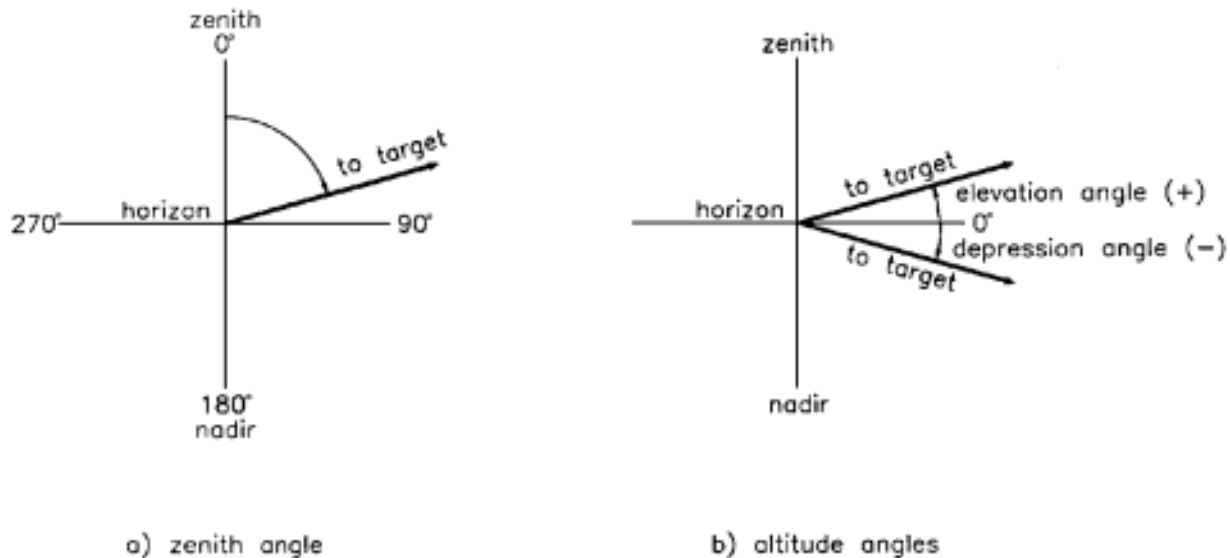
**Figure 146.1** Illustration of various types of horizontal angles.



**Vertical angles** are measured in a plane perpendicular to the horizontal. There are two fundamental types of vertical angles: zenith angles and altitude angles. A zenith angle (or zenith distance), illustrated in Fig. 146.2 (a), is the angle from the observer's zenith direction to the target. Zenith angles range from  $0^\circ$  to  $360^\circ$ , with  $90^\circ$  and  $270^\circ$  corresponding to the direction of the horizon in the direct and reverse position, respectively. An altitude angle, illustrated in Fig. 146.2 (b), is the angle from the observer's horizon to the target. Altitude angles above the horizon are called *elevation angles* and are considered positive by convention, whereas those below the

horizon are called *depression angles* and are considered negative.

**Figure 146.2** Profile view illustrating two types of vertical angles.



Surveyors determine angles with instruments known as *transits* or *theodolites*. Both instruments enable the user to perform the same basic functions, that is, to measure or establish horizontal and vertical angles. Generally, theodolites are more accurate and precise than transits, though this is not always the case. Some modern theodolites have electronic angle-reading systems and are incorporated into **total stations**, which also have the capability of measuring distances electronically. Total stations are particularly convenient due to their ability to feed recorded data directly to a field computer.

## 146.2 Meridians

In order to specify the horizontal component of direction, it is first necessary to specify the reference meridian. A meridian is an imaginary line that is selected as the nominal north-south indicator in the observer's horizon plane. It can be based on any one of several references: geodetic, astronomic, magnetic, grid, or assumed.

A geodetic (also called *geographic*) meridian is based on the north and south poles as defined by a particular latitude and longitude reference or graticule. It has been demonstrated that the rotational axis of the earth changes slightly over time, so in essence the geographic graticule is a "snapshot" in time. This reference becomes standardized by virtue of published coordinates for a network of monumented points, based on its definition.

An astronomic meridian is based on the rotational axis of the earth and the direction of gravity. It derives its name from the means by which it is typically established: astronomic observations. The angular difference between astronomic and geodetic meridians is expressed in terms of the Laplace equation. For most practical purposes in plane surveying, this difference is negligible and both meridians are collectively referred to as the "true" meridian.

A magnetic meridian is based on the magnetic north and south poles of the earth. These poles are



distinct from the geographic poles and change appreciably over time. The angle from true north to magnetic north is called the *magnetic declination* and is a function of the observer's location with respect to the poles. The effect of magnetic declination can be quite large—for example, in parts of Alaska, magnetic declination is greater than  $30^\circ$  east.

A map projection is a distorted rendition of a portion of the curved earth's surface on a surface that can be laid out flat. The projection has an inherent  $x, y$  coordinate system, with  $y$  in the general direction of north. In the projection any line parallel to the  $y$  axis is a grid meridian. They are different from the three meridians previously mentioned, in that grid meridians are parallel to each other, whereas the others are not.

Assumed (arbitrary) meridians are chosen for convenience. Here, a direction is arbitrarily specified for a line connecting two survey points. This direction is often chosen so as to approximate some specific meridian; however, no actual observation of a meridian is performed. There is an inherent risk associated with assumed meridians. If one or both of the survey points is lost, the assumed meridian becomes unrecoverable.

## 146.3 Direction

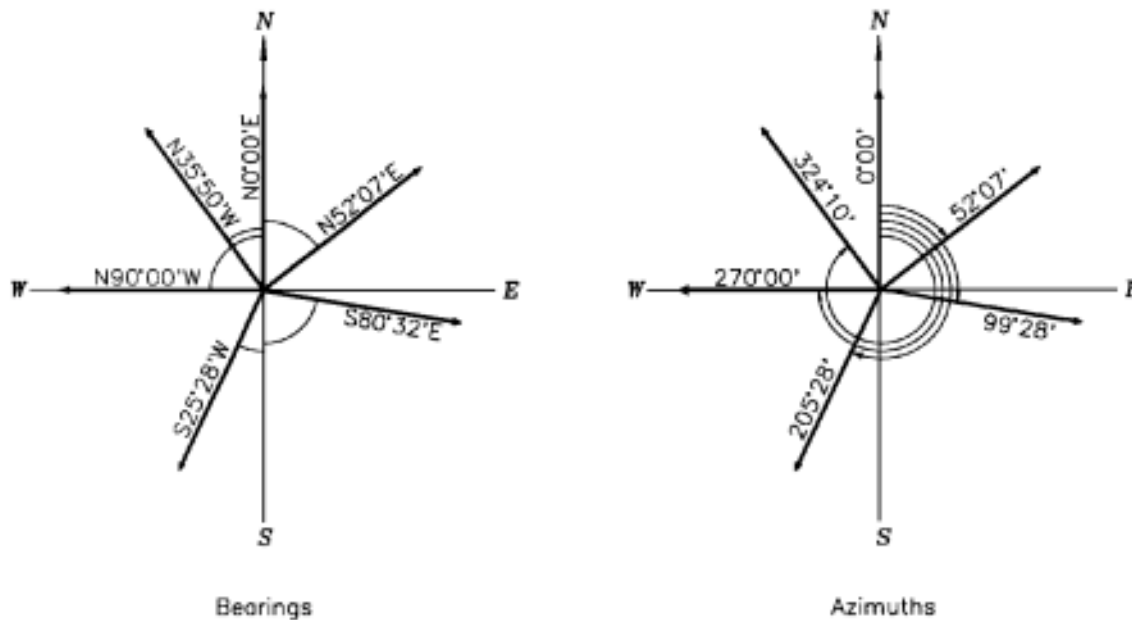
---

In mathematics, polar coordinates are often used to specify the position of a point. Here, the direction of the line from the origin to the point is based on the angle from the positive  $x$  axis, with counterclockwise angles being positive. In surveying, the direction of a line can be expressed either in terms of **bearing** or **azimuth**. Both forms depend on the meridian definition mentioned earlier. No matter which approach is used, it is important to clearly specify the meridian upon which the direction is based.

Bearing of a line is specified as an acute horizontal angle between the line and meridian, along with letters specifying the proper quadrant. It is expressed in the form of the letter N or S, followed by an angle (less than or equal to  $90^\circ$ ), followed by the letter E or W.

Azimuth is specified as the clockwise horizontal angle from the meridian to the line. The angular value is positive and less than  $360^\circ$ . Azimuths are commonly specified from north, though this is not a universally accepted standard. Some applications use azimuths that are referenced from south. Due to this possible ambiguity, one should clearly indicate whether a north or south reference is implied. [Figure 146.3](#) gives examples of bearings and azimuths (from north) for selected lines.

**Figure 146.3** Examples of bearings of selected lines and their equivalent azimuths.



## 146.4 Back Bearing and Back Azimuth

Back bearings are expressions of the opposite direction of a line. The expression is formed by starting with the original (forward) bearing and then changing the "sense" of the letters. An N is changed to an S (or vice versa), and an E is changed to a W (or vice versa); however, the angular value remains the same. For example, if the bearing from point 1 to point 2 is N47°15' W, its back bearing (i.e., the bearing from point 2 to point 1) is S47°15' E.

Back azimuths are also expressions of the opposite direction of a line. The expression is formed by adding 180° to or subtracting 180° from the original (forward) azimuth, keeping in mind that the result must be in the range of 0 to 360°. For example, if the azimuth from point 1 to point 2 is 312°45' , its back azimuth (i.e., the azimuth from point 2 to point 1) is 132°45' .

The foregoing simple relations for back bearings and back azimuths are applicable only to plane surveys of a limited extent, where it can be assumed that all meridians are parallel. In surveys covering a large area, earth curvature and meridian convergence are appreciable factors and therefore a more complicated relation must be used.

## 146.5 Applications

There are many applications in surveying that call for the use of directions. Property surveys, geodetic control surveys, transportation corridor (route) surveys, and topographic surveys are but a few. Most applications involving directions utilize plane trigonometry in the solution.

Computations involving addition and subtraction of angles, sine and cosine laws, right triangle relationships, and sum of angles in a closed polygon are routinely performed.

### Defining Terms

**Azimuth:** An expression for the direction of a line consisting of the clockwise horizontal angle ( $\geq 0^\circ$  and  $< 360^\circ$ ) from one end of the meridian. Azimuths from north are generally used;

however, some conventions employ azimuth from south.

**Bearing:** An expression for the direction of a line consisting of the horizontal angle ( $\leq 90^\circ$ ) that the line makes with the meridian in conjunction with prefixed and postfixed letters that specify the quadrant.

**Geodetic survey:** A survey in which the earth's true three-dimensional shape and gravity field are taken into account.

**Global Positioning System:** A system of satellites and ground receivers that enables users to determine geodetic coordinates of points to a high degree of accuracy. The system is under the control of the U.S. Department of Defense but has been used for civilian applications since the early 1980s.

**Horizontal angle:** An angle that is defined in a plane perpendicular to the direction of gravity.

**Meridian:** In a global context a meridian is the intersection of the plane containing the north pole, the south pole, and the observer's position with the spheroidal figure that approximates the earth. In a local context a meridian is a reference line that defines the north-south direction. This reference can be on a geodetic, astronomic, magnetic, grid, or assumed basis.

**Plane survey:** A survey of limited extent and accuracy in which the earth's surface is assumed to be a plane. This assumption permits the use of plane trigonometry in computations involving coordinates and other parameters.

**Total station:** A device used in surveying that incorporates an electronic theodolite with an electronic distance-measuring instrument and a computer. The device can automatically read and record horizontal and vertical angles and slope distances.

**Vertical angle:** An angle that is defined in a plane parallel to the direction of gravity.

## Reference

Wolf, P. R. and Brinker, R. C. 1994. *Elementary Surveying*, 9th ed. HarperCollins College, New York.

## Further Information

Bomford, G. 1980. *Geodesy*, 4th ed. Oxford University Press, New York. This book is considered to be a definitive text on the subject of geodesy, with a high degree of mathematical rigor.

Brinker, R. C. and Minnick, R. (Eds.). 1995. *The Surveying Handbook*, 2nd ed. Chapman and Hall, New York. A good general reference on virtually all topics of surveying.

Davis, R. E., Foote, F. S., Anderson, J. M., and Mikhail, E. M. 1981. *Surveying: Theory and Practice*, 6th ed. McGraw-Hill, New York. A good mathematical treatment of various surveying topics.

Bethel, J. "Photogrammetry and Topographic Mapping"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Photogrammetry and Topographic Mapping

---

147.1 Basic Concepts

147.2 Orientation and Model Setup

147.3 Data Collection for Topography

147.4 Data Processing for Topography

147.5 Data Presentation

**Jim Bethel**

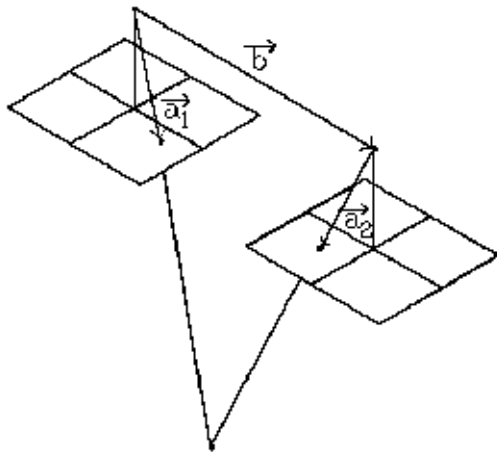
*Purdue University*

## 147.1 Basic Concepts

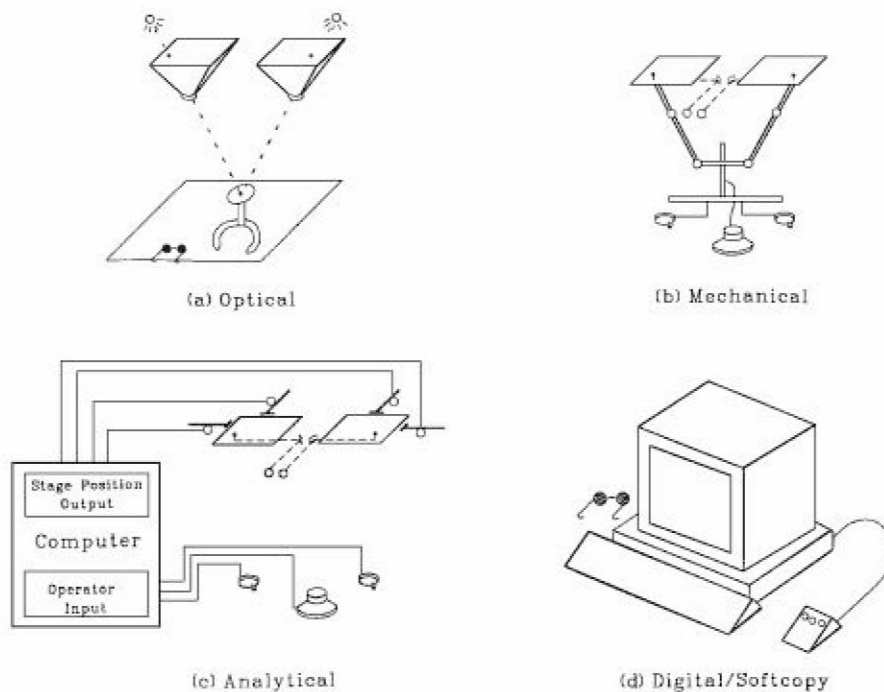
---

The term *photogrammetry* refers to the measurement of photographs and images for the purpose of determining the size, shape, position, and other spatial attributes of features appearing in the images. The most common application of this technique is aerial photogrammetry, in which nominally vertical photographs are used to produce topographic maps, which are often used for engineering design and land development. Aerial cameras are made to very exacting tolerances, and whatever small systematic errors may be present in the resulting photographs can be modeled mathematically so that very accurate ground positions and elevations can be inferred from photograph measurements. The mathematical basis of the imaging equations which relate object points (3-D) to image points (2-D) is that of a perspective projection, with a point (actually two points) in the lens assembly serving as the perspective center(s). A terrain point projected into two adjacent frame photographs is shown in [Fig. 147.1](#). The geometric relationship between the image and object spaces may be modeled in *analog* fashion by optical rays or by mechanically gimbaled steel rods. Today it is more commonly modeled mathematically in an *analytical* instrument or possibly in a *softcopy* workstation. Stereo data extraction instruments and systems are shown schematically in [Fig. 147.2](#).

**Figure 147.1** Terrain point projected into two frame photographs.



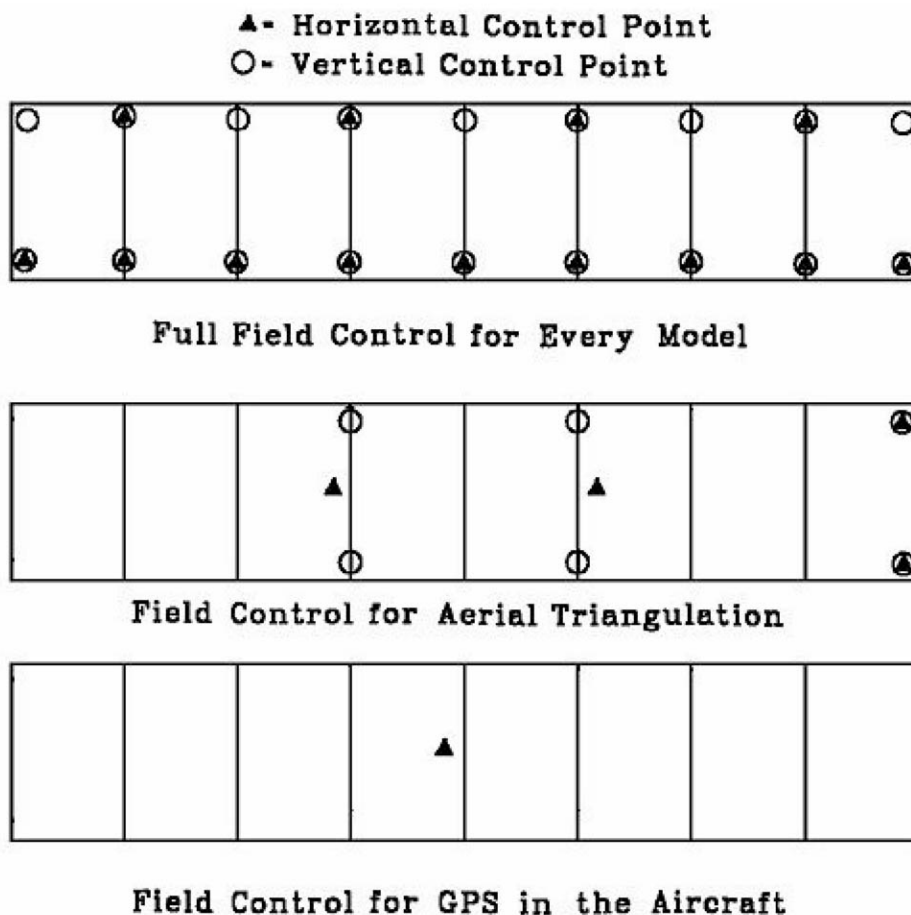
**Figure 147.2** Photogrammetric stereo data extraction systems.



Historically, the production of topographic maps from aerial photographs has been a very labor-intensive operation, beginning with the establishment of *ground control points*, proceeding through *aerial triangulation* to densify the ground control, and culminating in the meticulous tracing of planimetric features from adjacent *stereo pairs* and the tracing of contour lines to depict the shape of the landforms. Control point requirements for several scenarios are shown in [Fig. 147.3](#). In recent years, several trends have emerged which promise to make the mapping process more efficient. The use of GPS, the *global positioning system*, for ground control as well as for

direct observation of the *exposure stations* in the aircraft, is greatly simplifying the preliminary steps in the photogrammetric mapping process. High-level image processing of photographs converted to digital form is also beginning to replace the sometimes tedious operations of point selection, orientation, and even feature extraction.

**Figure 147.3** Variations of control requirements for a strip.



## 147.2 Orientation and Model Setup

Since photographs are inherently two-dimensional representations of a three-dimensional world, it takes a minimum of two photographs of the region of interest, taken from different points of view, to reconstruct the 3-D scene including the topographic landforms. Having the same scene viewed from two different locations gives rise to *parallax*, which our eyes interpret as a depth cue, much the same as in normal binocular vision. In order to position the two photographs of such a stereo pair, they must be aligned in the viewing instrument so that the flight paths coincide. Also, any tilts present in the camera at the instant of exposure must be reintroduced. This process is called *relative orientation*. When the *stereo model* is so oriented, then viewing of the entire overlap area

can be made without misalignments in the "cross-flight" direction, that is, without *y-parallax*.

Following relative orientation, one usually wishes to tie the "local" model coordinates to the object or ground coordinate system. Such a system is either a local (project-specific) Cartesian system, or one of the common map projection systems such as the state plane systems in the U.S. State plane systems are either transverse mercator projections or Lambert conic projections, depending on the shape and orientation of the zone. The process of relating the local model space to the object space with a given coordinate system is referred to as *absolute orientation*.

### 147.3 Data Collection for Topography

---

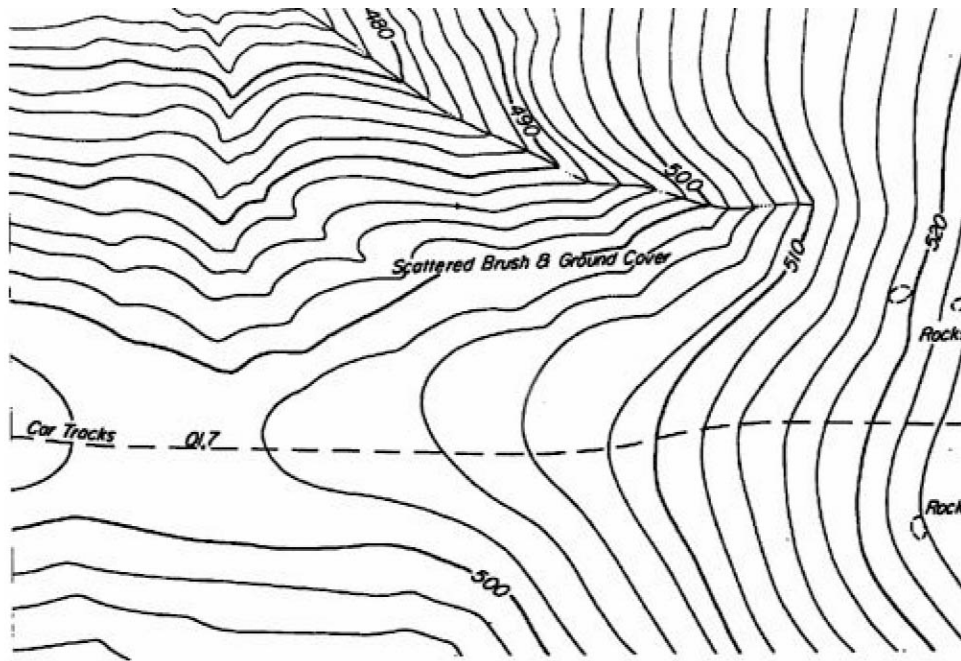
Terrain data for constructing a topographic map can be collected either in the field or from aerial photographs or other imagery. There is a certain threshold project size, below which it is more economical to collect data directly in the field, and above which it is more economical to first acquire aerial photographs and then collect the terrain data from the stereo photographs. This threshold project size is around 8 hectares (20 acres).

In previous decades, field procedures might have involved a *plane table*. Today field procedures most likely involve a *total station* with an electronic data collector. Such a system can record either raw angles and distances, or point coordinates, along with a point identification and a feature code. Using a total station, data points can be distributed as a *grid*, as points along a series of *cross sections*, or as a collection of random points dictated by the terrain. In both the cross section method and the random point method, field judgment is exercised to collect only significant planimetric or topographic features, or points that represent slope discontinuities. In addition, *break lines* may be collected to enforce sharp breaks in the slope continuity.

Not surprisingly, similar strategies are employed by the photogrammetrist in collecting terrain data on a *stereoplotter* or *digital (softcopy) stereo workstation*. Grids, cross sections, and random point collection are widely used in photogrammetric practice. Such data often take the form of a *digital terrain model (DTM)*, *digital elevation model (DEM)*, *digital height model (DHM)*, or *digital terrain elevation data (DTED)*. The photogrammetrist has other options as well. With the stereo view from above, direct tracing of contour lines is possible. In fact, until recently, this was the principal means of collecting topographic data from aerial photographs. A manually compiled contour map is shown in [Fig. 147.4](#). Another approach involves the collection system analyzing the data on the fly and interacting with the stereo operator to densify regions until all high-spatial-frequency terrain features are determined. This process is referred to as *progressive sampling* and, while intriguing, has never achieved wide acceptance. With the emerging dominance of digital imagery, automated methods of terrain surface extraction by image-processing techniques are gaining widespread use and acceptance. Such *matching* techniques, intended to duplicate the function of the human operator, may need to be tuned and adapted to different scales, terrain types, and image types.



**Figure 147.4** Manually compiled contour map.



Matching techniques often begin by aggregating the fine-resolution digital image into a coarser-resolution image, stopping the process when the aggregated image may be only a few *pixels* on a side. Such a progressive series of image resolutions is called an *image pyramid*. Matching at the coarse resolutions is effectively operating on a low-pass-filtered image, and thus reveals the low-spatial-frequency components of the terrain shape. Operating on the successively finer-resolution images, all the way down to the finest resolution, will reveal successively higher-spatial-frequency components of the terrain shape. Matching can be done by features or by raw grey levels or colors. It can be constrained by known image orientation data, or it can be unconstrained. It can take place in one dimension, as in *epipolar* resampled imagery, or it can take place with areas or area elements, as in *least squares* matching or *vertical line locus (VLL)* matching.

As automated terrain extraction from digital imagery becomes increasingly robust and reliable, other sources of digital imagery, besides scanned aerial photographs, may become very important. In particular, the recent relaxation of security concerns toward former Soviet Union countries has allowed the U.S. government to permit the dissemination of satellite imagery with a pixel size as small as 1 meter for civilian applications. Three commercial systems are currently planned to begin providing such data in the 1- to 3-meter range during 1996 and 1997. Such systems could have an enormous impact on the generation of DEM data in the years ahead. One caution is that the image geometry from such systems will be dynamic and time dependent and will therefore be much more difficult to model mathematically than simple frame aerial photographs. Thus digital softcopy workstations will become a necessity for handling such imagery rather than the relatively simple photogrammetric stereoplotters in wide use today.

A final note on data collection regarding GPS. Up to now field-collected elevations from GPS have been of only limited use. This is because engineers and land developers wish to have elevations in a system referenced to *mean sea level (MSL)*, whereas GPS height data is referenced

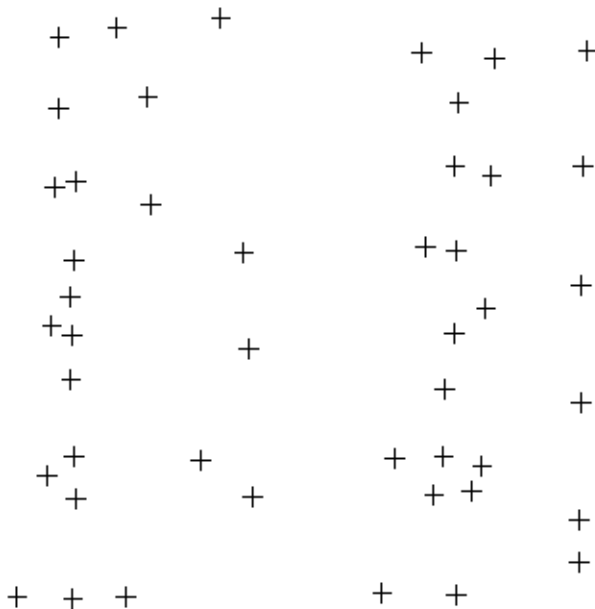
to the *ellipsoid* surface. The difference between the ellipsoid surface and the *geoid* surface is referred to as the *geoid undulation* or *geoid separation*. As our maps of geoid undulation become denser and more accurate, it will become increasingly possible to reduce GPS height data to MSL data. Thus another option will emerge for the rapid field collection of elevation data.

## 147.4 Data Processing for Topography

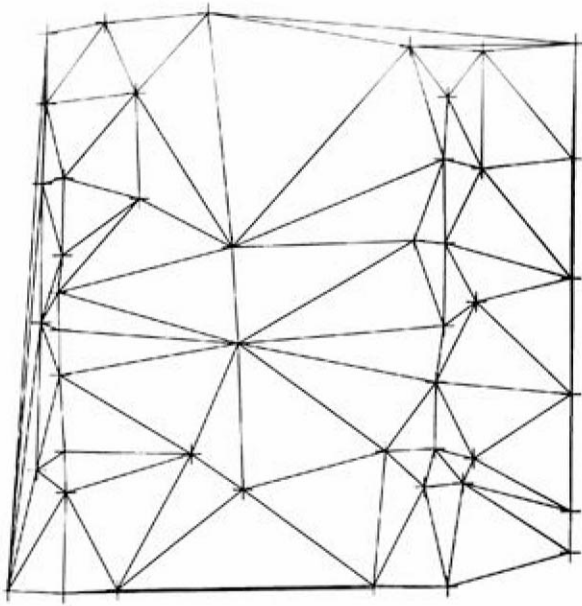
---

Data points collected either by field methods or by photogrammetric methods can be processed in a number of ways to make them useful for design engineers and others. Some systems are designed to work with a grid arrangement of the data points. If the raw data were not collected on such a grid, then the heights at the grid locations must be interpolated. This interpolation may be accomplished by *linear prediction*, by patchwise polynomials, or by a "moving-surface" method. If the data are in grid form, or have already been converted to grid form, then another interpolation task is necessary to generate heights at arbitrary locations. This could also be a moving-surface method, but grids are particularly well suited to techniques such as bilinear interpolation. If the data points are not in a grid, then they are often organized into a *triangulated irregular network (TIN)*. Representing the terrain surface by a TIN usually requires fewer data points than a grid, but there is a penalty in terms of access time. A small set of measured data points is shown in [Fig. 147.5](#), and the generated TIN is shown in [Fig. 147.6](#). The interpolated contours are shown in [Fig. 147.7](#).

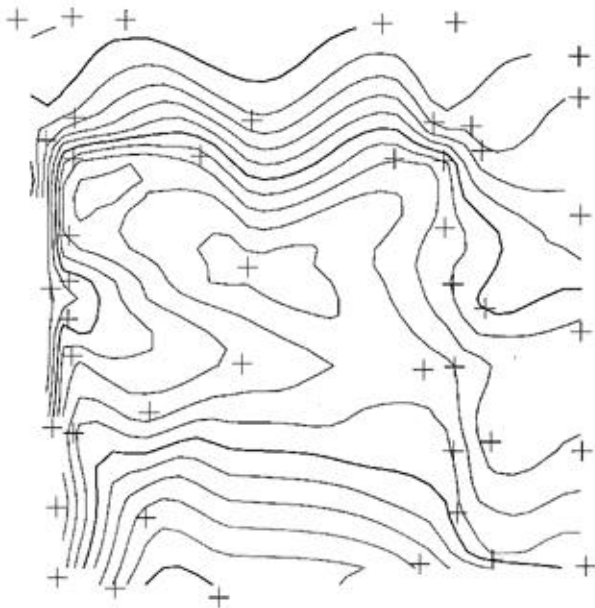
**Figure 147.5** Sample random point data set.



**Figure 147.6** TIN generated from points in Fig. 147.5.



**Figure 147.7** Interpolated contours from TIN.



The usual criterion for generating the triangular mesh in a TIN is that the triangles be as nearly equilateral as possible. For any given set of points there is a unique network of triangles, with the points as vertices, that most closely satisfies this criterion. This is called the *Delaunay*

*triangulation*. It may be generated from the data points by several algorithms. One of the simplest is the Watson algorithm outlined on the next page.

1. Create three fictitious points such that the defined triangle includes all of the data points.
2. Pick a data point.
3. Find all existing triangles (in the first pass there is only one, from step 1 above) whose circumscribed circle contains the point.
4. Make the union of all triangles from step 3, forming an "insertion polygon."
5. Destroy all internal edges in the insertion polygon, and connect the current point to all vertices of the polygon.
6. Go back to step 2, pick another point, and proceed until all points have been processed.
7. When done, eliminate any triangle containing a vertex from the three "artificial" points from step 1.

To enforce a break line, one can overlay the break line on the preliminary TIN and introduce new triangles as required to keep the break line continuous.

## 147.5 Data Presentation

---

Topographic data have traditionally been presented as contour lines on a hardcopy map. For engineering design work, presentation scales (i.e., map scales) in the U.S. have been 1:600 (1 inch = 50 feet), 1:1200 (1 inch = 100 feet), or 1:2400 (1 inch = 200 feet). In current practice such data might become a layer in a *geographic information system (GIS)*, and thus there may never be an archival document containing the topographic data. Many GIS systems are able to integrate DEM data, either grid or TIN, better than contour data. This is so because DEM data are more amenable to interpolation and manipulation than contour data. Derived information and products which can be produced from DEM data include volumes and earthwork quantities, flood levels, drainage patterns, arbitrary profiles, unobstructed line-of-sight diagrams, wireframe and shaded perspective views, image draped perspective views, and *orthophotographs*, or differentially rectified images.

Photogrammetry provides the most efficient way to generate topographic data for all but very small projects, where direct measurement in the field may be used. Photogrammetric techniques are undergoing a transition from strictly manual methods to increasingly automated methods. Likewise, there is a trend toward digital softcopy stereo workstations as opposed to the older hardcopy-based analog or analytical stereoplotter. Finally, it seems likely that a new generation of high-resolution commercial satellite imaging systems will gradually supplant film-based photographic methods, at least for small- and medium-scale mapping.

## References

- Burnside, C. D. 1985. *Mapping from Aerial Photographs*. John Wiley & Sons, New York.
- Chen, W.-F. (Ed.). 1995. *Civil Engineering Handbook*. CRC Press, Boca Raton, FL.

Kraus, K. 1993. *Photogrammetry*. Dummmler Verlag, Bonn, Germany.  
Leick, A. 1990. *GPS Satellite Surveying*. John Wiley & Sons, New York.  
Moffitt, F. H. and Mikhail, E. M. 1980. *Photogrammetry*, 3rd ed. Harper & Row, New York.  
Slama, C. C. (Ed.). 1980. *Manual of Photogrammetry*, 4th ed. American Society of  
Photogrammetry and Remote Sensing, Bethesda, MD.  
Wolf, P. R. 1983. *Elements of Photogrammetry*. McGraw-Hill, New York.

## **Further Information**

American Society for Photogrammetry and Remote Sensing (ASPRS), 5410 Grosvenor Lane,  
Suite 210, Bethesda, MD 20814-2160. Tel. 301-493-0290.  
American Congress on Surveying and Mapping (ACSM), 5410 Grosvenor Lane, Bethesda, MD  
20814-2122. Tel. 301-493-0200.  
*Journal of Surveying Engineering*, published quarterly by the American Society of Civil  
Engineers, 345 East 47th Street, New York, NY 10017-2398.  
*Photogrammetric Engineering and Remote Sensing*, published monthly by ASPRS.  
*The Photogrammetric Record*, journal published by the Photogrammetric Society, London, UK.  
*Surveying and Land Information Systems*, published quarterly by ACSM.

van Gelder, B. H. W. "Surveying Computations"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Surveying Computations

---

148.1 Principles of Multivariate Calculus

148.2 Principles of Linear Algebra

148.3 Model of Two Sets of Variables, Observations, and Parameters: The Mixed Model

148.4 Observations as a Function of Parameters Only: The Model of Observation Equations  
Notation

148.5 All Parameters Eliminated: The Model of Condition Equations

148.6 An Example: Traversing

Directional Measurements  $r_{ij}$  • Distance Measurements  $s_{ij}$

148.7 Dynamical Systems

**Boudewijn H. W. van Gelder**

*Purdue University*

Surveyors collect measurements (observations, data, etc.) for the purpose of determining a wide variety of variables (parameters, unknowns, etc.). These measurements may be distances, directions, angles, look angles, azimuths, elevation angles, height differences, time, and so on. From these observations the surveyor deduces results in terms of parameters, such as coordinates of an intersection point of two boundary lines or the height of a marker on a bridge. Rather than identifying a variety of (classical) survey problems such as point positioning through traversing or height determinations through spirit leveling, this chapter follows a more general approach by discussing a general relationship between observational data and parameters. In section 148.6 an example is presented.

## 148.1 Principles of Multivariate Calculus

---

A general relationship between two classes of variables may exist. One class of variables is denoted by  $l$  and the other by  $x$ . The two classes of variables refer to the two main classes a surveyor deals with: observations or the input data ( $l$ ) and the results or output data ( $x$ ). The variable  $l$  represents a group of  $n$  variables  $l_i$  with  $i = 1, \dots, n$ . The other variable  $x$  represents a group of  $u$  variables  $x_j$  with  $j = 1, \dots, u$ . There exists a functional relationship  $F$  between these two groups of variables. This functional relationship is nothing else than the mathematical model which expresses the assumed interdependency between the variables in question. Realize, though, that it is the surveyor who makes this enormously important judgment call on this mathematical relationship: like a car driver engages a gear, the surveyor engages at a certain moment a

mathematical model, or at least follows the spiritual father of the model provided in terms of some survey software package. The number of functional relationships does not need to be equal to  $n$  or  $u$ ; we may have  $F_k (k = 1, \dots, r)$  relationships. Summarizing, we have

$$F_k = F(l_i, x_j) \quad (148.1)$$

with

$$k = 1, \dots, r \quad (148.2)$$

$$i = 1, \dots, n \quad (148.3)$$

$$j = 1, \dots, u \quad (148.4)$$

A simple example will illustrate. Suppose we observe four pairs of coordinates  $\{X_l, Y_l\}$ , which we assume lie on a circle. So the model  $F_k$  is the equation of a circle, centered at an arbitrary coordinate  $\{X_c, Y_c\}$  with an arbitrary radius  $R$ :

$$F_k = (X_k - X_c)^2 + (Y_k - Y_c)^2 - R^2 \quad (148.5)$$

Geometry dictates that a circle is defined through three points (six coordinates), provided that those three points are not on a line. So four sets of coordinates will not quite fit the "circle" model. Small adjustments to the eight coordinates will be necessary to end up with one unique circle. In addition, we may have to slightly adjust the unknown circle centered at  $\{X_c, Y_c\}$  and its unknown radius  $R$ . We will probably have to slightly adjust any initial guess of these three variables  $\{X_c, Y_c, R\}$  to find the optimum circle passing through those four pairs of coordinates. These small adjustments to the variables  $l$  and  $x$ , which we call  $dl$  and  $dx$ , will make the variables fit the circle model just right. Adding the small corrections  $dl$  and  $dx$  to  $l$  and  $x$  will make the equations  $F_k$  become equal to zero:

$$F_k = F(l_i + dl_i, x_j + dx_j) = 0 \quad (148.6)$$

The problem of finding these hopefully small corrections  $dl_i$  and  $dx_j$  will be made much easier if we linearize the model according to the rules of multivariate calculus:

$$F_k(l_i + dl_i, x_j + dx_j) = F_k(l_i, x_j) + \frac{\partial F_k}{\partial l_i} dl_i + \frac{\partial F_k}{\partial x_j} dx_j + \dots = 0 \quad (148.7)$$

With the wealth of variables in Eq. (148.7),

- $r$  equations  $F_k$
- $n$  variables  $l_i$
- $u$  variables  $x_j$

at hand, it is much more practical to use the tools (vectors and matrices) of linear



algebra.

## 148.2 Principles of Linear Algebra

---

Equation (148.7) can be rewritten in terms of vectors and matrices. Defining the vectors (shown below in transposed form)  $\mathbf{L}$ ,  $\mathbf{X}$ , and  $\mathbf{W}$  and the matrices  $\mathbf{A}$  and  $\mathbf{B}$  according to

$$\mathbf{L}^T = [l_1, l_2, \dots, l_i, \dots, l_n]^T \quad (148.8)$$

$$\mathbf{X}^T = [x_1, x_2, \dots, x_j, \dots, x_u]^T \quad (148.9)$$

$$\mathbf{W}^T = [F_1, F_2, \dots, F_k, \dots, F_r]^T \quad (148.10)$$

and

$$\mathbf{A} = \begin{bmatrix} \frac{\partial F_1}{\partial l_1} & \dots & \frac{\partial F_1}{\partial l_n} \\ \vdots & \dots & \vdots \\ \frac{\partial F_r}{\partial l_1} & \dots & \frac{\partial F_r}{\partial l_n} \end{bmatrix} \quad (148.11)$$

$$\mathbf{B} = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \dots & \frac{\partial F_1}{\partial x_u} \\ \vdots & \dots & \vdots \\ \frac{\partial F_r}{\partial x_1} & \dots & \frac{\partial F_r}{\partial x_u} \end{bmatrix} \quad (148.12)$$

Matrix  $\mathbf{A}$  is the Jacobian of the model equations  $F$  with respect to the variables  $l$ , and consists of  $r$  rows and  $n$  columns. Similarly, the matrix  $\mathbf{B}$  is the Jacobian of the model equations  $F$  with respect to the variables  $x$ , and consists of  $r$  rows and  $u$  columns. The vector/matrix notation enables us to shorten Eq. (148.7) to

$$\mathbf{W} + \mathbf{B} d\mathbf{L} + \mathbf{A} d\mathbf{X} + \dots = 0 \quad (148.13)$$

Referring back to our circle fitting example, we have, for instance,

$$d\mathbf{L}^T = [dX_1, dY_1, dX_2, dY_2, dX_3, dY_3, dX_4, dY_4]^T \quad (148.14)$$

$$d\mathbf{X}^T = [dX_c, dY_c, dR]^T \quad (148.15)$$

The matrix element  $A_{3,2}$  (third row, second column) is equal to

$$A_{(\text{row}=3, \text{col}=2)} = \frac{\partial F_3}{\partial Y_c} = -2 \times (Y_3 - Y_c) \quad (148.16)$$

The matrix element  $B_{3,5}$  (third row, fifth column) is equal to

$$B_{(\text{row}=3, \text{col}=5)} = \frac{\partial F_3}{\partial X_3} = 2 \times (X_3 - X_c) \quad (148.17)$$

Matrix **A** is a full matrix; however, the **B** matrix is of the nature

$$\mathbf{B} = \begin{bmatrix} X & X & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & X & X & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & X & X & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & X & X \end{bmatrix} \quad (148.18)$$

with the  $X$ s denoting nonzero elements.

In this example, the running indices  $i, j, k$  [see Eqs. (148.3), (148.4), and (148.2), respectively] have ranges

$$\begin{aligned} i &= 1, \dots, 8 \\ j &= 1, \dots, 3 \\ k &= 1, \dots, 4 \end{aligned} \quad (148.19)$$

denoting eight observations and three parameters to be estimated, all being part of four equations.

### 148.3 Model of Two Sets of Variables, Observations, and Parameters: The Mixed Model

---

The linearized version of Eq. (148.6), Eq. (148.7), demands closer inspection of the definition of all variables involved. The surveyor has the benefit of observed values and approximate values for the parameters; however, they should not be mixed. With subscripts we indicate the nature of the variable. Subscript  $b$  denotes "observed value," subscript 0 denotes "approximate value," and subscript  $a$  denotes the value of the observable quantity or the (unknown) parameter which perfectly fits the model  $F$ .  $\mathbf{V}$  denotes the residual, which is the value that must be added to the observed value  $\mathbf{L}_b$  to obtain the value for  $\mathbf{L}_a$  which perfectly fits the model  $F$ . Similarly,  $\mathbf{X}$  denotes the correction which needs to be added to the approximate values of the parameter  $\mathbf{X}_0$  to obtain the value for the parameter  $\mathbf{X}_a$  which perfectly fits the model  $F$ .

The following steps need to be taken to arrive at the linearized model [Eq. (148.7) or (148.26)], starting from Eq. (148.6) or (148.20), with the linearization around the Taylor point  $\{\mathbf{L}_0, \mathbf{X}_0\}$  :

$$F(\mathbf{L}_a, \mathbf{X}_a) = 0 \quad (148.20)$$

or

$$F(\mathbf{L}_b + \mathbf{V}, \mathbf{X}_0 + \mathbf{X}) = 0 \quad (148.21)$$

$$F(\mathbf{L}_0, \mathbf{X}_0) + \frac{\partial F}{\partial \mathbf{L}}(\mathbf{L}_a - \mathbf{L}_0) + \frac{\partial F}{\partial \mathbf{X}}(\mathbf{X}_a - \mathbf{X}_0) = 0 \quad (148.22)$$

$$\mathbf{W}_0 + \mathbf{B} \times (\mathbf{L}_a - \mathbf{L}_b + \mathbf{L}_b - \mathbf{L}_0) + \mathbf{A} \times (\mathbf{X}_a - \mathbf{X}_0) = 0 \quad (148.23)$$

$$\mathbf{W}_0 + \mathbf{B} \times (\mathbf{V} + \mathbf{L}) + \mathbf{A} \times \mathbf{X} = 0 \quad (148.24)$$

$$\mathbf{A}\mathbf{X} + \mathbf{B}\mathbf{V} + (\mathbf{W}_0 + \mathbf{B}\mathbf{L}) = 0 \quad (148.25)$$

or, arriving at the linearized form of Eq. (148.20),

$$\mathbf{A}\mathbf{X} + \mathbf{B}\mathbf{V} + \mathbf{W} = 0 \quad (148.26)$$

with

$$\begin{aligned} \mathbf{A} &= \frac{\partial F}{\partial \mathbf{X}} \\ \mathbf{B} &= \frac{\partial F}{\partial \mathbf{L}} \\ \mathbf{X} &= \mathbf{X}_a - \mathbf{X}_0 \\ \mathbf{V} &= \mathbf{L}_a - \mathbf{L}_b \\ \mathbf{L} &= \mathbf{L}_b - \mathbf{L}_0 \\ \mathbf{W} &= \mathbf{W}_0 + \mathbf{B}\mathbf{L} \end{aligned} \quad (148.27)$$

In words, the variables denote

$\mathbf{L}_a$  : adjusted observations

$\mathbf{L}_b$  : observed values

$\mathbf{V}$  : residuals

$\mathbf{L}_0$  : approximate observations

$\mathbf{L}$  : residual observations

$\mathbf{X}_a$  : adjusted parameters

$\mathbf{X}_0$  : approximate parameters (initial guess)

$\mathbf{X}$  : unknown correction to parameters

$W_0$  : misclosure vector

$W$  : misclosure vector

$A$  : design matrix (partial derivative matrix for the parameters)

$B$  : partial derivative matrix for the observations

Equation (148.26) is a set of  $r$  equations with  $u$  unknowns. This is an inconsistent set of equations and cannot be solved since often  $r > u$ . The method of Lagrangian multipliers provides a set of  $u$  equations with  $u$  unknowns under the conditions that the sum of the squared residuals is minimum. This leads to a minimum variance estimate for the vector  $\mathbf{X}$  [see, e.g., (Hamilton, 1964; Strang, 1986; and Strang, 1988)]. The unknown vector  $\mathbf{X}$  can be computed from the  $u$  equations with  $u$  unknowns, also known as the "normal equations." Without derivation, the normal equations are

$$\mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} \mathbf{X} + \mathbf{A}^T \mathbf{M}^{-1} \mathbf{W} = 0 \quad (148.28)$$

or, in short,

$$\mathbf{N} \mathbf{X} + \mathbf{U} = 0 \quad (148.29)$$

with

$$\mathbf{N} = \mathbf{A}^T \mathbf{M}^{-1} \mathbf{A} \quad (148.30)$$

$$\mathbf{U} = \mathbf{A}^T \mathbf{M}^{-1} \mathbf{W} \quad (148.31)$$

and

$$\mathbf{M} = \mathbf{B} \mathbf{B}^T \quad (148.32)$$

The model which mixes the observations  $L_a$  and the parameters  $X_a$  can be further generalized, assuming a statistical model (variance/covariance matrix for the observations; see also **Chapter 143**):

$$\Sigma_{L_b} = \begin{bmatrix} \sigma_{l_1 l_1} & \sigma_{l_1 l_2} & \cdots & \sigma_{l_1 l_n} \\ \sigma_{l_2 l_1} & \sigma_{l_2 l_2} & \cdots & \sigma_{l_2 l_n} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{l_n l_1} & \sigma_{l_n l_2} & \cdots & \sigma_{l_n l_n} \end{bmatrix} \quad (148.33)$$

Factoring a common (standard unit weight) constant out, we get

$$\Sigma_{L_b} = \sigma_0^2 \mathbf{Q}_{L_b} = \begin{bmatrix} q_{l_1 l_1} & q_{l_1 l_2} & \cdots & q_{l_1 l_n} \\ q_{l_2 l_1} & q_{l_2 l_2} & \cdots & q_{l_2 l_n} \\ \vdots & \vdots & \cdots & \vdots \\ q_{l_n l_1} & q_{l_n l_2} & \cdots & q_{l_n l_n} \end{bmatrix} \quad (148.34)$$

The  $\mathbf{Q}$  matrix is called the weight coefficient matrix. In terms of the weight matrix  $\mathbf{P}$ , we get

$$\Sigma_{L_b} = \sigma_0^2 \mathbf{P}_{L_b}^{-1} = \begin{bmatrix} p_{l_1 l_1} & p_{l_1 l_2} & \cdots & p_{l_1 l_n} \\ p_{l_2 l_1} & p_{l_2 l_2} & \cdots & p_{l_2 l_n} \\ \vdots & \vdots & \cdots & \vdots \\ p_{l_n l_1} & p_{l_n l_2} & \cdots & p_{l_n l_n} \end{bmatrix}^{-1} \quad (148.35)$$

For weighted observations, the matrix  $\mathbf{M}$ , Eq. (148.32), is simply replaced by

$$\mathbf{M} = \mathbf{B} \mathbf{P}^{-1} \mathbf{B}^T = \frac{1}{\sigma_0^2} \mathbf{B} \Sigma_{L_b} \mathbf{B}^T \quad (148.36)$$

The least squares estimate for the solution vector can be obtained from

$$\begin{aligned} X_a &= X_0 + X \\ &= X_0 - [\mathbf{A}^T \mathbf{M}^{-1} \mathbf{A}]^{-1} \mathbf{A}^T \mathbf{M}^{-1} \mathbf{W} \\ &= X_0 - [\mathbf{A}^T (\mathbf{B} \mathbf{P}^{-1} \mathbf{B}^T)^{-1} \mathbf{A}]^{-1} \mathbf{A}^T (\mathbf{B} \mathbf{P}^{-1} \mathbf{B}^T)^{-1} \mathbf{W} \end{aligned} \quad (148.37)$$

The variance/covariance matrix of the parameter vector  $\mathbf{X}$  or  $X_a$ , after applying the law of propagation of errors, can be shown to be equal to

$$\Sigma_{X_a} = \Sigma_X = \sigma_0^2 (\mathbf{A}^T \mathbf{M}^{-1} \mathbf{A})^{-1} \quad (148.38)$$

The relationship between the a priori variance of unit weight  $\sigma_0^2$  and the a posteriori variance of unit weight  $\hat{\sigma}_0^2$ , the latter being computed from

$$\hat{\sigma}_0^2 = \frac{\mathbf{V}^T \mathbf{P} \mathbf{V}}{r-u} \quad (148.39)$$

can be tested according to

$$\frac{\chi_{r-u;1-\alpha/2}^2}{r-u} < \frac{\hat{\sigma}_0^2}{\sigma_0^2} < \frac{\chi_{r-u;\alpha/2}^2}{r-u} \quad (148.40)$$

Equation (148.40) reflects the probability that the ratio of the variances will fall within the specified bounds, and is equal to  $1 - \alpha$  ( = 95% if  $\alpha = 0.05$  or 5%). This test gives insight between the expected observational precision and the overall behavior of the residuals once a model has

been "engaged." Rejection of the test may also lead to rejection of the particular model—for instance, in our example, if the observations were actually to lie on an ellipse rather than a circle.

From the "mixed model"  $F(L_a, X_a) = 0$ , two special models can be derived, which will be treated in the following two sections.

## 148.4 Observations as a Function of Parameters Only: The Model of Observation Equations

---

From the  $r$  (linearized) equations with  $u$  unknowns, a special case results if it so happens that each individual observation can be expressed as a function of the unknowns only. In this case we have  $L_a = F(X_a)$ , which can be derived directly by rewriting Eq. (148.26) as

$$-BL - BV = AX + W_0 \quad (148.41)$$

If  $B$  is assumed to be equal to a negative unit matrix (a square matrix with zeros and  $-1$  as diagonal elements), Eq. (148.41) becomes simply, with  $W_0$  being absorbed in the  $L$  vector [see Eqs. (148.44) and (148.45)],

$$L + V = AX \quad (148.42)$$

with

$$\begin{aligned} A &= \frac{\partial F}{\partial X} \\ X &= X_a - X_0 \\ V &= L_a - L_b \\ L &= L_b - L_0 \\ L_0 &= F(X_0) \end{aligned} \quad (148.43)$$

When you start from the unlinearized model  $L_a = F(X_a)$ , we find similarly

$$\begin{aligned} L_a &= F(X_a) \\ L_b + V &= F(X_0 + X_a - X_0) \\ L_b + V &= F(X_0) + \frac{\partial F}{\partial X} \cdot (X_a - X_0) \\ L_b + V &= L_0 + A \cdot X \end{aligned} \quad (148.44)$$

Bringing the  $L_0$  vector to the left-hand side,

$$\begin{aligned} L_b - L_0 + V &= A \cdot X \\ L + V &= A \cdot X \end{aligned} \quad (148.45)$$

In words, the variables denote

$L_a$  : adjusted observations

$L_b$  : observed values

$V$  : residuals

$L_0$  : approximate observations

$L$  : residual observations

$X_a$  : adjusted parameters

$X_0$  : approximate parameters (initial guess)

$X$  : unknown correction to parameters

$A$  : design matrix (partial derivative matrix for the parameters)

The normal equations simplify to

$$NX + U = 0 \quad (148.46)$$

with

$$N = A^T P A \quad (148.47)$$

$$U = -A^T P L \quad (148.48)$$

The solution vector  $X_a$  is equal to

$$\begin{aligned} X_a &= X_0 + X \\ &= X_0 + [A^T P A]^{-1} A^T P L \end{aligned} \quad (148.49)$$

The variance/covariance matrix of the parameter vector  $X$  or  $X_a$ , applying the law of propagation of errors, can be shown to be equal to

$$\Sigma_{X_a} = \Sigma_X = \sigma_0^2 (A^T P A)^{-1} \quad (148.50)$$

The relationship between the a priori variance of unit weight  $\sigma_0^2$  and the a posteriori variance of unit weight  $\hat{\sigma}_0^2$ , the latter being computed from

$$\hat{\sigma}_0^2 = \frac{V^T P V}{n-u} \quad (148.51)$$

Note that the denominator of Eq. (148.51) represents the degrees of freedom, which are equal to  $n - u$  since we deal with  $n$  (linearized) equations with  $u$  unknowns.

The method of observation equations is also known as "adjustment of indirect observations"

[see, e.g., (Mikhail, 1976)].

## Notation

The linearized observation equation, Eq. (148.45),

$$L + V = AX \quad (148.52)$$

appears under a variety of notations in the literature. For instance, Mikhail [1976] and Mikhail and Gracie [1981] use

$$l + \nu = -B\Delta \quad (148.53)$$

In the statistics literature one often finds

$$y - \varepsilon = X\beta \quad (148.54)$$

or, as in Gelb [1974],

$$z - \nu = Hx \quad (148.55)$$

Also, a more tensor-oriented notation may be found, as in Baarda [1967]:

$$x^i + \varepsilon^i = a_{\alpha}^i Y^{\alpha} \quad (148.56)$$

## 148.5 All Parameters Eliminated: The Model of Condition Equations

---

From the  $r$  (linearized) equations with  $u$  unknowns, another special case may be derived by eliminating the  $u$  unknowns from the linearized model, Eq. (148.26):

$$AX + BV + W = 0 \quad (148.57)$$

After elimination, we obtain  $r - u$  equations which reflect the mathematical relationship between the observations only. The equations are of the type

$$B'V + W' = 0 \quad (148.58)$$

They reflect the existing conditions between the observables—hence the name of the method. The classical example in surveying is that in a (not too large) triangle, the three measured angles have to sum to  $\pi$ . Another example concerns the loop closures between the leveled height differences, which have to sum to zero in each loop.



Starting from the nonlinearized model  $F(\mathbf{L}_a) = 0$ , we find

$$\begin{aligned}
 F(\mathbf{L}_a) &= 0 \\
 F(\mathbf{L}_b + \mathbf{V}) &= 0 \\
 F(\mathbf{L}_0 + \mathbf{L}_b - \mathbf{L}_0 + \mathbf{V}) &= 0 \\
 F(\mathbf{L}_0 + \mathbf{L} + \mathbf{V}) &= 0 \\
 F(\mathbf{L}_0) + \frac{\partial F}{\partial \mathbf{L}} \cdot (\mathbf{L} + \mathbf{V}) &= 0 \\
 \mathbf{W}_0 + \mathbf{B} \cdot \mathbf{L} + \mathbf{B} \cdot \mathbf{V} &= 0 \\
 \mathbf{W} + \mathbf{B}\mathbf{V} &= 0
 \end{aligned} \tag{148.59}$$

with

$$\begin{aligned}
 \mathbf{B} &= \frac{\partial F}{\partial \mathbf{L}} \\
 \mathbf{V} &= \mathbf{L}_a - \mathbf{L}_b \\
 \mathbf{L} &= \mathbf{L}_b - \mathbf{L}_0 \\
 \mathbf{W}_0 &= F(\mathbf{L}_0) \\
 \mathbf{W} &= \mathbf{W}_0 + \mathbf{B}\mathbf{L}
 \end{aligned} \tag{148.60}$$

In words, the variables denote

$\mathbf{L}_a$ : adjusted observations

$\mathbf{L}_b$ : observed values

$\mathbf{V}$ : residuals

$\mathbf{L}_0$ : approximate observations

$\mathbf{L}$ : residual observations

$\mathbf{W}_0$ : misclosure vector

$\mathbf{W}$ : misclosure vector

$\mathbf{B}$ : partial derivative matrix for the observations

## 148.6 An Example: Traversing

Various methods are presented in the survey literature [see, e.g., [Wolf and Brinker \(1995\)](#), chapters 12 and 13] to adjust data collected as part of a traverse. The (two-dimensional) least squares example discussed illustrates the formation of observation equations and can be adapted to reflect any traverse method. The example here involves the measurements of directions and distances along a traverse which stretches between two known points  $A$  and  $B$ , in coordinates  $\{x_A, y_A\}$  and  $\{x_B, y_B\}$ . In these terminal points two closing directions are measured to two known azimuth

markers, the points  $P$  and  $Q$ , respectively. The traverse is to solve for parameters such as the unknown coordinates of points 1 through  $n$ , the orientation unknowns for each point where directional measurements took place. Finally we assume the existence of an unknown scale factor  $\lambda$  between the distance measurement equipment and the distances implied by the known coordinates of points  $A$ ,  $B$ ,  $P$ , and  $Q$ .

## Directional Measurements $r_{ij}$

In point  $A$  two directions are measured, a backsight direction to  $P$  and a foresight direction to point 1. In point 1 two directions are measured, to the previous point  $A$  and to point 2; in point 2 directions are measured to 1 and 3; and so on. In the next to last point  $n$  a backsight direction is measured to point  $n - 1$  and a foresight direction to point  $B$ . In point  $B$  two directions are measured, to  $n$  and to the azimuth marker  $Q$ . The directional measurements can be written as a function of differences of azimuths  $Az_{ij}$  between points  $i$  and  $j$  and the unknown azimuths of the directional zero orientations  $o_i$ . The latter refer to the (unknown) azimuths of the "zero" reading on the horizontal circle of the theodolite. This so-called zero "reading" is not a reading on the circle at all, but the result of an analysis of a series of directional measurements taken in the point in question. The azimuths  $Az_{ij}$  are in turn a function of the coordinate unknowns  $\{x_i, y_i\}$  and  $\{x_j, y_j\}$ .

The first two observation equations, Eqs. (148.61) and (148.62), generated by the two directional measurements in point  $A$  are

$$r_{AP} = Az_{AP} - o_A = \arctan \left( \frac{x_P - x_A}{y_P - y_A} \right) - o_A \quad (148.61)$$

Note that  $o_A$  in Eq. (148.61) is the only unknown since we adopted the coordinates of  $A$  and  $P$ . However, the second (forward) direction in  $A(r_{A1})$  is dependent of three unknowns,  $o_A, x_1, y_1$ :

$$r_{A1} = Az_{A1} - o_A = \arctan \left( \frac{x_1 - x_A}{y_1 - y_A} \right) - o_A \quad (148.62)$$

The next two directional measurements, Eqs. (148.63) and (148.64), in point 1 are dependent of five unknowns,  $o_1, x_1, y_1, x_2$ , and  $y_2$ :

$$r_{1A} = Az_{1A} - o_1 = \arctan \left( \frac{x_A - x_1}{y_A - y_1} \right) - o_1 \quad (148.63)$$

For the foresight direction,

$$r_{12} = Az_{12} - o_1 = \arctan \left( \frac{x_2 - x_1}{y_2 - y_1} \right) - o_1 \quad (148.64)$$

In point  $i$ , in the middle of the traverse, we have the backsight direction

$$r_{i,i-1} = Az_{i,i-1} - o_i = \arctan \left( \frac{x_{i-1} - x_i}{y_{i-1} - y_i} \right) - o_i \quad (148.65)$$

and the foresight direction

$$r_{i,i+1} = Az_{i,i+1} - o_i = \arctan \left( \frac{x_{i+1} - x_i}{y_{i+1} - y_i} \right) - o_i \quad (148.66)$$

In the last point of the traverse, point  $B$ , we have

$$r_{Bn} = Az_{Bn} - o_B = \arctan \left( \frac{x_n - x_B}{y_n - y_B} \right) - o_B \quad (148.67)$$

and the foresight direction to azimuth marker  $Q$ ,

$$r_{BQ} = Az_{BQ} - o_B = \arctan \left( \frac{x_Q - x_B}{y_Q - y_B} \right) - o_B \quad (148.68)$$

So far, these directional measurements have generated  $2(n + 2)$  observational equations with  $(n + 2)$  directional unknowns and  $2n$  unknown coordinates, totaling  $3n + 2$  unknown parameters. The problem would not be solvable:  $2n + 4$  equations with  $3n + 2$  unknowns, a fact known to a surveyor because of geometric considerations alone. The addition of distance measurements will make traversing an efficient survey tool.

Note that one may have to add or subtract multiples of 360 degrees to keep the directional measurements  $r_{ij}$  between 0 and 360 degrees.

## Distance Measurements $s_{ij}$

The distance measurements with unknown (common) scale  $\lambda$  can be written in terms of the following functions. Since no distance measurements between point  $A$  and the azimuth marker  $P$  were assumed, we have in point  $A$  only one (forward) distance measurement

$$s_{A1} = \lambda[(x_1 - x_A)^2 + (y_1 - y_A)^2]^{1/2} \quad (148.69)$$

In point 1 we have a backsight distance  $s_{1A}$  and a foresight distance  $s_{12}$ , generating the following two observation equations:

$$s_{1A} = \lambda[(x_A - x_1)^2 + (y_A - y_1)^2]^{1/2} \quad (148.70)$$

and

$$s_{12} = \lambda[(x_2 - x_1)^2 + (y_2 - y_1)^2]^{1/2} \quad (148.71)$$

In point  $i$ , in the middle of the traverse, we have for the backsight distance

$$s_{i,i-1} = \lambda[(x_{i-1} - x_i)^2 + (y_{i-1} - y_i)^2]^{1/2} \quad (148.72)$$

and the foresight distance,

$$s_{i,i+1} = \lambda[(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2]^{1/2} \quad (148.73)$$

In the last point of the traverse, in point  $B$ , we have for the backsight distance

$$s_{Bn} = \lambda[(x_n - x_B)^2 + (y_n - y_B)^2]^{1/2} \quad (148.74)$$

We assumed that no foresight distance measurement to azimuth marker  $Q$  took place.

The distance measurements added  $2(n + 1)$  observation equations and only one additional unknown, the scale factor  $\lambda$ . The  $2n$  coordinates  $\{x_i, y_i\}$  were already included in the previous directional observation equations. Summing the equations for both the directions and distances, we have for this particular traverse  $2(n + 2) + 2(n + 1) = 4n + 6$  observations with  $(3n + 2) + 1 = 3n + 3$  unknowns. The degrees of freedom are in this case  $(4n + 6) - (3n + 3) = n + 3$  for the traverse under mentioned measurement conditions.

## 148.7 Dynamical Systems

---

Modern survey techniques incorporate the element *time* in two different aspects: first of all, in classical survey systems the assumption is made that we deal with stationary systems. That is, the random behavior of, say, the residuals is invariant of time. Secondly, during a kinematic survey—for instance, having an aircraft make aerial photographs equipped with a GPS receiver—we have to deal with estimating a vector of unknowns, say the position of the GPS antenna fixed on top of the airplane's fuselage, which is not independent of time anymore. We get for the variance/covariance matrix of the observations, Eq. (148.33), and the linearized version of Eq. (148.52), respectively

$$\Sigma_{L_b} = \Sigma_{L_b(t)}(t) \quad (148.75)$$

and

$$L(t) + V(t) = A(t)X(t) \quad (148.76)$$

The vector  $X(t)$  reflects the state of the parameters at epoch  $t$ . At the same time, a different model may be at hand which describes the rate of change of this vector (vector velocity),

$$\dot{\mathbf{X}}(t) = F(t)\mathbf{X}(t) + \varepsilon' \quad (148.77)$$

One may similarly be able to write the state of the vector  $\mathbf{X}$  at epoch  $(t + dt)$  as a function of the state at  $t$ , according to

$$\mathbf{X}(t + dt) = \Phi(t + dt, t)\mathbf{X}(t) + \varepsilon \quad (148.78)$$

The (Jacobian) matrix  $\Phi$  (see also section 143.7 elsewhere in this book) is called the state transition matrix. A new estimate  $\mathbf{X}(t + dt)$  is computed from a new measurement through Eq. (148.76) and through the use of the previous estimate  $\mathbf{X}(t)$  through Eq. (148.77) or (148.78).

Equations (148.75) through (148.78) lead to dynamical estimation models. One of the better-known estimation (filtering) models has been developed by Kalman and others. These models are developed in the so-called time domain (as opposed to the frequency domain). The reader is referred to the vast literature in this area [see, e.g., (Gelb, 1974) and others].

## References

- Baarda, W. 1967. Statistical concepts in geodesy. *Publ. Geodesy*. N.S. 2(4).  
 Gelb, A. 1974. *Applied Optimal Estimation*. The MIT Press, Cambridge, MA.  
 Hamilton, W. C. 1964. *Statistics in Physical Science: Estimation, Hypothesis Testing, and Least Squares*. Ronald Press, New York.  
 Mikhail, E. M. 1976. *Observations and Least Squares*. IEP-A Dun-Donnelley, New York.  
 Mikhail, E. M. and Gracie, G. 1981. *Analysis and Adjustment of Survey Measurements*. Van Nostrand Reinhold, New York.  
 Strang, G. 1986. *Introduction to Applied Mathematics*. Wellesley-Cambridge Press, Wellesley, MA.  
 Strang, G. 1988. *Linear Algebra and Its Applications*, 3rd ed. Saunders College Publishing, Fort Worth, TX.  
 Wolf, P. R. and Brinker, R. C. 1994. *Elementary Surveying*. HarperCollins College Publishers, New York.  
     Ch. 12: Traversing  
     Ch. 13: Traverse Computations

## Further Information

### Textbooks and Reference Books

For additional reading and more background, from the very basic to the advanced level, consult specific chapters in a variety of textbooks on geodesy, satellite geodesy, physical geodesy, surveying, photogrammetry, or statistics itself. The reader is referred to the following textbooks (in English):

- Bjerhammer, E. A. 1973. *Theory of Errors and Generalized Matrix Inverses*. Elsevier Scientific Publishing, New York.

- Bomford, G. 1980. *Geodesy*. Clarendon Press, Oxford.  
 Ch. 1: Triangulation, Traverse, and Trilateration (Field Work)  
 Ch. 2: Computation of Triangulation, Traverse, and Trilateration  
 Ap. D: Theory of Errors
- Carr, J. R. 1995. *Numerical Analysis for the Geological Sciences*. Prentice Hall, Englewood Cliffs, NJ.
- Davis, R. E., Foote, F. S., Anderson, J. M., and Mikhail, E. M. 1981. *Surveying: Theory and Practice*. McGraw-Hill Publishing Co., New York.  
 Ch. 2: Survey Measurements and Adjustments  
 Ap. B: Least-Squares Adjustment
- Escobal, P. R. 1976. *Methods of Orbit Determination*. John Wiley & Sons, New York.  
 Ap. IV: Minimum Variance Orbital Parameter Estimation
- Fraleigh, J. B. and Beauregard, R. A. 1987. *Linear Algebra*. Addison-Wesley Publishing Co., Reading, MA.  
 Ch. 5: Applications of Vector Geometry and of Determinants  
 Sec 5.2: The Method of Least Squares
- Heiskanen, W. A. and Moritz, H. 1967. *Physical Geodesy*. W. H. Freeman & Co., San Francisco.  
 Ch. 7: Statistical Methods in Physical Geodesy
- Hirvonen, R. A. 1965. *Adjustment by Least Squares in Geodesy and Photogrammetry*. Frederick Ungar Publishing Co., New York.
- Hofmann-Wellenhof, B., Lichtenegger, H., and Collins, J. 1995. *GPS: Theory and Practice*. Springer-Verlag, New York.  
 Ch. 9: Data Processing
- Kaula, W. M. 1966. *Theory of Satellite Geodesy: Applications of Satellites to Geodesy*. Blaisdell Publishing Co., Waltham, MA.  
 Ch. 5: Statistical Implications  
 Ch. 6: Data Analysis
- Koch, K. R. 1988. *Parameter Estimation and Hypothesis Testing in Linear Models*. Springer-Verlag, New York.
- Kraus, K. 1993. *Photogrammetry*. Ferd. Dümmlers Verlag, Bonn, Germany.  
 Ap. 4.2-1: Adjustment by the Method of Least Squares
- Leick, A. 1995. *GPS: Satellite Surveying*. John Wiley & Sons, New York.  
 Ch. 4: Adjustment Computations  
 Ch. 5: Least-Squares Adjustment Examples  
 Ap. B: Linearization  
 Ap. C: One-Dimensional Distributions
- McCormac, J. C. 1995. *Surveying*. Prentice Hall, Englewood Cliffs, NJ.  
 Ch. 2: Introduction to Measurements  
 Ch. 11: Traverse Adjustment and Area Computation
- Menke, W. 1989. *Geophysical Data Analysis: Discrete Inverse Theory*. Academic Press, San Diego.
- Moffitt, F. H. and Bouchard, H. 1992. *Surveying*. HarperCollins Publishers, New York.  
 Ap. A: Adjustment of Elementary Surveying Measurements by the Method of Least Squares

- Ap. B: The Adjustment of Instruments
- Mueller, I. I. and Ramsayer, K. H. 1979. *Introduction to Surveying*. Frederick Ungar Publishing Co., New York.
- Ch. 5: Adjustment Computation by Least Squares
- Papoulis, A. 1985. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill Publishing Co., New York.
- Tienstra, J. M. 1966. *Theory of Adjustment of Normally Distributed Observations*. Argus Publishing Co., Amsterdam.
- Uotila, U. A. 1985. *Adjustment Computations Notes*. Department of Geodetic Science and Surveying, Ohio State University, Columbus.
- Van'cek, P. and Krakiwsky, E. J. 1982. *Geodesy: The Concepts*. North-Holland Publishing Co., Amsterdam.
- Ch. 11: Classes of Mathematical Models
- Ch. 12: Least-Squares Solution of Overdetermined Models
- Ch. 13: Assessment of Results
- Ch. 14: Formulation and Solving of Problems
- Wolf, P. R. 1983. *Elements of Photogrammetry*. McGraw-Hill Publishing Co., New York.
- Ap. A: Random Errors and Least Squares Adjustment
- Wolf, P. R. 1987. *Adjustment Computations: Practical Least Squares for Surveyors*. Landmark Enterprises, Rancho Cordova, CA.
- Wolf, P. R. and Brinker, R. C. 1994. *Elementary Surveying*. HarperCollins College Publishers, New York.
- Ch. 2: Theory of Measurements and Errors
- Ap. C: Propagation of Random Errors and Least-Squares Adjustment

## Journals and Organizations

The latest results from research of least squares applications in geodesy, surveying, mapping, and photogrammetry are published in a variety of journals.

Two international magazines under the auspices of the International Association of Geodesy, both published by Springer-Verlag (Berlin/Heidelberg/New York), are:

*Bulletin Géodésique*

*Manuscripta Geodetica*

Geodesy- and geophysics-related articles can be found in:

American Geophysical Union, Washington, D.C.: *EOS* and *Journal of Geophysical Research*

Royal Astronomical Society, London: *Geophysical Journal International*

Statistical articles related to kinematic GPS can be found in:

Institute of Navigation: *Navigation*

Many national mapping organizations publish journals in which recent statistical applications in geodesy/surveying/mapping/photogrammetry are documented:

American Congress of Surveying and Mapping: *Surveying and Land Information Systems* and *Cartography and Geographic Information Systems*

American Society of Photogrammetry and Remote Sensing: *Photogrammetric Engineering & Remote Sensing*

American Society of Civil Engineers: *Journal of Surveying Engineering*

Deutscher Verein für Vermessungswesen: *Zeitschrift für Vermessungswesen*, Konrad Wittwer  
Verlag, Stuttgart

Canadian Institute of Geomatics: *Geomatica*

Royal Society of Chartered Surveyors: *Survey Review*

Institute of Surveyors of Australia: *Australian Surveyor*

Worth special mention are the following trade magazines:

*GPS World*, published by Advanstar Communications, Eugene, OR

*P.O.B. (Point of Beginning)*, published by P.O.B. Publishing Co., Canton, MI

*Professional Surveyor*, published by American Surveyors Publishing Co., Arlington,  
VA

*Geodetical Info Magazine*, published by Geodetical Information & Trading Centre bv., Lemmer,  
the Netherlands

National mapping organizations such as the U.S. National Geodetic Survey (NGS) regularly  
make software available (free and at cost). Information can be obtained from:

National Geodetic Survey

Geodetic Services Branch

National Ocean Service, NOAA

1315 East-West Highway, Station 8620

Silver Spring, MD 20910-3282



van Gelder, B. H. W. "Satellite Surveying"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

**149.1 A Satellite Orbiting the Earth****149.2 The Orbital Ellipse****149.3 Relationship between Cartesian and Keplerian Orbital Elements****149.4 Orbit of a Satellite in a Noncentral Force Field****149.5 The Global Positioning System (GPS)**

Positioning • Limiting Factors • Modeling and the GPS Observables • GPS Receivers • GPS Base Station • GIS, Heights, and High-Accuracy Reference Networks

**149.6 Gravity Field and Related Issues**

One-Dimensional Positioning: Heights and Vertical Control • Two-Dimensional Positioning: East/North and Horizontal Control • Three-Dimensional Positioning: Geocentric Positions and Full 3-D Control

**Boudewijn H. W. van Gelder**

*Purdue University*

The global positioning system (GPS) has become a tool used in a variety of fields both within and outside engineering.

Positioning has become possible with accuracies ranging from the subcentimeter level—for high-accuracy geodetic applications as used in state, national, and global geodetic networks, deformation analysis in engineering, and geophysics—to the hectometer level in navigation applications. Similar to the space domain, a variety of accuracy classes may be assigned to the time domain: GPS provides position and velocity determinations averaged over time spans from subseconds (instantaneous) to one or two days. Stationary applications of the observatory type are used in GPS tracking for orbit improvement.

---

**149.1 A Satellite Orbiting the Earth**

---

The path of an earth-orbiting satellite is similar to that of a planet around the sun. In history the solution to the motion of planets around the sun was found before its explanation. Johannes Kepler discovered certain regularities in the motions of planets around the sun. Through the analysis of his own observations and those made by Tycho Brahe, he formulated the following three laws:

*First law (1609):*

The orbit of each planet around the sun is an ellipse. The sun is in one of the two focal points.

*Second law* (1609):

The line from sun to planet sweeps out equal areas in equal time periods.

*Third law* (1611):

The ratio between the square of a planet's orbital period and the third power of its average distance from the sun is constant.

Kepler's third law leads to the famous equation

$$n^2 a^3 = GM \quad (149.1)$$

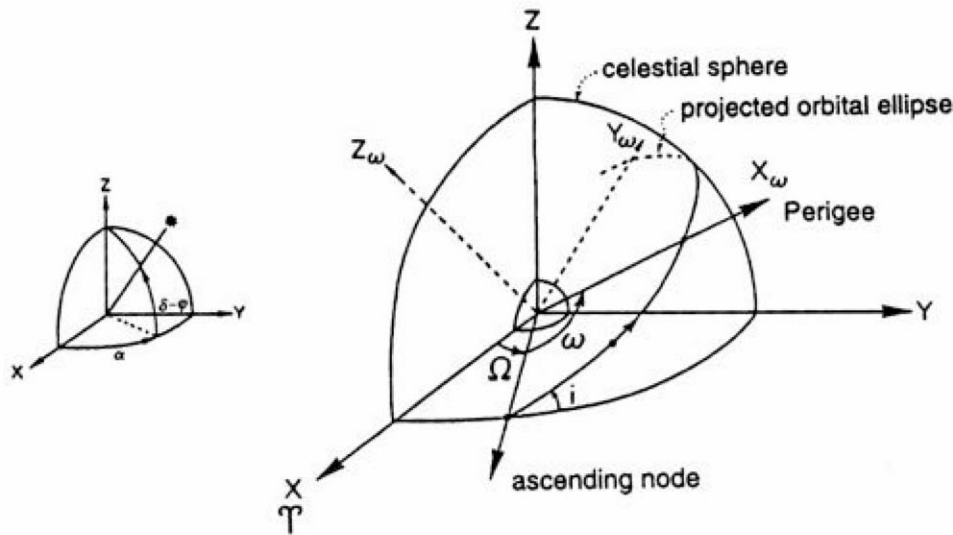
where  $n$  is the average angular rate and  $a$  the semimajor axis of the orbital ellipse.

In 1665-1666 Newton formulated his more fundamental laws of nature (which were only published after 1687) and showed that Kepler's laws follow from them.

## 149.2 The Orbital Ellipse

In a (quasi-)inertial frame the ellipse of an earth-orbiting satellite has to be positioned: the focal point will coincide with the center of mass (CoM) of the earth. Instead of picturing the ellipse itself, we project the ellipse on a celestial sphere centered at the CoM. On the celestial sphere we also project the earth's equator (see Fig. 149.1).

**Figure 149.1** Celestial sphere with projected orbital ellipse and equator.



The orientation of the orbital ellipse requires three orientation angles with respect to the inertial frame XYZ: two for the orientation of the plane of the orbit,  $\Omega$  and  $I$ , and one for the orientation of the ellipse in the orbital plane, for which one refers to the point of closest approach, the perigee,  $\omega$ .

$\Omega$  represents the right ascension ( $\alpha$ ) of the ascending node. The ascending node is the (projected) point where the satellite rises above the equator plane.

$I$  represents the inclination of the orbital plane with respect to the equator

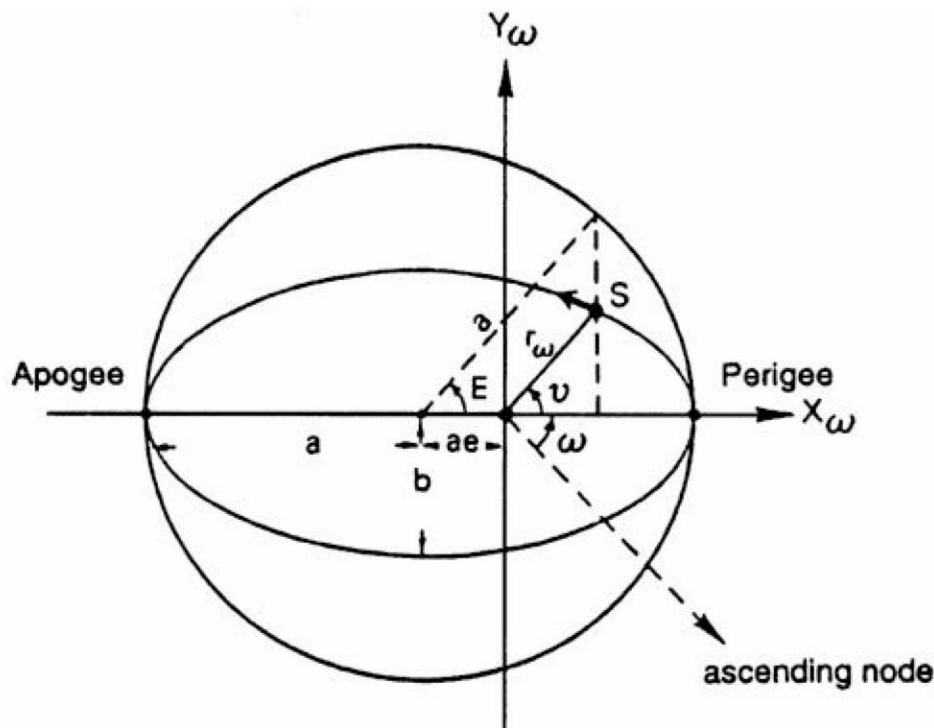
plane.

$\omega$  represents the argument of perigee—the angle from the ascending node (in the plane of the orbit) to the perigee (for planets, the perihelion), which is that point where the satellite (planet) approaches the closest to the earth (sun), or, more precisely, the center of mass of the earth (sun).

Similar to the Earth's ellipsoid, we define the orbital ellipse by a semimajor axis  $a$  and eccentricity  $e$ . In orbital mechanics it is unusual to describe the shape of the orbital ellipse by its flattening.

The position of the satellite in the orbital plane is depicted in Fig. 149.2.

**Figure 149.2** The position of the satellite ( $S$ ) in the orbital plane.



In Fig. 149.2 the major variables are defined as follows:

$a$  = the semimajor axis of the orbital ellipse

$b$  = the semiminor axis of the orbital ellipse

$e$  = the eccentricity of the orbital ellipse, with

$$e^2 = \frac{a^2 - b^2}{a^2} \quad (149.2)$$

$\nu$  = the true anomaly, sometimes denoted by  $f$

$E$  = the eccentric anomaly

The relation between the true anomaly and the eccentric anomaly can be derived

as

$$\tan\left(\frac{E}{2}\right) = \sqrt{\frac{1-e}{1+e}} \cdot \tan\left(\frac{\nu}{2}\right) \quad (149.3)$$

The Cartesian coordinates of the satellite position are

$$X_I = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \mathcal{R}_3(-\omega) \cdot \mathcal{R}_1(-I) \cdot \mathcal{R}_3(-\Omega) \cdot \begin{pmatrix} a \cdot (\cos E - e) \\ a \cdot \sqrt{1-e^2} \cdot \sin E \\ 0 \end{pmatrix} \quad (149.4)$$

In Eq. (149.4) the Cartesian coordinates are expressed in terms of the six so-called Keplerian elements:  $a$ ,  $e$ ,  $I$ ,  $\Omega$ ,  $\omega$ , and  $E$ . If we know the position of the satellite at an epoch  $t_0$  through  $\{a, e, I, \Omega, \omega, E_0\}$ , we are capable of computing the position of the satellite at an arbitrary epoch  $t$  through Eq. (149.4) if we know the relationship in time between  $E$  and  $E_0$ . In other words, how does the angle  $E$  increase with time?

We define an auxiliary variable (angle)  $M$  that increases linearly in time with the mean motion  $n [= (GM/a^3)^{1/2}]$  according to Kepler's third law. The angle  $M$ , the mean anomaly, may be expressed as function of time by

$$M = M_0 + n \cdot (t - t_0) \quad (149.5)$$

Through *Kepler's equation*,

$$M = E - e \cdot \sin E \quad (149.6)$$

the (time) relationship between  $M$  and  $E$  is given. Kepler's equation is the direct result of the enforcement of Kepler's second law ("equal area law").

Combining Eqs. (5) and (6) gives an equation that expresses the relationship between a given eccentric anomaly  $E_0$  (or  $M_0$  or  $\nu_0$ ) at an epoch  $t_0$  and the eccentric anomaly  $E$  at an arbitrary epoch  $t$ :

$$E - E_0 = e \cdot (\sin E - \sin E_0) + n \cdot (t - t_0) \quad (149.7)$$

The transformation is complete when we express the velocity vector  $\{\dot{X}, \dot{Y}, \dot{Z}\}$  in terms of those Keplerian elements. Differentiating Eq. (149.4) with respect to time and combining the position and velocity components into one expression, we get

$$[X_I | \dot{X}_I] = \begin{bmatrix} X | \dot{X} \\ Y | \dot{Y} \\ Z | \dot{Z} \end{bmatrix} \quad (149.8)$$

$$= \mathcal{R}_3(-\omega) \cdot \mathcal{R}_1(-I) \cdot \mathcal{R}_3(-\omega) \begin{bmatrix} a(\cos E - e) & | & -a\dot{E} \sin E \\ a\sqrt{1-e^2} \sin E & | & a\dot{E} \sqrt{1-e^2} \cos E \\ 0 & | & 0 \end{bmatrix} \quad (149.9)$$

The remaining variable  $\dot{E}$  is obtained through differentiation of Eq. (149.6):

$$\dot{E} = \frac{n}{1 - e \cdot \cos E} \quad (149.10)$$

Now all six Cartesian orbital elements (state vector elements) are expressed in terms of the six Keplerian elements.

## 149.3 Relationship between Cartesian and Keplerian Orbital Elements

---

To compute the inertial position of a satellite in a central force field, it is simpler to perform a time update in the Keplerian elements than in the Cartesian elements. The time update takes place through Eqs. (149.5), (149.6), and (149.7).

Schematically, the following procedure is to be followed:

$$\begin{array}{ll} t_0 : & \{X, Y, Z, \dot{X}, \dot{Y}, \dot{Z}\} \\ & \downarrow \quad \text{Conversion to Keplerian elements} \\ t_0 : & \{a, e, I, -\omega, E_0\} \\ & \downarrow \quad \text{Equation of Kepler, Eq. (149.7)} \\ t_1 : & \{a, e, I, -\omega, E_1\} \\ & \downarrow \quad \text{Conversion to Cartesian elements} \\ t_1 : & \{X, Y, Z, \dot{X}, \dot{Y}, \dot{Z}\} \end{array}$$

The conversion from Keplerian elements to state vector elements was treated in the previous section. For the somewhat more complicated conversion from position and velocity vector to Keplerian representation, the reader is referred to textbooks such as [Escobal, 1976]. Basically we "invert" Eqs. (149.8) and (149.9) by solving for the six elements  $\{a, e, I, -\omega, E\}$  in terms of the six state vector elements.

## 149.4 Orbit of a Satellite in a Noncentral Force Field

The equations of motion for a real satellite are more difficult than suggested by Eqs. (149.8) and (149.9). First of all, we do not deal with a central force field: the earth is not a sphere, nor does it have a radial symmetric density. Second, we deal with other forces, chiefly the gravity of the moon and the sun, atmospheric drag, and solar radiation pressure. Equations (149.8) and (149.9) get a more general meaning if we suppose that a potential function is being generated by the sum of the forces acting on the satellite:

$$V = V_c + V_{nc}^t + V_{\text{sun}}^t + V_{\text{moon}}^t + \cdots \quad (149.11)$$

where  $V_c$  is the central part of the earth's gravitational potential,

$$V_c = \mu/|X| \quad (149.12)$$

and  $V_{nc}^t$  is the noncentral and time-dependent part of the Earth's gravitational field. (The upper index  $t$  has been added to various potentials to reflect their time variability with respect to the inertial frame.)

The equations of motion to be solved are

$$\begin{aligned} \ddot{\mathbf{X}} &= \nabla(V_c + V_{nc}^t + V_{\text{sun}}^t + V_{\text{moon}}^t + \cdots) \\ &= \nabla V_c + \nabla V_{nc}^t + \nabla V_{\text{sun}}^t + \nabla V_{\text{moon}}^t + \cdots \end{aligned} \quad (149.13)$$

For the earth's gravitational field, we have (in an earth-fixed frame)

$$V_c + V_{nc} = \frac{\mu}{r} \left[ 1 + \sum_{l=1}^{\infty} \sum_{m=0}^l \left( \frac{a_e}{r} \right)^l \cdot (C_{lm} \cos m\lambda + S_{lm} \sin m\lambda) \cdot P_{lm}(\sin \phi) \right] \quad (149.14)$$

With Eq. (149.14) one is able to compute the potential at each point  $\{\lambda, \phi, r\}$  necessary for the integration of the satellite's orbit. The coefficients  $C_{lm}$  and  $S_{lm}$  of the spherical harmonic expansion are in the order of  $10^{-6}$  except for  $C_{20}$  ( $l = 2, m = 0$ ), which is about  $10^{-3}$ . This has to do with the fact that the earth's equipotential surface at mean sea level can be best approximated by an ellipsoid of revolution. One has to realize that the coefficients  $C_{lm}, S_{lm}$  describe the shape of the potential field and not the shape of the physical earth, despite a high correlation between the two.  $P_{lm}(\sin \phi)$  are the associated Legendre functions of the first kind, of degree  $l$  and order  $m$ ;  $a_e$  is some adopted value for the semimajor axis (equatorial radius) of the earth. [For values of  $a_e$ ,  $\mu (= GM)$ , and  $C_{20} (= -J_2)$ , see the following: [IAG, 1971](#); [IAG, 1980](#); [IAG, 1984](#); [IAG, 1988a](#); [IAG, 1988b](#); [DMA, 1988](#); [IERS, 1992](#); and [Cohen and Taylor, 1988](#).]

The equatorial radius  $a_e$ , the geocentric gravitational constant  $GM$ , and the dynamic form factor  $J_2$  characterize the earth as an ellipsoid of revolution with an equipotential surface.

If we restrict ourselves to the central part ( $\mu = GM$ ) and the dynamic flattening ( $C_{20} = -J_2$ ),

then Eq. (149.14) becomes

$$V_c + V_{nc} = \frac{\mu}{r} \left[ 1 + \frac{J_2 a_e^2}{2r^2} \cdot (1 - 3 \sin^2 \phi) \right] \quad (149.15)$$

with

$$\sin \phi = \sin \delta = \frac{z}{r} \quad (149.16)$$

where  $\phi$  is the latitude and  $\delta$  the declination (see [Fig. 149.2](#)).

The solution expressed in Keplerian elements shows periodic perturbations and some dominant secular effects. An approximate solution using only the latter effects is (position only)

$$X_I = \mathcal{R}_3[-(\dot{\Omega}_0 + \dot{\Omega} \Delta t)] \cdot \mathcal{R}_1(-I) \cdot \mathcal{R}_3[-(\omega_0 + \dot{\omega} \Delta t)] \cdot X_\omega \quad (149.17)$$

with

$$\Delta t = t - t_0 \quad (149.18)$$

$$\dot{\Omega} = -\frac{3}{2} \frac{J_2 a_e^2}{a^2 (1 - e^2)^2} n \cos I \quad (149.19)$$

$$\dot{\omega} = \frac{3}{2} \frac{J_2 a_e^2}{a^2 (1 - e^2)^2} n (2 - 2^{1/2} \sin^2 I) \quad (149.20)$$

$$n = n_0 \cdot \left[ 1 + \frac{3}{2} \frac{J_2 a_e^2 \sqrt{1 - e^2}}{a^2 (1 - e^2)^2} (1 - 1^{1/2} \sin^2 I) \right] \quad (149.21)$$

with

$$n_0 = \sqrt{\frac{GM}{a^3}} \quad (149.22)$$

Whenever



$I = 0^\circ$	we have	an equatorial orbit
$0^\circ < I < 90^\circ$		a direct orbit
$I = 90^\circ$		a polar orbit
$90^\circ < I < 180^\circ$		a retrograde orbit
$I = 180^\circ$		a retrograde equatorial orbit

Equation (149.19) shows that the ascending node of a direct orbit slowly drifts to the west. For a satellite at about 150 km above the earth's surface, the right ascension of the ascending node decreases about  $9^\circ$  per day.

The satellites belonging to the global positioning system have an inclination of about  $55^\circ$ . Their nodal regression rate is about  $-0.04187^\circ$  per day.

## 149.5 The Global Positioning System (GPS)

---

The Navstar GPS space segment consists of 1 Block I satellite, 9 Block II satellites, and 15 Block IIA satellites as of October 1994 [GPS World, 1994]. This means that the full Block II satellite constellation, in six orbital planes at an height of about 20 000 kilometers, is complete. With this number of satellites, 3-D positioning is possible every hour of the day. However, care must be exercised, since an optimum configuration for 3-D positioning is not available on a full day's basis.

In the meantime, GPS receivers, ranging in cost between \$300 and \$20 000 , are readily available. Over 100 manufacturers are marketing receivers, and the prices are still dropping! Magazines such as *Geodetical Info Magazine*, *GPS World*, *P.O.B.*, and *Professional Surveyor* regularly publish information on the latest models [see, for example, the recent GPS equipment surveys in (Reilly, 1994) and in (Chan and Schorr, 1994)].

GPS consumer markets have been rapidly expanding. In the areas of land, marine, and aviation navigation; of precise surveying; of electronic charting; and of time transfer the deployment of GPS equipment seems to have become indispensable. This holds for military as well as civilian users.

### Positioning

Two classes of positioning are recognized: Standard Positioning Service (SPS) and Precise Positioning Service (PPS). In terms of positional accuracies one has to distinguish between SPS with and without selective availability (SA) on the one hand and PPS on the other. Selective availability deliberately introduces clock errors and ephemeris errors in the data being broadcast by the satellite. The current accuracy of the (civil) signal without SA is in the order of 20-40 m. With SA-implemented SPS accuracy is degraded to 100 m. At the time of this writing (October 1994) the satellite constellation guarantees 100-m SPS accuracy to civilian and commercial users 95% of the time.

In several applications GPS receivers are interfaced with other positioning systems, such as inertial navigation systems (INS), hyperbolic systems, and even automatic braking systems (ABS) in cars.

GPS receivers in combination with various equipment are able to provide the answers to such general questions as [Wells and Kleusberg, 1990]:

All of these questions may refer to an observer either at rest (*static* positioning) or in motion (*kinematic* positioning). The questions may be answered immediately (*real-time* processing, often misnamed DGPS, which stands for differential GPS) or after the fact (*batch* processing).

The various options are summarized in Table 149.1 [Wells and Kleusberg, 1990]. The accuracies for time dissemination are summarized in Table 149.2.

**Table 149.1** Accuracies of Various GPS Positioning Modes

Absolute positioning	
SPS with SA	100 m
SPS without SA	40 m
PPS	20 m
Relative differential positioning	
Differential SPS	10 m
Carrier-smoothed code	2 m
Ambiguity-resolved carrier	10 cm
Surveying between fixed points	1 mm to 10 cm

*Note:* The accuracy of differential modes is dependent on interreceiver distance.

**Table 149.2** Accuracies of Various GPS Time Dissemination Modes

Time and time interval	
With SA	500 ns
Without SA, correct position	100 ns
Common mode, common view	< 25 ns

## Limiting Factors

Physics of the environment, instruments, broadcast ephemeris, and the relative geometry between orbits and networks all form limiting factors on the final accuracy of the results. Dilution of precision (DOP) is used as a scaling factor between the observational accuracy and positioning accuracy [Wells, 1986]. For reasons of safety and accuracy one should avoid periods in which the DOP factor is larger than 6.

The atmosphere of the earth changes the speed and the geometrical path of the electromagnetic signals broadcast by the GPS satellites. In the uppermost part of the atmosphere (the ionosphere) charged particles vary in number spatially as well as temporally. The so-called ionospheric refraction errors may amount to several tens of meters. Since this effect is frequency dependent,

the first-order effect can be largely eliminated by the use of dual-frequency receivers. The lower part of the atmosphere (the troposphere) causes refraction errors of several meters. Fortunately, the effect can be modeled rather well by measuring the atmospheric conditions at the measuring site.

GPS instruments are capable of measuring one or a combination of the following signals:

- C/A code, with an accuracy of a few meters

- P code, with an accuracy of a few decimeters

- Carrier phase, with an accuracy of a few millimeters

In addition to this measurement noise, receiver clock errors have to be modeled as to-be-solved-for parameters. It is this synchronization parameter between satellite time and receiver time that makes it necessary to have at least four satellites in view in order to get a 3-D fix.

Because of the high frequency of the GPS signals, multipath effects may hamper the final accuracy; the signal arriving at the receiver through a reflected path may be stronger than the direct signal. By careful antenna design and positioning, multipath effects are reduced. The phase center of the antenna needs to be carefully calibrated with respect to a geometric reference point on the antenna assembly. However, because of the varying inclination angle of the incoming electromagnetic signals, effects of a moving phase center may be present at all times.

Information on the orbit of the satellite, as well as the orbital geometry relative to the network/receiver geometry, influences the overall positioning accuracy. The information the satellite broadcasts on its position and velocity is necessarily the result of a process of prediction. This causes the broadcast ephemeris to be contaminated with extrapolation errors. Typical values are

- Radial Error: about 5 m

- Across-track error: about 10 m

- Along-track error: about 15 m

Also, the on-board satellite clock is not free of errors. Orbital and satellite clock errors can be largely taken care of by careful design of the functional model.

As mentioned before, deliberate contamination of the broadcast ephemeris and satellite time degrades the system accuracy to several tens of meters. PPS users are able to use the P code on two frequencies and have access to the SA code. Consequently, they are capable of eliminating the ionospheric effects and of removing deliberately introduced orbital and satellite clock errors. The resulting measurement error will be about 5 m. SPS users are able to use the C/A code (on one frequency only); they do not have access to the SA code. Consequently, the ionospheric error can only be roughly modeled, and these users are stuck with the deliberate errors. The resulting measurement error may be as large as 50 m.

Translocation techniques such as having a stationary receiver continuously supporting the other (roving) receivers will reduce the measurement error to well below the 10 m for SPS users.

Differencing techniques applied to the carrier phase measurements are successfully used to

eliminate a wide variety of errors, provided the receivers are not too far apart. In essence, two close-by receivers are influenced almost equally by (deliberate) orbital errors and by part of the atmosphere error. Differencing of the measurements of both receivers will cancel a large portion of the first-order effects of these errors.

## Modeling and the GPS Observables

Developing well-chosen functional models  $F$ , relating the GPS measurements  $\mathbf{L}$  to the modeled parameters  $X$ , enables users to fit GPS perfectly to their needs. A wide class of applications, from monitoring the subsidence of oil rigs in the open sea to real-time navigation of vehicles collecting geoinformation, belong to the range of possibilities opened up by the introduction of GPS.

In satellite geodesy, one has traditionally modeled the state of the satellite—a vector combining the positional ( $X$ ) and the velocity ( $\dot{X}$ ) information. Nowadays GPS provides geodesists, or geoscientists in general, with a tool by which the state of the observer, also in terms of position ( $x$ ) and velocity ( $\dot{x}$ ), can be determined with high accuracy and often in real time. The GPS satellite geodetic model has evolved to

$$\mathbf{L} = F(X, \dot{X}, x, \dot{x}, \mathbf{p}, t) \quad (149.23)$$

with

$\mathbf{L}$  = C/A code, P code, or carrier phase observations,  $i = 1, \dots, n$

$X$  = 3-D position of the satellite at epoch  $t$

$\dot{X}$  = 3-D velocity of the satellite at epoch  $t$

$x$  = 1-D, 2-D, or 3-D position of the observer at epoch  $t$

$\dot{x}$  = 1-D, 2-D, or 3-D velocity of the observer at epoch  $t$

$\mathbf{p}$  = vector of modeled (known or unknown) parameters,  $j = 1, \dots, u$

$t$  = epoch of measurement taking

Various differencing operators  $D^k$ , up to order 3, are applied to the original observations in order to take full benefit of the GPS measurements. The difference operator  $D^k$  may be applied in the observation space spanned by the vector  $\mathbf{L}$  [Eq. (149.24)], or the  $D^k$  operator may be applied in the parameter space  $x$  [Eq. (149.25)]:

$$D^k[\mathbf{L}] = D^k[F(X, \dot{X}, x, \dot{x}, \mathbf{p}, t)] \quad (149.24)$$

$$\mathbf{L} = F(X, \dot{X}, D^1(x, \dot{x}), \mathbf{p}, t) \quad (149.25)$$

The latter method is sometimes referred to as "delta positioning." This is a difficult way of saying that one may either construct so-called derived observations from the original observations by differencing techniques, or model the original observations, compute parameters (e.g., coordinates) in this way, and subsequently start a differencing technique on the results obtained from the roving receiver and the base receiver.

## Pseudo Ranging

We restrict the discussion to the C/A-based pseudo-range observables. The ranges are called "pseudo" because this technique is basically a one-way ranging technique with two independent clocks: the offset  $\delta t_E^S$  between the satellite clock  $S$  and the receiver clock  $E$  yields one additional parameter to be solved for. Writing the observation equation in the earth-fixed reference frame, we have

$$pr = \sqrt{(x^S - x_E)^2 + (y^S - y_E)^2 + (z^S - z_E)^2} - c \cdot \delta t_E^S \quad (149.26)$$

Inspection of the partials,

$$\frac{\partial pr}{\partial x^S} = \frac{x^S - x_E}{pr} = -\frac{\partial pr}{\partial x_E} \quad (149.27)$$

$$\frac{\partial pr}{\partial y^S} = \frac{y^S - y_E}{pr} = -\frac{\partial pr}{\partial y_E} \quad (149.28)$$

$$\frac{\partial pr}{\partial z^S} = \frac{z^S - z_E}{pr} = -\frac{\partial pr}{\partial z_E} \quad (149.29)$$

$$\frac{\partial pr}{\partial(\delta t_E^S)} = -c \quad (149.30)$$

reveals that:

- The coordinates of the stations are primarily obtained in a frame determined by the satellites or, better, by their broadcast ephemeris.
- Partial derivatives evaluated for neighboring stations are practically identical, so the coordinates of one station need to be adopted.

## Phase (Carrier Wave) Differencing

For precise engineering applications the phase of the carrier wave is measured. Two wavelengths are available in principle:

$$\begin{aligned} L_1: \quad \lambda_1 &= \frac{c}{f_1} \quad \text{with } f_1 = 1.57542 \text{ GHz} \\ &\cong 19.0 \text{ cm} \end{aligned} \quad (149.31)$$

and

$$L_2: \quad \lambda_2 = \frac{c}{f_2} \quad \text{with } f_2 = 1.22760 \text{ GHz} \quad (149.32)$$

$$\cong 24.4 \text{ cm}$$

For phase measurements the following observation equation can be set up:

$$\text{Range} = \phi + N \cdot \lambda_l, \quad l = 1, \dots, 2 \quad (149.33)$$

or

$$\Phi_E^S = \sqrt{(x^S - x_E)^2 + (y^S - y_E)^2 + (z^S - z_E)^2} - N_E^S \cdot \lambda_l \quad (149.34)$$

where  $\Phi_E^S$  is the phase observable in a particular  $S$ - $E$  combination, and  $N_E^S$  is the integer multiple of wavelengths in the range: the ambiguity. Phase measurements can be done with probably 1% accuracy. This yields an observational accuracy—in case the ambiguity  $N$  can be properly determined—in the millimeter range!

Using the various differencing operators on the phase measurements:

$D^{k=1}$  yields single differences:

- Between receiver differences,  $\Delta\Phi$ , eliminating or reducing satellite-related errors
- Between satellite differences,  $\nabla\Phi$ , eliminating or reducing receiver-related errors
- Between epoch differences,  $\delta\Phi$ , eliminating phase ambiguities per satellite-receiver combination

$D^{k=2}$  yields double differences:

- Between receiver/satellite differences,  $\nabla\Delta\Phi$ , eliminating or reducing satellite- and receiver-related errors, and so forth

$D^{k=3}$  yields triple differences:

- Between epoch/receiver/satellite differences,  $\delta\nabla\Delta\Phi$ , eliminating or reducing satellite/receiver-related errors, and ambiguities

Receivers that use carrier wave observations have, in addition to the electronic components that do the phase measurements, a counter that counts the complete cycles between selected epochs. GPS analysis software uses the triple differences to detect and possibly repair cycle slips occurring during loss of lock.

Design specifications and receiver selection are dependent on the specific project accuracy requirements. In the U.S. the Federal Geodetic Control Committee has adopted various specifications [FGCC, 1989].

## GPS Receivers

A variety of receivers are on the market. Basically, they can be grouped in the four classes listed in [Table 149.3](#).

**Table 149.3** Accuracy Grades of Civilian/Commercial GPS Receivers

Navigation grade	40–100 m	C/A code, in stand-alone mode
Mapping (GIS) grade	2–5 m	C/A code, in differenced mode
Surveying grade	1–2 cm within 10 km	C/A code + phase, differenced
Geodesy grade	5–15 mm over any distance	C/A + P code + phase, differenced

The observations of the first three types of receivers are subjected to models that can be characterized as *geometric* models. The position of the satellite is considered to be known, based mostly on the information taken from the broadcast ephemeris. The known positions are, of course, not errorless. First of all, the positions are predicted and thus contain errors because of an extrapolation process in time. Second, the positions being broadcast may be corrupted by intentional errors (due to SA). Differencing techniques are capable of eliminating most of the error if the separation between base station and roving receiver is not too large.

Millimeter-accurate observations from geodesy-grade receivers are often subjected to analysis through models of the *dynamic* type. Software packages containing dynamic models are very elaborate and allow for some kind of orbit improvement estimation process.

Reilly [1994] and Chan and Schorr [1994] list overviews of recent GPS receivers and supporting software packages.

## GPS Base Station

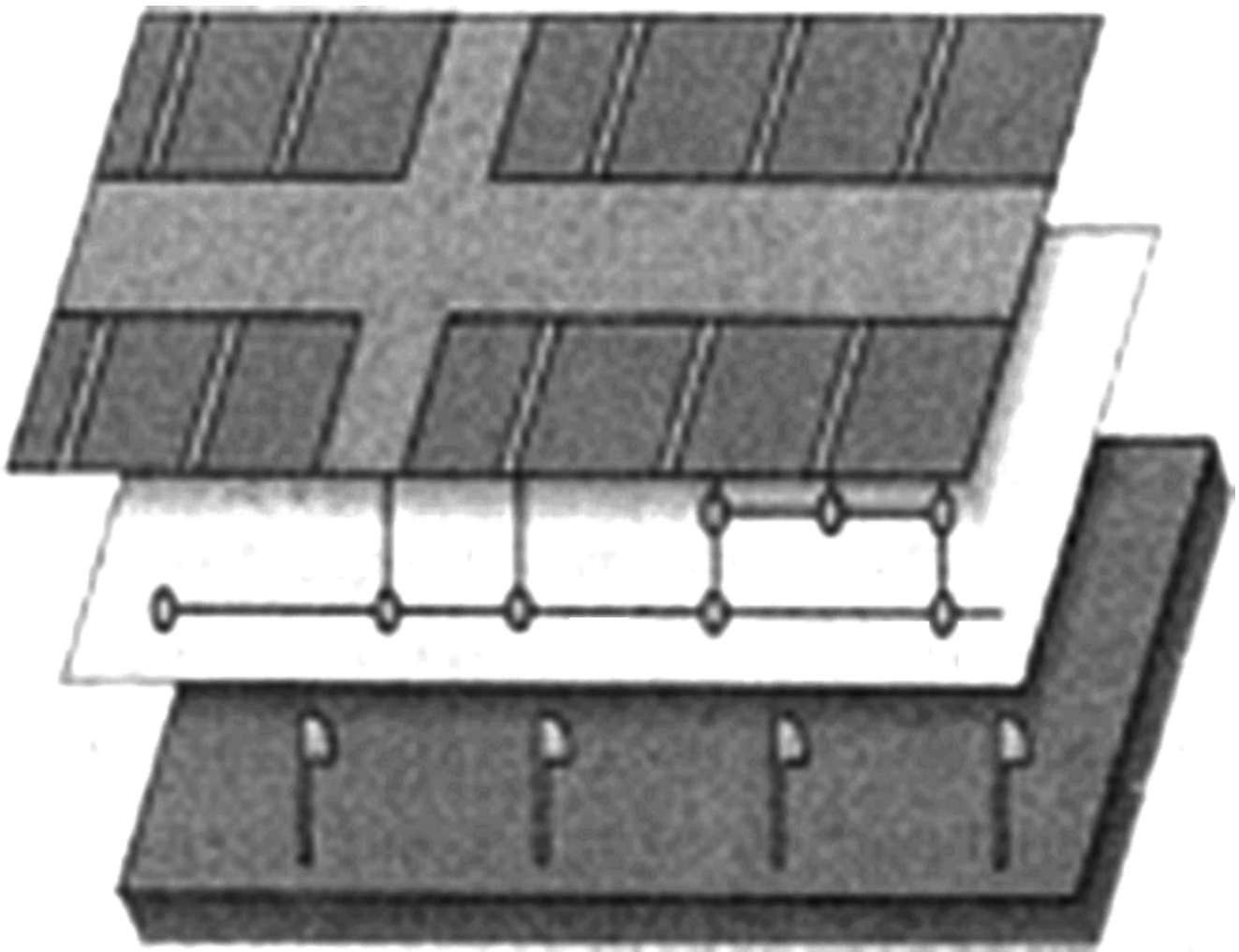
GPS, like most other classical survey techniques, has to be applied in a differential mode if one wants to obtain reliable relative positional information. This implies that, for most applications of GPS in geodesy—surveying and mapping, photogrammetry, GIS, and so forth—one has to have at least two GPS receivers at one's disposal. If one of the receivers occupies a known location during an acceptable minimum period, then one may obtain accurate coordinates for the second receiver *in the same frame*. In surveying/geodesy applications it is preferable to include three stations with known horizontal coordinates and at least four with known vertical (orthometric) heights. In most GIS applications a receiver is left at one particular site. This station serves as a so-called base station.

## GIS, Heights, and High-Accuracy Reference Networks

In order to reduce influences from satellite-related errors and atmospheric conditions in geodesy, surveying, and geographic information systems (GIS) applications, GPS receivers are operated in a differential mode. Whenever the roving receiver is not too far from the base station receiver, errors at high altitudes (satellite and atmosphere) are more or less canceled if the "fix from the field" is

differenced with the "fix from the base."

The Washington editor of *GPS World*, Hale Montgomery, writes, "As a peripheral industry, the reference station business has grown almost into an embarrassment of riches, with stations proliferating nationwide and sometimes duplicating services." William Strange, Chief Geodesist at The National Geodetic Survey (NGS), is quoted as saying, "Only about 25 full-service, fixed stations would be needed to cover the entire United States" [Montgomery, 1993]. A group of the interagency Federal Geodetic Control subcommittee has compiled a list of about 90 base stations operating on a more or less permanent basis. If all GIS/GPS base stations being planned or in operation are included, the feared proliferation will be even larger. From the point of view of the U.S. tax-paying citizen, "duplication of services" may be wasteful; on the other hand, decentralization of services may often be more cost-effective than all-encompassing projects run by even more all-encompassing agencies.



#### GEOGRAPHIC INFORMATION SYSTEMS: THE BASICS

Geographic Information Systems are databases that store both location information and descriptive data. This information can be displayed in a computerized map format. Most GIS databases store features in a layer format so that different types of information can be sorted, searched, edited, displayed, and output by type.



In this simple example, various types of data that relate to an urban asset management database are depicted. However, any type of geographic data could be stored and organized this way in a GIS. Here, three layers of a database are shown: a road map, an underground power line map on the middle layer, and street light poles on the third layer. Both the location data and descriptive information, such as road type, pole condition, and cable type, can be gathered with GPS and easily transferred to this GIS.

### **How GPS Works**

The Global Positioning System is a constellation of 24 satellites that orbit the earth at an altitude of 20 200 kilometers, constantly emitting GPS signals. GPS receivers on earth calculate their positions by making distance measurements to four or more satellites. Individual distance measurements to each satellite are determined by analyzing the time it takes for a signal to travel from a satellite, whose location is known through monitoring, to a GPS receiver. Using some relatively simple geometry, the receiver determines its position. GPS mapping systems utilize Differential GPS (DGPS) techniques to obtain even better accuracy, in the decimeter to 5-meter range. (Courtesy of Trimble Navigation.)

From the geodetic point of view, the duplication of services (base station-generated fixes in the field) will proliferate the coordinate fields and the reference frames they supposedly are tied to. The loss of money and effort in the years to come in trying to make sense out of these most likely nonmatching point fields may be far larger than the money lost in "service-duplicating" base stations. If we are not careful, the "coordinate-duplicating" base stations will create a chaos among GIS-applying agencies.

Everyone is convinced of the necessity to collect GIS data in one *common frame*. Formerly, the GIS community was satisfied with positions of 3- to 5-m accuracy. Manufacturers are aggressively marketing GPS/GIS equipment with 0.5-m accuracy (with a price tag of 20 000 per receiver, and remember you need at least two). The increased demand for accuracy requires that a reference frame be in place that lasts at least two decades. This calls for a consistent reference that is one and probably two orders of magnitude more accurate than presently available.

The accuracy of the classical horizontal control was on the order of one part in 100 000 (1 cm over 1 km). GPS is a survey tool with an accuracy of one part per 1 000 000 (1 cm over 10 km). Many states have put new high-accuracy reference networks (HARNs) in place to accommodate the accuracy of GPS surveys. Even for GIS applications where 0.5-m accuracies are claimed for the roving receivers, one may speak of 1 ppm surveys whenever those rovers operate at a distance of 500 km from their base station.

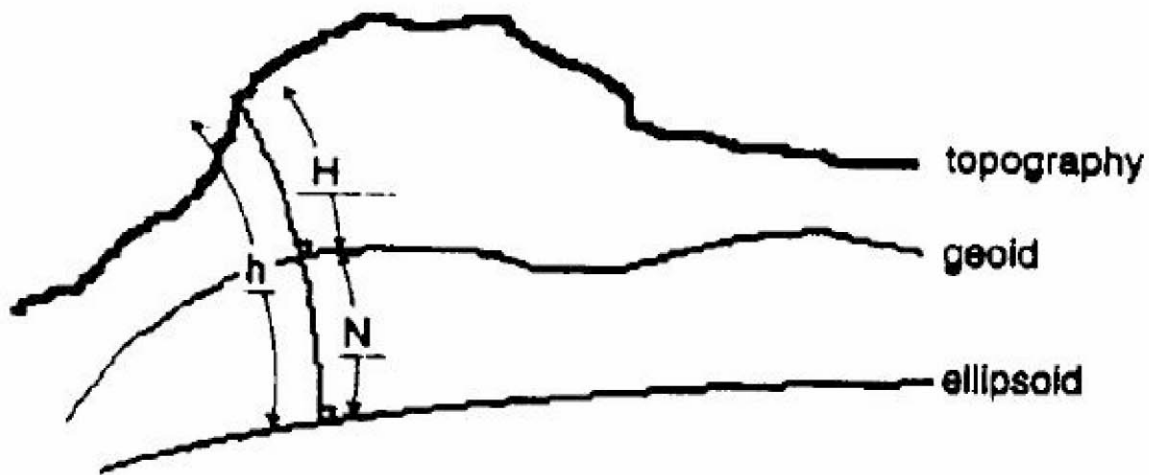
It should not be forgotten that GPS is a geometric survey tool yielding results in terms of earth-fixed coordinate differences  $x_{ij}$ . From these coordinate differences expressed in curvilinear coordinates we obtain at best somewhat reproducible ellipsoidal height differences. These height differences are *not* easily converted to orthometric height differences of equal accuracy. The latter height differences are of interest in engineering and GIS applications, as discussed in the next section.

## 149.6 Gravity Field and Related Issues

### One-Dimensional Positioning: Heights and Vertical Control

Some of the most accurate measurements surveyors are able to make are the determinations of height differences by spirit leveling. Since a leveling instrument's line of sight is tangent to the potential surface, one may say that leveling actually determines the height differences with respect to equipotential surfaces. If one singles out one particular equipotential surface at mean sea level (the so-called geoid), then the heights a surveyor determines are actually *orthometric heights* (see Fig. 149.3).

**Figure 149.3** Orthometric heights.



Leveling in a closed loop is a check on the actual height differences not in a metrical sense but in a potential sense: the distance between equipotential surfaces varies due to local gravity variations.

In spherical approximation the potential at a point A is

$$V = -\frac{GM}{r} = -\frac{GM}{R + h} \quad (149.35)$$

The gravity is locally dependent on the change in the potential per height unit,  
or

$$\frac{dV}{dr} = g = \frac{GM}{r^2} \quad (149.36)$$

The potential difference  $dV$  between two equipotential surfaces is

$$dV = g \cdot dr \quad (149.37)$$

Consequently, if one levels in a loop, one has

$$\sum dV = \sum g \cdot dr = 0 \quad (149.38)$$

or

$$\oint dV = \oint g \cdot dr = 0 \quad (149.39)$$

This implies that for each metrically leveled height difference  $dr$ , one has to multiply this difference by the local gravity. Depending on the behavior of the potential surfaces in a certain area and the diameter of one's project, one has to "carry along a gravimeter" while leveling.

The variations of local gravity vary depending on the geology of the area. Variations in the order  $10^{-7}g$  may yield errors as large as 10 mm for height differences in the order of several hundred meters. For precise leveling surveys ( $\leq 0.1$  mm/km), gravity observations have to be made with an interval of:

2 to 3 km in relatively "flat" areas

1 to 2 km in hilly terrain

1/2 to 1 1/2 km in mountainous regions

For more design criteria on leveling and gravity surveys, see [FGCC, 1989] and Table 149.4.

**Table 149.4** FGCC Vertical Control Accuracy Standards (Differential Leveling)

First Order		
Class I		$b^* < 0.5$
Class II		0.7
Second Order		
Class I		1.0
Class II		1.3
Third Order		
		2.0

$*b = S/\sqrt{d}(\text{mm}/\sqrt{\text{km}})$ , where

S = standard deviation of elevation difference between control points (mm)

d = approximate horizontal distance along leveled route (km)

GPS surveys yield at best ellipsoidal height differences. These are rather meaningless from the engineering point of view. Therefore, extreme caution should be exercised when GPS height

information, even after correction for geoidal undulations, is to be merged with height information from leveling. For two different points  $i$  and  $j$ ,

$$h_i = H_i + N_i \quad (149.40)$$

$$h_j = H_j + N_j \quad (149.41)$$

Subtracting Eq. (149.40) from (149.41), we find the ellipsoidal height differences  $h_{ij}$  (from GPS) in terms of the orthometric height differences  $H_{ij}$  (from leveling) and the geoidal height differences  $N_{ij}$  (from gravity surveys):

$$h_{ij} = H_{ij} + N_{ij} \quad (149.42)$$

where

$$h_{ij} = h_j - h_i \quad (149.43)$$

$$H_{ij} = H_j - H_i \quad (149.44)$$

$$N_{ij} = N_j - N_i \quad (149.45)$$

For instance, with the National Geodetic Survey's software program GEOID93 geoidal height differences are as accurate as 10 cm over 100 km for the conterminous U.S. For GPS leveling, this means that GPS may compete with third-order leveling as long as the stations are more than 5 km apart.

In principle, any equipotential surface can act as a vertical datum. The National Geodetic Vertical Datum of 1929 (NGVD29) is not a true mean sea-level datum. Problems may arise in merging GPS heights, gravity surveys, and orthometric heights referring to NGVD29. Heights referring to the NGVD88 datum will be more suitable for use with GPS surveys. In the U.S. about 600 000 vertical control stations are in existence.

## Two-Dimensional Positioning: East/North and Horizontal Control

In classical geodesy the measurements in height (leveling) had to be separated from the horizontal measurements (directions, angles, azimuths, distances). To allow for the curvature of the earth and the varying gravity field, the horizontal observations were reduced first to the geoid, taking into account the orthometric heights. Subsequently, it was desired to take advantage of geometrical properties between the once-reduced horizontal observations, and the observations had to be reduced once more, from the geoid to the ellipsoid. An ellipsoid approximates the geoid up to 0.01%; the variations of the geoid are nowhere larger than 150 m. On the ellipsoid, which is a precise mathematical figure, one could check, for instance, whether the sum of the three angles equaled a prescribed value.

So far, geodesists have relied on a biaxial ellipsoid of revolution. A semimajor axis  $a_e$  and a

semiminor axis  $b_e$  define the dimensions of the ellipsoid. Rather than using this semiminor axis, one may specify the flattening of the ellipsoid:

$$f = \frac{a_e - b_e}{a_e} \approx \frac{1}{298.257 \dots} \quad (149.46)$$

For a semimajor axis of about 6378.137 km, this implies that the semiminor axis is  $6378.137/298.257 \approx 22$  kilometers shorter than  $a_e$ .

Distance measurements need to be reduced to the ellipsoid. Angular measurements made with theodolites, total stations, and other instruments need to be corrected for several effects:

- The direction of local gravity does not coincide with the normal to the ellipsoid.
- The direction of the first axis of the instrument coincides with the direction of the local gravity vector. Notwithstanding this effect, the earth's curvature causes nonparallelism of first axes of one arcsecond for each 30 m.
- The targets aimed at generally do not reside on the ellipsoid.

The noncoincidence of the gravity vector and the normal is called "deflection of the vertical." Proper knowledge of the behavior of the local geopotential surfaces is needed for proper distance and angle reductions. Consult Vanicek and Krakiwsky [1982], for example, for the mathematical background of these reductions. The FGCC adopted the accuracy standards given in Table 149.5 for horizontal control using classical geodetic measurement techniques [FGCC, 1984]. In the U.S. over 270 000 horizontal control stations exist.

**Table 149.5** FGCC Horizontal Control Accuracy Standards (Classical Techniques)

First Order	
1:100 000 (10 mm/km)	
Second Order	
Class I	1:50 000 (20 mm/km)
Class II	1:20 000 (50 mm/km)
Third Order	
Class I	1:10 000 (100 mm/km)
Class II	1:5 000 (200 mm/km)

## Three-Dimensional Positioning: Geocentric Positions and Full 3-D Control

Modern 3-D survey techniques, most noticeably GPS, allow for immediate 3-D relative positioning. 3-D coordinates are equally accurately expressed in ellipsoidal, spherical, or Cartesian coordinates. Care should be exercised to properly label curvilinear coordinates as spherical (geographic) or ellipsoidal (geodetic). Table 149.6 shows the large discrepancies between the two.

At the mid-latitudes they may differ by more than 11'. This could result in a north-south error of 20 km. When merging GIS data sets one should be aware of the meaning "LAT/LON" in any instance. Consult [Stem, 1991] for the use of U.S. state plane coordinates. Curvilinear coordinates, their transformations, and their use are discussed in a variety of textbooks; see the section "Further Information" or articles such as [Soler, 1976], [Leick and van Gelder, 1975], [Soler and Hothem, 1988], and [Soler and van Gelder, 1987].

**Table 149.6** Geographic (Spherical) Latitude as a Function of Geodetic Latitude

Geodetic Latitude			Geographic Latitude			Geodetic Minus Geographic Latitude		
Degrees	Minutes	Seconds	Degrees	Minutes	Seconds	Degrees	Minutes	Seconds
00	0	0.000	00	00	00.000	00	00	00.000
10	0	0.000	09	56	03.819	00	03	56.181
20	0	0.000	19	52	35.868	00	07	24.132
30	0	0.000	29	50	01.089	00	09	58.911
40	0	0.000	39	48	38.198	00	11	21.802
50	0	0.000	49	48	37.402	00	11	22.598
60	0	0.000	59	49	59.074	00	10	00.926
70	0	0.000	69	52	33.576	00	07	26.424
80	0	0.000	79	56	02.324	00	03	57.676
90	0	0.000	90	00	00.000	00	00	00.000

Despite their 3-D characteristics, networks generated by GPS are the weakest in the vertical component, not only because of the lack of physical significance of GPS's determined heights, as described in the preceding subsection, but also because of the geometrical distribution of satellites with respect to the vertical: no satellite signals are received from "below the network." This lopsidedness makes the vertical the worst determined component in 3-D.

Because of the inclination of the GPS satellites there are places on earth, most notoriously the mid-latitudes, where there is not an even distribution of satellites in the azimuth sense. For instance, in the northern mid-latitudes we never have as many satellites to the north as we have to the south [see, for example, (Santerre, 1991)]. This makes the latitude the second best determined curvilinear coordinate. For space techniques the FGCC has proposed the classification in Table 149.7.

**Table 149.7** FGCC 3-D Accuracy Standards (Space System Techniques)

AA Order (Global)
3 mm + 1: 100 000 00 (1 mm/100 km)
A Order (Primary)
5 mm + 1: 10 000 000 (1 mm/10 km)
B Order (Secondary)
8 mm + 1: 1 000 00 (1 mm/km)
C Order (Dependent)
10 mm + 1: 100 000 (10 mm/km)

## References

- Chan, L. and Schorr, J. 1994. GPS world receiver survey. *GPS World*. 5(1):38-56.
- Cohen, E. R. and Taylor, B. N. 1988. The fundamental physical constants. *Phys. Today*. 41(8):9-13.
- Defense Mapping Agency (DMA). 1988. *Department of Defense World Geodetic System: Its Definition and Relationships with Local Geodetic Systems*. DMA Technical Report 8350.2 (revised 1 March 1988).
- Escobal, P. R. 1976. *Methods of Orbit Determination*. John Wiley & Sons, New York.
- FGCC. 1984. *Standards and Specifications for Geodetic Control Networks*. (Reprint version February 1991.) Federal Geodetic Control Committee. Rockville, MD.
- FGCC. 1989. *Geometric Geodetic Accuracy Standards and Specifications for using GPS Relative Positioning Techniques*. Version 5.0. Federal Geodetic Control Committee. Rockville, MD.
- GPS World. 1994. Satellite almanac overview. *GPS World*. 5(10):60.
- International Association of Geodesy (IAG). 1971. *Geodetic Reference System 1967*. Publication Spéciale No. 3. IAG, Paris.
- International Association of Geodesy (IAG). 1980. Geodetic reference system 1980 (compiled by H. Moritz). *Bull. Géodésique*. 54(3):395-405.
- International Association of Geodesy (IAG). 1984. Geodetic reference system 1980 (compiled by H. Moritz). *Bull. Géodésique*. 58(3):388-398.
- International Association of Geodesy (IAG). 1988a. Geodetic reference system 1980 (compiled by H. Moritz). *Bull. Géodésique*. 62(3):348-358.
- International Association of Geodesy (IAG). 1988b. Parameters of common relevance of astronomy, geodesy, and geodynamics (compiled by B.H. Chovitz). *Bull. Géodésique*. 62(3):359-367.
- International Earth Rotation Service (IERS). 1992. *IERS Standards (1992)*, ed. D. D. McCarthy. IERS Technical Note 12. Central Bureau of the IERS, Observatoire de Paris.
- Leick, A. and van Gelder, B. H. W. 1975. *On Similarity Transformations and Geodetic Network Distortions Based on Doppler Satellite Coordinates*. Reports of the Department of Geodetic Science, No. 235. Ohio State University, Columbus.
- Montgomery, H. 1993. City streets, airports, and a station roundup. *GPS World*. 4(2):16-19.
- NATO. 1988. *Standardization Agreement on NAVSTAR Global Positioning System (GPS), System Characteristics* Preliminary Draft. STANAG 4294 (revision: 15 April 1988).
- Reilly, J. P. 1994. P.O.B. 1994 GPS equipment survey. *P.O.B.* 19(5):75-86.
- Santerre, R. 1991. Impact of GPS satellite sky distribution. *Manuscripta Geodetica*. 61(1):28-53.
- Soler, T. 1976. *On Differential Transformations between Cartesian and Curvilinear (Geodetic) Coordinates*. Reports of the Department of Geodetic Science, No. 236. Ohio State University, Columbus.

- Soler, T. and Hothem, L. D. 1988. Coordinate systems used in geodesy: Basic definitions and concepts. *J. Surv. Engi.* 114(2):84-97.
- Soler, T. and van Gelder, B. H. W. 1987. On differential scale changes and the satellite Doppler z-shift. *Geophys. J. R. Astron. Soc.* 91:639-656.
- Stem, J. E. 1991. *State Plane Coordinate System of 1983*. NOAA Manual NOS NGS 5. Rockville, MD.
- Wells, D. and Kleusberg, A. 1990. GPS: A multipurpose system. *GPS World*. 1(1):60-63.
- Wells, D. (Ed.). 1986. *Guide to GPS Positioning*. Canadian GPS Associates, Fredericton, New Brunswick.

## Further Information

### Textbooks and Reference Books

For additional reading and more background, from the very basic to the advanced level, in geodesy, satellite geodesy, physical geodesy, mechanics, orbital mechanics, and relativity, the reader is referred to the following textbooks (in English):

- Bomford, G. 1980. *Geodesy*. Clarendon Press, Oxford.
- Goldstein, H. 1965. *Classical Mechanics*. Addison-Wesley Publishing Co., Reading, MA.
- Heiskanen, W. A. and Moritz, H. 1967. *Physical Geodesy*. W. H. Freeman & Co., San Francisco.
- Hofmann-Wellenhof, B., Lichtenegger, H., and Collins, J. 1995. *GPS: Theory and Practice*. Springer-Verlag, New York.
- Jeffreys, H. 1970. *The Earth: Its Origin, History and Physical Constitution*. Cambridge University Press, Cambridge.
- Kaula, W. M. 1966. *Theory of Satellite Geodesy: Applications of Satellites to Geodesy*. Blaisdell Publishing Co., Waltham, MA.
- Lambeck, K. 1988. *Geophysical Geodesy: The Slow Deformations of the Earth*. Clarendon Press, Oxford.
- Leick, A. 1995. *GPS: Satellite Surveying*. John Wiley & Sons, New York.
- Maling, D. H. 1993. *Coordinate Systems and Map Projections*. Pergamon Press, New York.
- Melchior, P. 1978. *The Tides of the Planet Earth*. Pergamon Press, New York.
- Moritz, H. 1990. *The Figure of the Earth: Theoretical Geodesy and the Earth's Interior*. Wichmann, Karlsruhe.
- Moritz, H. and Mueller, I. I. 1988. *Earth Rotation: Theory and Observation*. Frederick Ungar Publishing Co., New York.
- Mueller, I. I. 1969. *Spherical and Practical Astronomy, as Applied to Geodesy*. Frederick Ungar Publishing Co., New York.
- Munk, W. H. and MacDonald, G. J. F. 1975. *The Rotation of the Earth: A Geophysical Discussion*. Cambridge University Press, Cambridge.
- Seeber, G. 1993. *Satellite Geodesy: Foundations, Methods, and Applications*. Walter de Gruyter, New York.
- Soffel, M. H. 1989. *Relativity in Astrometry, Celestial Mechanics and Geodesy*. Springer-Verlag, New York.



Torge, W. 1991. *Geodesy*. Walter de Gruyter, New York.  
Vanicek, P. and Krakiwsky, E. J. 1982. *Geodesy: The Concepts*. North-Holland Publishing Co., Amsterdam.

### **Journals and Organizations**

The latest results from research in geodesy are published in two international magazines under the auspices of the International Association of Geodesy, both published by Springer-Verlag (Berlin/Heidelberg/New York):

*Bulletin Géodésique*

*Manuscripta Geodetica*

Geodesy- and geophysics-related articles can be found in:

American Geophysical Union (Washington, DC): *EOS* and *Journal of Geophysical Research*

Royal Astronomical Society (London): *Geophysical Journal International*

Kinematic GPS-related articles can be found in:

Institute of Navigation: *Navigation*

American Society of Photogrammetry and Remote Sensing: *Photogrammetric Engineering & Remote Sensing*

Many national mapping organizations publish journals in which recent results in geodesy/surveying/mapping are documented:

American Congress of Surveying and Mapping:

*Surveying and Land Information Systems*

Cartography and Geographic Information Systems

American Society of Civil Engineers: *Journal of Surveying Engineering*

Deutscher Verein für Vermessungswesen: *Zeitschrift für Vermessungswesen*, Konrad Wittwer Verlag, Stuttgart

The Canadian Institute of Geomatics: *Geomatica*

The Royal Society of Chartered Surveyors: *Survey Review*

Institute of Surveyors of Australia: *Australian Surveyor*

Worth special mention are the following trade magazines:

*GPS World*, published by Advanstar Communications, Eugene, OR

*P.O.B. (Point of Beginning)*, published by P.O.B. Publishing Co., Canton, MI

*Professional Surveyor*, published by American Surveyors Publishing Co., Arlington, VA

*Geodetical Info Magazine*, published by Geodetical Information & Trading Centre bv., Lemmer, the Netherlands

National mapping organizations such as the U.S. National Geodetic Survey (NGS) regularly make geodetic software available (free and at cost). Information can be obtained from:

National Geodetic Survey

Geodetic Services Branch

National Ocean Service, NOAA  
1315 East-West Highway, Station 8620  
Silver Spring, MD 20910-3282

Thompson J. F. "Surveying Applications for Geographic Information Systems"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Surveying Applications for Geographic Information Systems

---

- 150.1 GIS Fundamentals
- 150.2 Monumentation or Control Surveying
- 150.3 Topographic Surveying
- 150.4 Future GIS Surveying Applications

**James F. Thompson**

*Thompson Professional Group, Inc.*

The utilization of **geographic information systems** (GISs) for the facility management of information and data has become widespread throughout the U.S. and the rest of the world. Inherently, a GIS can be quite complex in terms of the amount of data to be managed; however, the accuracy and usefulness of a given GIS is directly related to the quality and quantity of information and data properly collected, input, and managed within the system itself.

There are several ways that information and data can be obtained for a GIS. These would include such means as the digital input of information from previously prepared maps, plans, and so forth, and many other methods of data collection and input that can involve remote sensing and the routine communication of relational databases. But usually, the foundation and fundamental building blocks of a GIS are based upon the surveying applications utilized for the system.

This chapter discusses the various land surveying applications for GISs and the most commonly used techniques for collecting the field information. Future applications as based upon ever-developing technology are discussed as well.

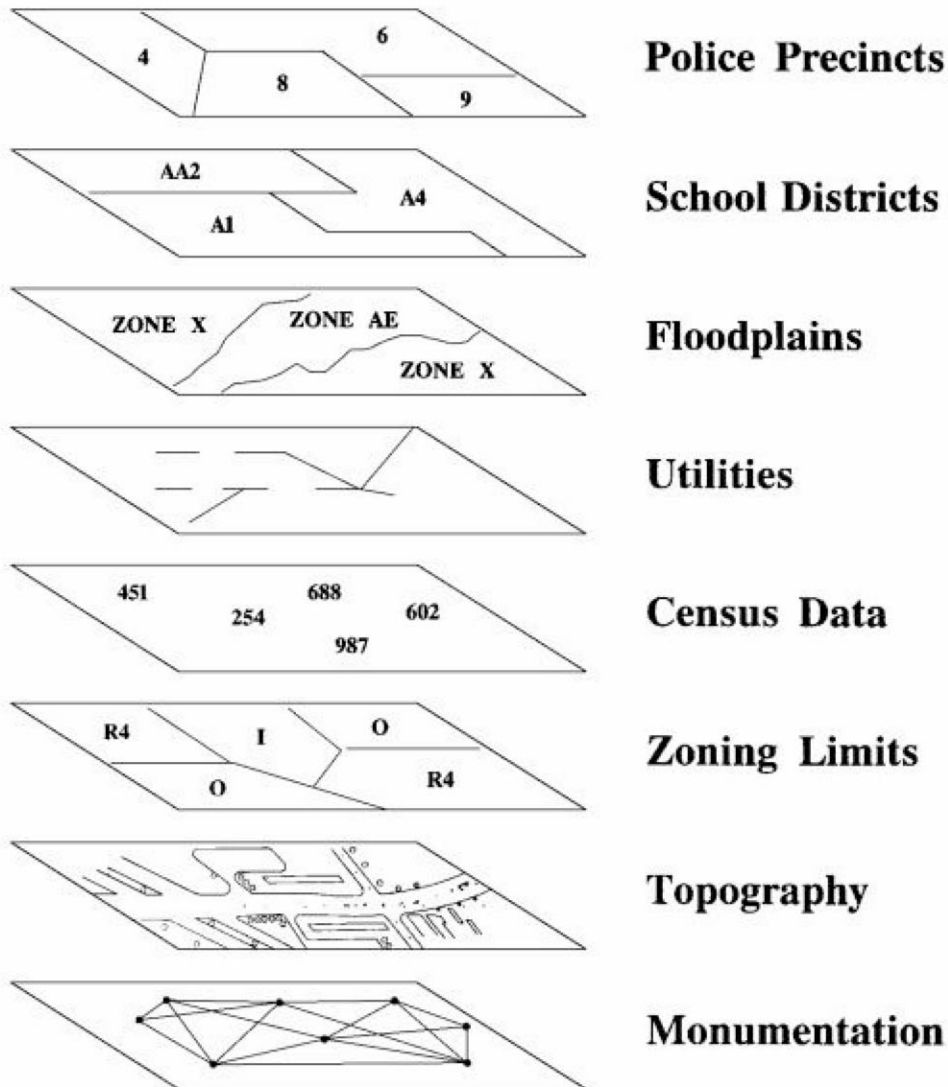
### 150.1 GIS Fundamentals

---

A GIS can be generally described as a computer-based system that stores and manages information and data as related to a geographic area. In other words, data are managed by a graphic computer system to readily provide information on a specific geographic point or area of interest. A GIS is typically divided into graphic layers or levels to better manage the information. [Figure 150.1](#) represents a simple GIS configuration where information of different types is *stacked* to form an organizational structure for the system itself. This graphic information is linked to one or more databases to form a relationship between the graphic information and supporting data. Within a GIS every graphic element, together with its supporting data or attributes, has its own unique geographic coordinates. As such, a GIS can be used to query a database and find information on a

geographic feature, element, or area.

**Figure 150.1** Simplified GIS layers.



The complete discussion of the GIS subject matter can be quite involved and is beyond the scope of this chapter. What is important to note is that the fundamental geographic information itself is based upon the *surveying* data used as the foundation of the GIS. This surveying information can be assembled as based purely on existing plats, maps, and so forth without any detailed field work; however, these "paper surveys" typically will not yield an accurate reflection of existing conditions. It is extremely important that a GIS be founded and maintained on solid and accurate surveying information, because the unique address of a given graphic element will only be as accurate as the overall surveying methods employed for the development of the GIS.

Although there are certainly several distinct categories of surveying, there are two fundamental

land surveying categories that are commonly and easily applied to a GIS. The first is **monumentation** (or control) **surveying**, and the second is **topographic surveying**. Each has a vital function in the development of a GIS and plays a significant role in establishing the inherent accuracy of a GIS. In addition, both categories are directly linked, as further discussed in the following sections.

## 150.2 Monumentation or Control Surveying

---

This singular subject of establishing a monumentation network is one of the most important issues relating to the surveying applications for a GIS. The monumentation network is the first building block of a GIS and sets the accuracy standard for the remaining information to be compiled and managed within the system. Should the level of accuracy of the monumentation network be compromised, the remaining surveying information obtained will be at least as inaccurate, since this other information will be slaved to the control network.

A monumentation system for a GIS is a network of control points or monuments, established in the field, that are referenced not only to each other, but also to a common datum. This common datum, in the U.S., is typically that which has already been established by the National Geodetic Survey (NGS) of the federal government. The NGS has established and maintains a network of first-order monuments with assigned horizontal coordinates across the country. In the past first-order was conventionally thought of as having an accuracy level of 1 part in 100000; however, with the advancements in surveying equipment technology, first-order monuments commonly have a much higher level of accuracy. In addition, vertical elevation information is often available for many NGS monuments. It is important to understand that these NGS monuments provide a common thread that bonds geographic areas across the country. Each state has its own coordinate system, as dependent upon the map projection used, which is generally a result of the shape of the state or area to be projected. **Map projections** form the basis for planar mapping of geographic areas. A surveyor establishing the monumentation network for a given area must be knowledgeable of the projection and resultant **state plane coordinate system** used in the area.

When we establish a geodetic monumentation network for a GIS, we are typically densifying the existing NGS monumentation system in our area of interest. Often, many NGS monuments are recovered in the field and then used as reference points within the densified monumentation network. The result of the monumentation survey is a network of readily identifiable points throughout an area that can be easily used for land surveying and **aerial photogrammetry**. The fundamental issue at hand is that, once the monumentation network is established, a common datum will exist throughout an area that can be used to reference all subsequent surveying data. Even past surveying information can be moved to this newly established datum via a coordinate transformation.

The actual procedures for establishing a geodetic monumentation network can be divided into several stages that build upon each other:

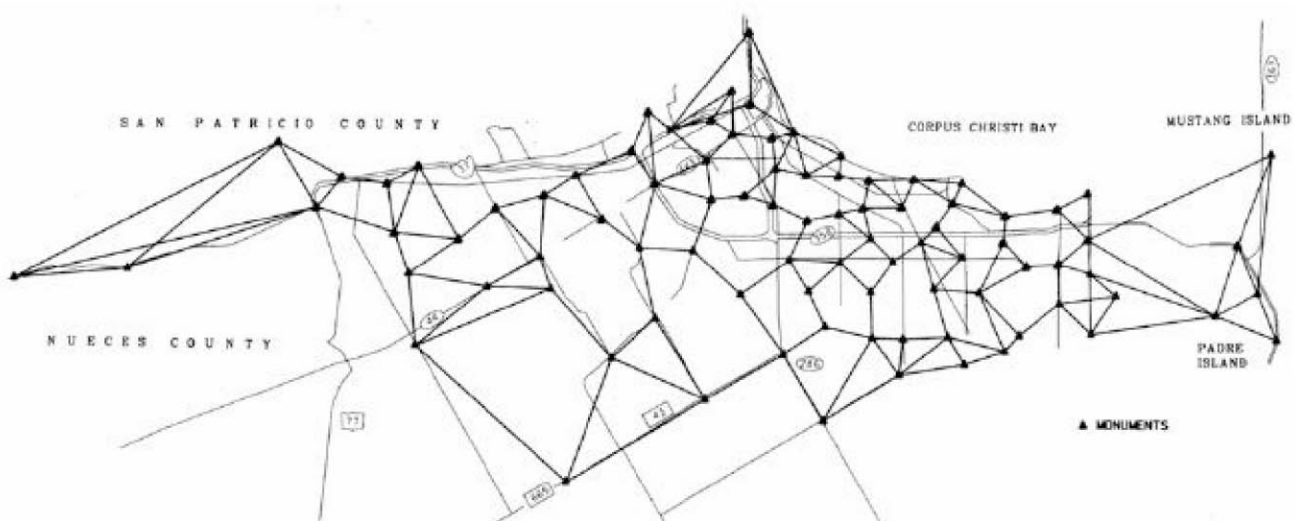
1. Research existing NGS monuments and field recover the monuments.
2. Design a network layout as referenced to the recovered NGS monuments.

3. Plan the field survey activities to establish the monuments.
4. Execute the field survey activities.
5. Mathematically adjust the network and calculate the coordinates of the monuments.

Recovering the existing NGS monuments in the field can be very time consuming and, at times, frustrating. Many monuments have not been recovered or used for several years, thus often making them difficult to find. Once recovered, these NGS monuments and their previously established coordinates will be used to reference and constrain the new monuments to be established.

Once the positions of the recovered NGS monuments are known in relation to the area of interest to serve as the geographic region supported by a GIS, a monumentation network is designed that provides the layout of the new monuments. This layout is composed of the existing NGS monuments, the new monuments to be established, and the base lines connecting all of the monuments. The strength and accuracy of a monumentation network is dependent on the design of the layout and, specifically, the base lines within the layout. Today, the most common means used to perform the survey for the establishment of a geodetic monumentation network is the **global positioning system** (GPS). GPS surveying is a method of using transmitting satellites and portable field receivers to accurately determine the coordinates of a point. (GPS surveying is discussed in detail in **Chapter 149**.) A base line between two monuments is created when separate GPS receivers are set on both points and they receive credible satellite data simultaneously. [Figure 150.2](#) illustrates a designed and successfully implemented base-line and monument layout for a municipal GIS application. In the design of the network layout, consideration must be given to the field procedures required to create the needed base lines throughout the monumentation network.

**Figure 150.2** Base-line and monument layout for a municipal GIS application.



Once the desired monumentation layout is designed, the field activities themselves must be carefully planned. Without proper planning, the GPS field surveys can be disastrous. Assuming that all goes well during the GPS field surveys and that good data were received from the satellites,

the resultant coordinates of the new monuments can be calculated. This entire process is often called *balancing*, *adjusting*, or *constraining* the network. During this process, the entire monumentation network is referenced to the NGS datum, and state plane coordinates are assigned to each monument. State plane coordinates differ from surface coordinates as would be needed for conventional plane table land surveying. Notice that state plane coordinates are used in map projections to account for the curvature of the earth. Every point has a scale factor that must be applied to the state plane coordinates (northing and easting) of that point to convert the point's state plane coordinates to surface coordinates at sea level. To account for the difference of the terrain's elevation that normally differs from sea level, a sea level correction factor must be applied as well. The combination of a scale factor and a sea level correction factor is commonly termed a **combined scale factor**. The following equations are used to convert state plane coordinates to surface coordinates:

$$\text{Combined scale factor} = \text{Scale factor} \times \text{Sea level correction factor}$$

$$\text{Surface coordinate} = \text{State plane coordinate} / \text{Combined scale factor}$$

It is extremely important that the relationship between surface and state plane coordinates is understood. A conventional land surveyor does not measure distances, for example, along the map projection utilized, but, rather, along the true terrain surface being surveyed.

After the horizontal coordinates of the monuments are established, vertical elevations can be determined for the points, such as to support surveys that are in need of relational elevations. Once again, a similar method of balancing or adjusting the network is utilized as based upon the known elevations of points within the monumentation network. Many advancements have been made in accurately calculating vertical elevations using GPS surveying. This is mostly a result of the ability to more accurately account for the true shape of and anomalies in the earth's surface. Although GPS surveying is a good technique to use for establishing the vertical elevations of monuments within a network, conventional field surveying or leveling between the monuments is common and works well.

## 150.3 Topographic Surveying

---

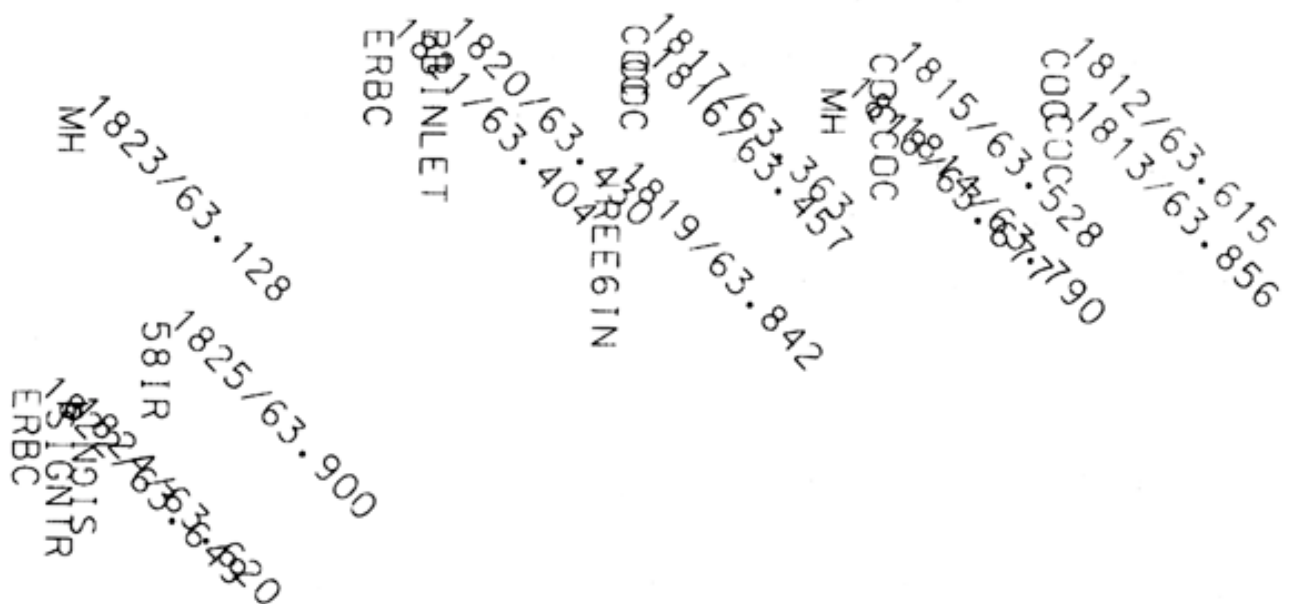
The actual collection of field information that shall be used in a GIS is performed through topographic surveying. The category of topographic surveying as applied to a GIS also generally includes other categories of surveys, such as route, utility, and location surveys. The data collected from such a survey include not only the location of surface features (such as telephone poles, streets, trees, rivers, buildings, houses, and other readily identifiable surface topographic features) but also the rights-of-way, easements, property ownership lines, and underground utilities (such as pipelines and sewers). In a sense, the surveyor is the *eyes* of a GIS. What the surveyor includes in the survey in terms of topography, utilities, property corners, and so forth will dictate how much information is available within the GIS. As such, it is important to include as much information as desired or needed in the topographic surveys, which will serve to feed required data to the GIS. As future needs arise, more topographic information can be obtained and included in a GIS.



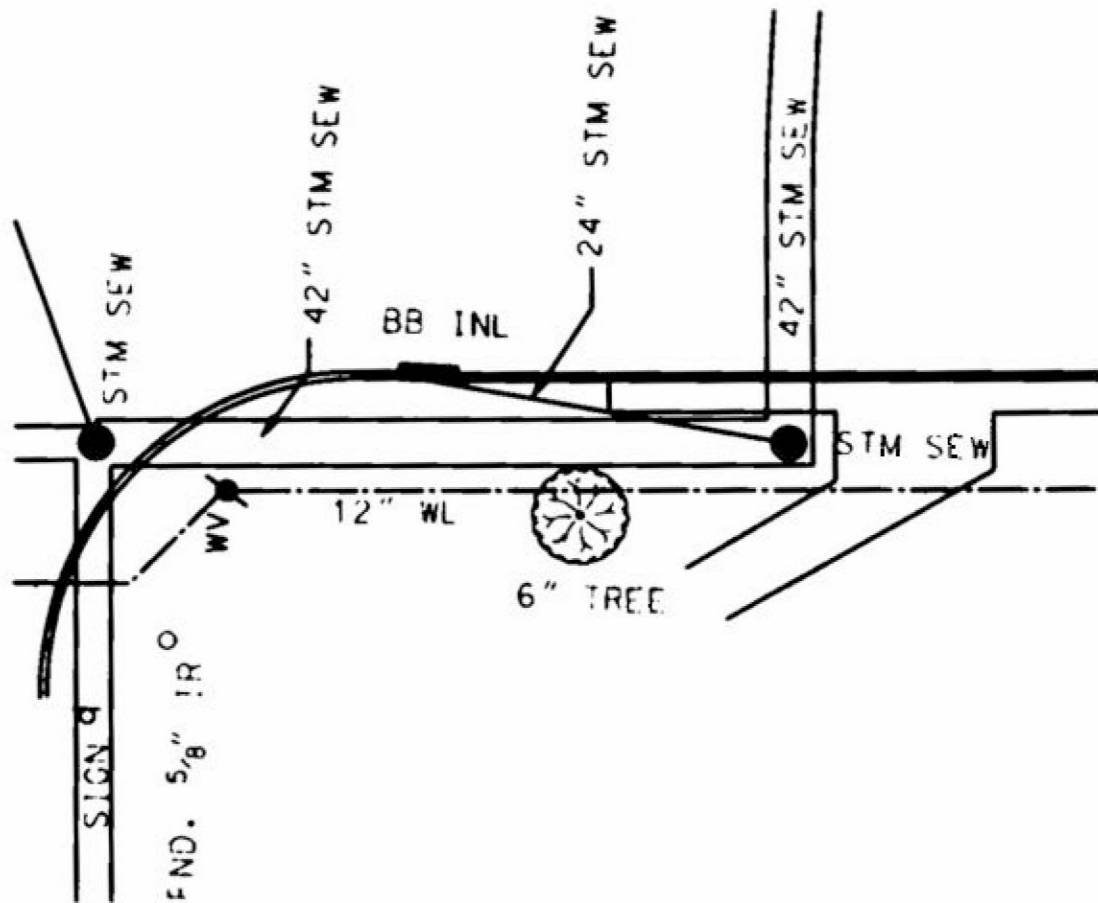
The methods used to perform topography surveying can vary tremendously. The most common method used is conventional land surveying; however, even that method has changed dramatically with advanced technology. Electronic data collectors (in lieu of hand-written field books) are quite common today, and even GPS receivers are being used routinely in topography surveys. In conjunction with conventional land surveys, aerial photogrammetry is used extensively to obtain topographic information needed to supply data to a GIS. Although conventional land surveying and aerial photogrammetry are some of the most popular means of collecting topographic data for GIS applications, other methods, such as **satellite imagery**, are applied as well.

In performing a conventional topographic survey, a surveyor will traverse between the geodetic network monuments, obtaining as much topographic information as desired. Again, the information sought is dependent on the demands of the GIS that dictate the detail needed in terms of the survey data obtained. The traversed line of the survey between the monuments is balanced in terms of angles and distances as based upon the known coordinates of the monuments; this will ensure that coordinates of the obtained topographic features are properly determined in relation to the datum used for the GIS. [Figure 150.3](#) represents a series of points obtained for a GIS using conventional topographic surveying techniques. [Figure 150.4](#) illustrates the same data in their final form, as would be input graphically into the GIS. Every point has its own unique address and attribute. A fire hydrant, for example, will have its own specific coordinates. A relational database can be linked to the fire hydrant graphic element within the GIS—which would provide such information as when the hydrant was installed and when last checked to ensure its proper operating condition.

**Figure 150.3** Series of points obtained using conventional topographic surveying techniques.



**Figure 150.4** Points converted to topographic features to be input graphically into a GIS.



Aerial photogrammetry, an extremely powerful and cost-effective method of obtaining a great amount of topographic information, is used routinely in GIS applications with conventional land surveying. Like the ground surveys performed, the aerial surveys are linked to the GIS monumentation network by referencing the aerial photographs to known points on the ground that can be clearly identified in the photographs. This is typically done by painting (or in many cases, taping) a large white or contrasting colored mark on the ground in the shape of an "X" or a "V" that can be clearly identified in the photograph. These marks are then tied to the monument system conventionally with ground surveys. The end result is an aerial photograph, or series of photographs, rectified to accurately illustrate the existing topography on the proper coordinate system. Once the aerial photographs are produced, they are digitized into graphic elements that are directly incorporated into a GIS. With the horizontal position of the topographic features, the elevation contours and many specific point elevations can be determined as well.

A combination of conventional land surveys and aerial photogrammetry can be used quite effectively in obtaining most of the topographic features needed for a given GIS application; however, other more advanced surveying methods such as radar and satellite observations can be employed as well. Everything from the broad expanse of wooded areas to the finite detail of utility locations can be surveyed, identified, and incorporated into a GIS. A visionary approach should be

taken regarding the surveying applications of a GIS. It is this approach that will drive the accuracy and integrity of the GIS itself.

## 150.4 Future GIS Surveying Applications

---

When dealing with constantly evolving technology, it often seems that the future is already upon us. In the not-too-distant future, it is expected that GIS technology will be so widespread that essentially every area of the country will be included into a GIS in some fashion. Cities will be linked with other cities, states with other states, and countries with other countries. Automobiles will travel without human interface, because every road and highway will have its route defined with pinpoint accuracy. Every home, building, fence corner, and manhole will have its own unique address in terms of coordinates. All of this information, as managed by a GIS, must be originally established and properly maintained by the surveyor. Hopefully, this task will be made easier by surveying methods of the next generation, which might include handheld survey transceivers and other advanced technologies.

### Defining Terms

**Aerial photogrammetry:** The surveying of surface features through the use of photographs, as taken from an aerial perspective (i.e., an airplane).

**Combined scale factor:** A factor that can be used to convert state plane coordinates to surface terrain coordinates, considering the map projection utilized and the relative elevation of the point of interest.

**Geographic information system:** A computer-based system that stores and manages information and data as related to a geographic area.

**Global positioning system:** A method of surveying that uses transmitting satellites and portable field receivers to accurately determine the coordinates of a point.

**Map projection:** The projection of the earth's surface to a plane (or map), considering the curvature of the earth.

**Monumentation surveying:** The establishment or recovery of a horizontal and/or vertical coordinate control network as based upon a layout of monuments or control points.

**Satellite imagery:** Images of the earth's surface as obtained from orbiting satellites.

**State plane coordinate system:** A coordinate system used throughout the U.S. that is based on the map projection of one or more zones within each state.

**Topographic surveying:** The surveying of topographic features through conventional land surveying, aerial photogrammetry, or other means.

### References

- American Congress on Surveying and Mapping. 1993. *Surveying and Land Information Systems*. 53(4). Bethesda, MD.
- Davis, R. E., Foote, F. S., Anderson, J. M., and Mikhail, E. M. 1981. *Surveying Theory and Practice*. McGraw-Hill, New York.

Federal Geodetic Control Committee. 1984. *Standards and Specifications for Geodetic Control Networks*. NOAA, Rockville, MD.

## **Further Information**

One of the largest sources of information pertaining to surveying applications for geographic information systems is the American Congress of Surveying and Mapping (ACSM), 5410 Grosvenor Lane, Bethesda, MD 20814-2122. The ACSM publishes journals relating to surveying, mapping, and land information.

The NGS Information Center maintains a tremendous amount of valuable information on geodetic surveying, horizontal and vertical coordinate listings, and other applicable publications. This information may be obtained from NOAA, National Geodetic Survey, N/CG17, 1315 East-West Highway, Room 9202, Silver Spring, MD 20910-3282.

Kiefer, R. W. "Remote Sensing"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

151.1 Electromagnetic Energy

151.2 Atmospheric Effects

151.3 Remote Sensing Systems

Aerial Photography • Multispectral, Thermal, and Hyperspectral Scanners • Side-Looking Radar

151.4 Remote Sensing from Earth Orbit

151.5 Digital Image Processing

**Ralph W. Kiefer**

*University of Wisconsin, Madison*

**Thomas M. Lillesand**

*University of Wisconsin, Madison*

**Remote sensing** involves the use of airborne and space-imaging systems to inventory and monitor earth resources. Broadly defined, remote sensing is any methodology employed to study the characteristics of objects from a distance. Using various remote sensing devices, we remotely collect *data* that can be analyzed to obtain *information* about the objects, areas, or phenomena of interest. This chapter discusses sensor systems that record energy over a broad range of the **electromagnetic spectrum**, from ultraviolet to microwave wavelengths.

Remote sensing affords a practical means for frequent and accurate monitoring of the earth's resources from a site-specific to global basis. This technology is aiding in assessing the impact of a range of human activities on our planet's air, water, and land. Data obtained from remote sensors have provided information necessary for making sound decisions and formulating policy in a host of resource development and land use applications. Remote sensing techniques have also been used in numerous special applications. Expediting petroleum and mineral exploration, locating forest fires, providing information for hydrologic modeling, aiding in global crop production estimates, monitoring population growth and distribution, and determining the location and extent of oil spills and other water pollutants are but a few of the many and varied applications of remote sensing that benefit humankind on a daily basis. It should be pointed out that these applications almost always involve some use of **ground truth** or on-the-ground observation. That is, remote sensing is typically a means of extrapolating from, not replacing, conventional field observation.

## 151.1 Electromagnetic Energy

The sun and various other sources radiate electromagnetic energy over a range of wavelengths. Light is a particular type of **electromagnetic radiation** that can be seen or sensed by the human eye. All electromagnetic energy, whether visible or invisible, travels in the form of sinusoidal waves. Wavelength ranges of special interest in remote sensing are shown in [Table 151.1](#).

**Table 151.1** Components of the Electromagnetic Spectrum

Wavelength	Spectral Region
0.3 to 0.4 $\mu\text{m}$	Ultraviolet
0.4 to 0.7 $\mu\text{m}$	Visible: 0.4 to 0.5 $\mu\text{m}$ = blue 0.5 to 0.6 $\mu\text{m}$ = green 0.6 to 0.7 $\mu\text{m}$ = red
0.7 to 1.3 $\mu\text{m}$	Near infrared
1.3 to 3.0 $\mu\text{m}$	Mid-infrared
3 to 14 $\mu\text{m}$	Thermal infrared
1 mm to 1 m	Microwave

When electromagnetic energy is incident upon an object on the earth's surface, it can interact with the object in any or all of three distinct ways. The incident energy can be reflected, transmitted, or absorbed. The absorbed component goes into heating the body and is subsequently re-emitted from the object. The particular mix of these three possible interactions is dependent upon the physical nature of objects. For example, healthy vegetation normally appears green because the blue and red components of the incident light are absorbed by chlorophyll present in plant leaves. In contrast, concrete surfaces strongly reflect blue, green, and red wavelengths nearly equally and appear light gray. Remote sensors record such variations in energy interaction (both in visible and invisible wavelengths) in order to discriminate between earth surface features and to assist in quantifying their condition.

All objects at a temperature greater than absolute zero radiate energy according to the formula

$$M = \sigma \varepsilon T^4 \quad (151.1)$$

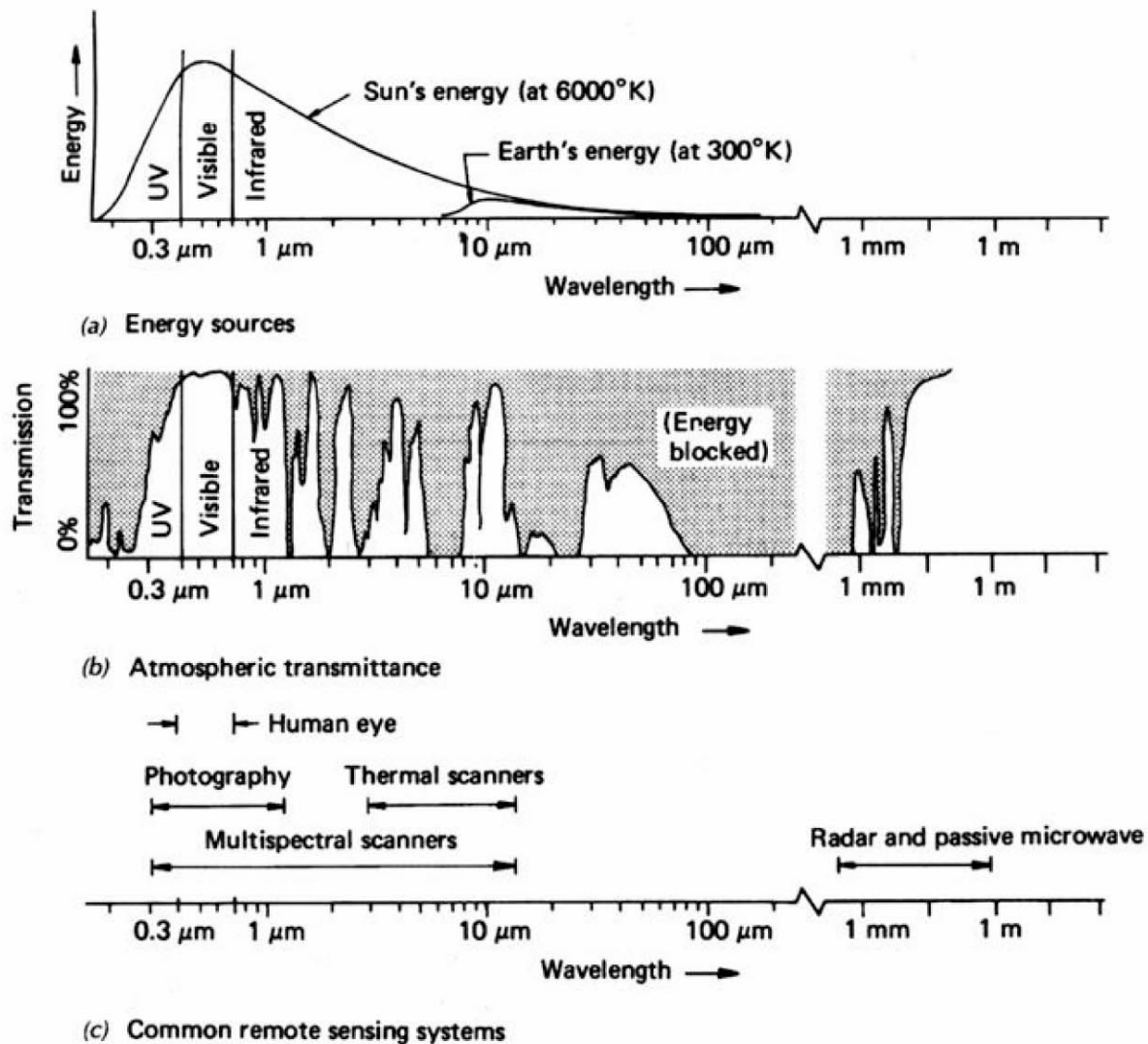
where  $M$  is the total radiant exitance (radiated energy) from the surface of a material ( $\text{W m}^{-2}$ ),  $\sigma$  is the *Stefan-Boltzmann constant* ( $5.6697 \cdot 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$ ),  $\varepsilon$  is the emissivity (efficiency of radiation) of the material, and  $T$  is the absolute temperature (K) of the emitting material. The general shape of the resulting curves of **emitted energy** (radiant exitance) versus wavelength is shown in [Fig. 151.1\(a\)](#). The wavelength at which the amount of emitted energy is a maximum is related to its temperature by *Wien's displacement law*, which states that

$$\lambda_m = \frac{A}{T} \quad (151.2)$$

where  $\lambda_m$  is the wavelength of maximum spectral radiant exitance ( $\mu\text{m}$ ),  $A = 2898 \mu\text{m K}$ , and  $T$  = temperature (K). Note that the sun's energy has a peak wavelength of  $0.48 \mu\text{m}$  (in the visible part

of the spectrum), whereas the earth's energy has a peak wavelength of  $9.67 \mu\text{m}$  (invisible to the human eye, but able to be sensed by a thermal scanner).

**Figure 151.1** Spectral characteristics of (a) energy sources, (b) atmospheric effects, and (c) sensing systems. Note that wavelength scale is logarithmic. (Source: Lillesand, T. M. and Kiefer, R. W. 1994. *Remote Sensing and Image Interpretation*, 3rd ed. John Wiley & Sons, New York. With permission.)



**Reflected energy** is recorded when objects are sensed in sunlight in the ultraviolet, visible, near-infrared or mid-infrared portions of the spectrum. Radiated energy is recorded in the thermal infrared portion of the electromagnetic spectrum using radiometers and thermal scanners. This allows, for example, the detection and recording of the heated effluent from a power plant as it flows into a lake at a cooler temperature.

Remote sensing systems can be active or passive systems. The examples cited earlier are passive



systems in that they record reflected sunlight or emitted radiation. Radar systems are called *active systems* because they supply their own energy. Pulses of microwave energy are transmitted from radar systems toward objects on the ground, and the backscattered energy is then received by radar antennas and used to form images of the strength of the radar return from various objects.

## 151.2 Atmospheric Effects

---

Because the atmosphere contains a wide variety of suspended particles, it offers energy interaction capabilities just as "ground" objects do. The extent to which the atmosphere transmits electro-magnetic energy is dependent upon wavelength, as shown in [Fig. 151.1\(b\)](#). The sensing systems typically used in various wavelength ranges are shown in [Fig. 151.1\(c\)](#). Energy in the ultraviolet wavelengths is scattered greatly, which limits the use of ultraviolet wavelengths from aerial or space platforms. The atmosphere is transparent enough in the visible, near-infrared, and mid-infrared wavelengths to permit aerial photography and multispectral sensing in these wavelengths. In this region the blue wavelengths are scattered the most and the mid-infrared wavelengths are scattered the least. In the thermal infrared region there are two "windows" where the atmosphere is relatively transparent: 3–5  $\mu\text{m}$  and 8–14  $\mu\text{m}$  wavelength (most aerial thermal scanning is done in the 8–14  $\mu\text{m}$  band). At microwave wavelengths the atmosphere is extremely transparent, and many radar systems can be operated in virtually all weather conditions.

## 151.3 Remote Sensing Systems

---

### Aerial Photography

Aerial photographs can be taken on any of several film types, from a variety of flying heights. Mapping cameras typically use an image size of 230 by 230 mm. Smaller-format cameras (70 mm and 35 mm) can be used where large-area coverage with great geometric fidelity is not required. The interpretability of aerial photographs is highly dependent on the selection of film type and image scale.

Principally because of cost considerations, the films most widely used for aerial photography are the black and white (b/w) films. Panchromatic films are b/w films sensitive to the visible portion of the electromagnetic spectrum (0.4 to 0.7  $\mu\text{m}$ ). The sensitivity of black and white infrared films includes both the visible part of the spectrum and also the wavelengths 0.7 to 0.9  $\mu\text{m}$  (near infrared). It is important to note that infrared energy of these wavelengths does not represent heat emitted from objects, but simply reflected infrared energy to which the human eye is insensitive.

Color and color infrared films are also widely used for aerial photography. Although the cost of using these films is greater than for black and white films, they provide greater information content due to the human eye's ability to discriminate substantially more colors than shades of gray. Normal color films have three separate emulsion layers sensitive to blue, green, and red wavelengths, respectively. Color infrared films have three similar emulsion layers, but they are sensitive to green, red, and near-infrared wavelengths, respectively. Again, as in the case of b/w infrared films, it is reflected sunlight, not emitted energy, that is photographed with color infrared

film.

*Digital cameras* use a camera body and lens but record image data with charge-coupled devices (CCDs) rather than film. The electrical signals generated by these detectors are stored digitally, typically using media such as computer disks. Although this process is not "photography" in the traditional sense (images are not recorded directly onto photographic film), it is often referred to as "digital photography." Of course, hard copy photographs can also be converted into an array of digital picture elements (**pixels**) using some form of image scanner.

*Video cameras* are sometimes used as a substitute for small-format (70 mm and 35 mm) cameras. Video camera data are recorded on videotape, typically in HI-8 format.

The scale of photographs affects the level of useful information they contain. Small-scale photographs (1:50 000 or smaller) are used for reconnaissance mapping, large-area resource assessment, and large-area planning. Medium-scale photographs (1:12 000 to 1:50 000) are used for the identification, classification, and mapping of such items as tree species, agricultural crop types, vegetation communities, and soil types. Large-scale photographs (larger than 1:12 000) are used for the intensive monitoring of specific items such as surveys of the damage caused by plant diseases, insects, or tree blow-downs. Applications such as hazardous waste site assessment often require very-large-scale photographs.

The principles of **photogrammetry** can be used to obtain approximate distances and ground elevations from aerial photographs using relatively unsophisticated equipment and simple geometric concepts, as well as to obtain extremely precise maps and measurements using sophisticated "soft copy" instrumentation, digital images, and complex computational techniques.

## Multispectral, Thermal, and Hyperspectral Scanners

*Multispectral scanners* are electro-optical devices that sense selectively in multiple spectral bands using electronic detectors rather than film. They sense one small area on the ground at a time and, through scanning, build up two-dimensional images of the terrain for a swath beneath an aircraft or spacecraft. Through this process, they collect rows and columns of image data that can be computer processed. As shown in [Fig. 151.1](#), multispectral scanners can sense in a much broader spectral range than film. Multispectral scanners are the sensing devices used in the Landsat and SPOT satellites (discussed later).

*Thermal scanners* are electro-optical devices that sense in the thermal infrared portion of the electromagnetic spectrum. They do not record the true internal temperature of objects (kinetic temperature), but rather their apparent temperature based on the radiation from their top surfaces (radiant temperature). Because they sense energy emitted (rather than reflected) from objects, thermal scanning systems can operate day or night. Multiple thermal bands can be sensed simultaneously, as in the case of NASA's *Thermal Infrared Multispectral Scanner*. Successful interpretations of thermal imagery have been made in such diverse tasks as determining rock type and structure, locating geological faults, mapping soil type and moisture conditions, determining the thermal characteristics of volcanoes, studying evapotranspiration from vegetation, locating cold water springs, determining the extent and characteristics of thermal plumes in lakes and rivers, delineating the extent of active forest fires, and locating underground coal mine fires.

*Hyperspectral scanners* acquire multispectral images in many very narrow, contiguous spectral

bands throughout the visible, near-infrared, and mid-infrared portions of the electromagnetic spectrum. These systems typically collect 200 or more channels of data, which enables the construction of an effectively continuous reflectance spectrum for every pixel in the scene (as opposed to the 4–6 broad spectral bands used by the Landsat and SPOT satellites).

## Side-Looking Radar

An increasing amount of valuable environmental and resource information is being acquired by active radar systems that operate in the microwave portion of the spectrum. Microwaves are capable of penetrating the atmosphere under virtually all conditions. Depending on the specific wavelengths involved, microwave energy can penetrate clouds, fog, light rain, and smoke. Side-looking radar uses an antenna pointed to the side of the aircraft or spacecraft. Because the sensor is mounted on a moving platform, it is able to produce continuous strips of imagery depicting very large ground areas located adjacent to the flight line.

Microwave reflections from earth materials bear no direct relationship to their counterparts in the visible portion of the spectrum. For example, many surfaces that appear rough in the visible portion of the spectrum may appear smooth as seen by microwaves (e.g., a white sand beach). The appearance of various objects on radar images depends principally on the orientation of the terrain relative to the aircraft or spacecraft (important because this is a side-looking sensor), the object's surface roughness, its moisture content, and its metallic content.

Radar image interpretation has been successful in applications as varied as mapping major rock units and surficial materials, mapping geologic structure (folds, faults, and joints), discriminating vegetation types (natural vegetation and crops), determining sea ice type and condition, and mapping surface drainage patterns (streams and lakes).

## 151.4 Remote Sensing from Earth Orbit

---

The use of satellites as sensor platforms has made possible the acquisition of repetitive multispectral data of the earth's surface on a global basis. The principal earth resources satellites to date have been the U.S. Landsat and the French SPOT systems. The application of satellite image interpretation has already been demonstrated in many fields, such as agriculture, botany, cartography, civil engineering, environmental modeling and monitoring, forestry, geography, geology, geophysics, habitat assessment, land resource analysis, land use planning, oceanography, range management, and water resources.

The Landsat satellites image each spot on the earth's surface once each 16 days, providing for frequent, synoptic, repetitive, global coverage. The Landsat multispectral scanner currently in operation is the *Thematic Mapper*, which senses in six bands of the spectrum, from the blue through the mid-infrared, with a ground resolution cell size of 30 by 30 m. A seventh band senses in the thermal infrared with a ground resolution cell size of 120 by 120 m.

The SPOT satellites image each spot on the earth's surface once each 26 days, but the satellite can be aimed (using ground commands) as much as 27° from nadir to allow more frequent viewing opportunities (e.g., at a latitude of 45° a total of 11 viewing opportunities exist during each 26-day cycle). Stereoscopic imaging is also possible due to the off-nadir viewing capabilities. The

multispectral scanner of the current SPOT satellite senses in three bands of the spectrum (green, red, and near infrared) with a ground resolution cell size of 20 by 20 m. In its "panchromatic" (black and white) mode it senses in one broad band with a ground resolution cell size of 10 by 10 m.

The Canadian Radarsat satellite is planned for launch in 1995. This satellite will provide radar data with many possible swath widths and resolutions on a 1–3 day repeat coverage (depending on latitude). The primary applications for which Radarsat has been designed include ice reconnaissance, coastal surveillance, land cover mapping, and agricultural and forest monitoring.

Other earth resources satellites include those of the European Space Agency and the National Space Agency of Japan. Likewise, several other countries operate, or are planning to launch, earth resources satellites.

## 151.5 Digital Image Processing

---

The digital data acquired by multispectral and thermal scanners, radar systems, and digital cameras are typically computer processed to produce images through *digital image processing*. Through various image-processing techniques, digital images can be enhanced for viewing and human image interpretation. Digital data can also be processed using computer-based image classification techniques to prepare various thematic maps, such as land cover maps. Digital image-processing procedures are normally integrated with the functions of geographic information systems (GIS).

### Defining Terms

**Electromagnetic radiation:** The transmission of energy in the form of waves having both an electric and a magnetic component.

**Electromagnetic spectrum:** Electromagnetic radiation is most simply characterized by its frequency or wavelength. When electromagnetic waves are so ordered, the resulting array is called the *electromagnetic spectrum*. The spectrum is normally considered to be bounded by cosmic rays at the short wavelength end and by microwaves at the long wavelength end.

**Emitted energy:** The energy radiated by an object resulting from its internal molecular motion (heat). All objects above "absolute zero" in temperature radiate energy.

**Ground truth (or reference data):** Field observations or other information used to aid or verify the interpretation of remotely sensed data.

**Photogrammetry:** The science, art, and technology of obtaining reliable measurements, maps, digital elevation models, thematic GIS data, and other derived products from photographs.

**Pixel:** The cell representing each combination of row and column (picture element) in a digital data set.

**Reflected energy:** That component of incident energy that is reflected from an object.

**Remote sensing:** Studying the characteristics of objects from a distance by recording and analyzing electromagnetic energy, typically from ultraviolet to microwave wavelengths.

## References

- American Society of Photogrammetry. 1980. *Manual of Photogrammetry*, 4th ed. ASP, Falls Church, VA.
- American Society of Photogrammetry. 1983. *Manual of Remote Sensing*, 2nd ed. ASP, Falls Church, VA.
- American Society for Photogrammetry and Remote Sensing. 1995. *Manual of Photographic Interpretation*, 2nd ed. ASPRS, Bethesda, MD.
- American Society for Photogrammetry and Remote Sensing. 1995. *Manual of Remote Sensing<sup>3/4</sup>Earth Observing Platforms and Sensors*, 3rd ed. CD-ROM. ASPRS, Bethesda, MD.
- Avery, T. E. and Berlin, G. L. 1992. *Fundamentals of Remote Sensing and Airphoto Interpretation*, 5th ed. Macmillan, New York.
- Campbell, J. B. 1987. *Introduction to Remote Sensing*. Guilford, New York.
- Elachi, C. 1987. *Introduction to the Physics and Techniques of Remote Sensing*. John Wiley & Sons, New York.
- Jensen, J. R. 1986. *Introductory Digital Image Processing: A Remote Sensing Perspective*. Prentice Hall, Englewood Cliffs, NJ.
- Lillesand, T. M. and Kiefer, R. W. 1994. *Remote Sensing and Image Interpretation*, 3rd ed. John Wiley & Sons, New York.
- Sabins, F. F., Jr. 1987. *Remote Sensing: Principles and Interpretation*, 2nd ed. Freeman, New York.
- Wolf, P. R. 1983. *Elements of Photogrammetry*, 2nd ed. McGraw-Hill, New York.

## Further Information

The leading professional society dealing with remote sensing and photogrammetry is the American Society for Photogrammetry and Remote Sensing, 5410 Grosvenor Lane, Suite 210, Bethesda, MD 20814. The society publishes the monthly journal *Photogrammetric Engineering & Remote Sensing*, as well as many books and special publications.

Information on the availability of cartographic and image data throughout the U.S., including aerial photographs and satellite images, can be obtained from the Earth Science Information Center (ESIC), U.S. Geological Survey, 507 National Center, Reston, VA 22092.

Information on the availability of Landsat data on a worldwide basis can be obtained from EOSAT Corporation, 4300 Forbes Boulevard, Lanham, MD 20706.

Information on the availability of SPOT data on a worldwide basis can be obtained from SPOT Image, 16 bis, avenue Edouard-Belin, 31030 Toulouse Cedex, France. For North America contact SPOT Image Corporation, 1897 Preston White Drive, Reston, VA 22091.

Information on the availability of Radarsat data can be obtained from Radarsat International Inc., Building D, Suite 200, 3851 Shell Road, Richmond, British Columbia V6X 2W2, Canada.

Raven, F. H. "Control Systems"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



The Hubble Space Telescope was launched on April 24, 1990, with an optical flaw. In December 1993 after five days of extravehicular activity the telescope was successfully repaired. Additional servicing activities included the installation of new solar arrays that eliminate the jitter experienced as the telescope pans from day to night. This allows the telescope to point within 0.007 arcsec accuracy.

This photograph was taken on December 9, 1993, as the Hubble Space Telescope began its separation from the Space Shuttle *Endeavour* after it spent a week and a half berthed in the space vehicle's cargo bay. Part of the earth's horizon is visible in the upper right corner. (Photo courtesy of NASA.)

# XXIV

## Control Systems

---

**Francis H. Raven**

*University of Notre Dame*

**152 Feedback** *W. L. Brogan*

Characteristics of a Feedback Control System • Fundamentals of Feedback for Time-invariant Linear Systems • Illustrative Applications • Feedback in Nonlinear Systems • Summary of the Typical Steps in the Design of Feedback Systems

**153 Root Locus** *W. L. Brogan*

The Root Locus Concept • Root Locus Details • Generation of Root Locus Plots • Examples • Summary and Conclusions

**154 Nyquist Criterion and Stability** *N. S. Nise*

Concept and Definition of Frequency Response • Plotting Frequency Response • Stability • Nyquist Criterion for Stability • Gain Design for Stability via the Nyquist Criterion • Stability via Separate Magnitude and Phase Plots (Bode Plots)

**155 System Compensation** *F. H. Raven*

Correlation between Transient and Frequency Response • Determining  $K$  to Yield a Desired  $M_p$  • Gain Margin and Phase Margin • Series Compensation • Internal Feedback • Compensation on the  $S$  Plane

**156 Process Control** *T. E. Marlin*

Control Performance and Design Decisions

**157 Digital Control** *R. G. Jacquot*

Feedback Control • Digital Control • Microcontroller Architecture • Linear Digital Control • Digital Control Stability Analysis and Design • Computer-Aided Design

**158 Robots and Control** *R. G. Bonitz and T. C. Hsia*

Independent Joint Control • Method of Computed Torque • Cartesian-Space Control

**159 State Variable Feedback** *T. L. Vincent*

Linear State Space Control Systems • Controllability and Observability • Eigenvalue Placement • Observer Design

THIS SECTION PRESENTS THE BASIC CONCEPTS of control system theory and



examines the inherent structure of feedback control systems. Emphasis is given to basic techniques and methods for determining the transient behavior of systems. The chapters show how to design and modify systems to obtain the desired system performance.

The basic features of feedback control systems are explained and illustrated in **Chapter 152**. These features include improved stability and transient response, reduced sensitivity to parameter variations, better steady state accuracy, minimization of the effect of external disturbances, and the ability to automatically follow an input signal.

The transient response is governed by the location of the roots of the characteristic equation. **Chapter 153** shows how to plot the path of all the roots of the characteristic equation as the system gain  $K$  is varied. Thus, the value of  $K$  which yields the desired transient response is determined directly from the root-locus plot for the system.

Frequency response methods provide a different vantage point than transient response methods for viewing feedback control systems. **Chapter 154** introduces frequency response concepts and the determination of stability using the Nyquist stability criterion. The transfer function for a system may be determined experimentally by frequency response methods.

The correlation criteria which relate transient response to frequency response are explained in **Chapter 155**. Various system compensation techniques for changing the frequency response to achieve the desired transient behavior for the system are described. Finally, it is shown how the root-locus plot may be modified by various compensators to obtain good transient response.

**Chapter 156** addresses the application of control theory to the process industries. These industries typically involve continuous processes in industries such as chemical, petroleum, pulp and paper, steel, and electrical power generation. These processes tend to be highly nonlinear and difficult to model accurately. The complex and time-varying dynamics of such systems require all control designs to explicitly consider robustness to ensure stability and performance over the expected range of operating conditions.

With the advent of relatively inexpensive microcomputer systems, low-cost computer systems are now being incorporated into automotive and home appliance systems, for example. **Chapter 157** describes the basic procedures for designing such systems. Digital control systems are inherently more accurate and provide greater flexibility for controlling the transient response than analog systems.

The fundamental control problem in robotics is to determine the actuator signals required to ensure that the desired motion is executed in accordance with specified

performance criteria. Because a robot's dynamics are described by a set of coupled nonlinear differential equations, the solution may be quite complicated. As described in **Chapter 158**, the planning of the manipulator trajectory to achieve a desired motion is integrally linked to the control problem.

State variable feedback is used for the design of both nonlinear and linear control systems. However, to use state variable feedback, every state variable must be either measured or estimated. Since measuring every state variable is impractical for most control applications, a state estimator must be included. The design of state variable feedback systems is the topic of **Chapter 159**.

Brogan, W. L. "Feedback"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 152

## Feedback

---

- 152.1 Characteristics of a Feedback Control System
- 152.2 Fundamentals of Feedback for Time-invariant Linear Systems
- 152.3 Illustrative Applications
- 152.4 Feedback in Nonlinear Systems
- 152.5 Summary of the Typical Steps in the Design of Feedback Systems

**William L. Brogan**

*University of Nevada, Las Vegas*

"Feedback, one of the most fundamental processes existing in nature, is present in almost all dynamic systems, including those within man, among men, and between men and machines" [DiStefano *et al.*, 1967]. The reflex action of the neural-muscular system when the hand touches a hot stove is one example. The corrections applied to an automobile steering wheel when the driver sees a deviation from the desired path is another. Performance reviews in the workplace or the classroom may lead to altered work assignments and either rewards or admonishments. The home furnace comes on whenever the thermostat-measured temperature falls sufficiently far below the desired temperature.

Each of these examples has a primary system or dynamic unit that produces an output, and a sensor or measurement unit that produces a feedback signal based upon that output. The input to the system depends on a comparison of the feedback signal with the goals for the system. The difference between a commanded input and the feedback signal is frequently used to drive the system. This error-correcting concept is the essence of feedback control. The history of feedback is summarized by Franklin *et al.* [1986], Dorf [1989], and D'Azzo and Houpis [1988].

### 152.1 Characteristics of a Feedback Control System

---

When applied properly, the effects of feedback are as follows:

1. Improved stability and transient response
2. Ability to automatically track a specified input
3. Higher steady state accuracy
4. Reduced sensitivity to parameter variations
5. Better disturbance rejection capability

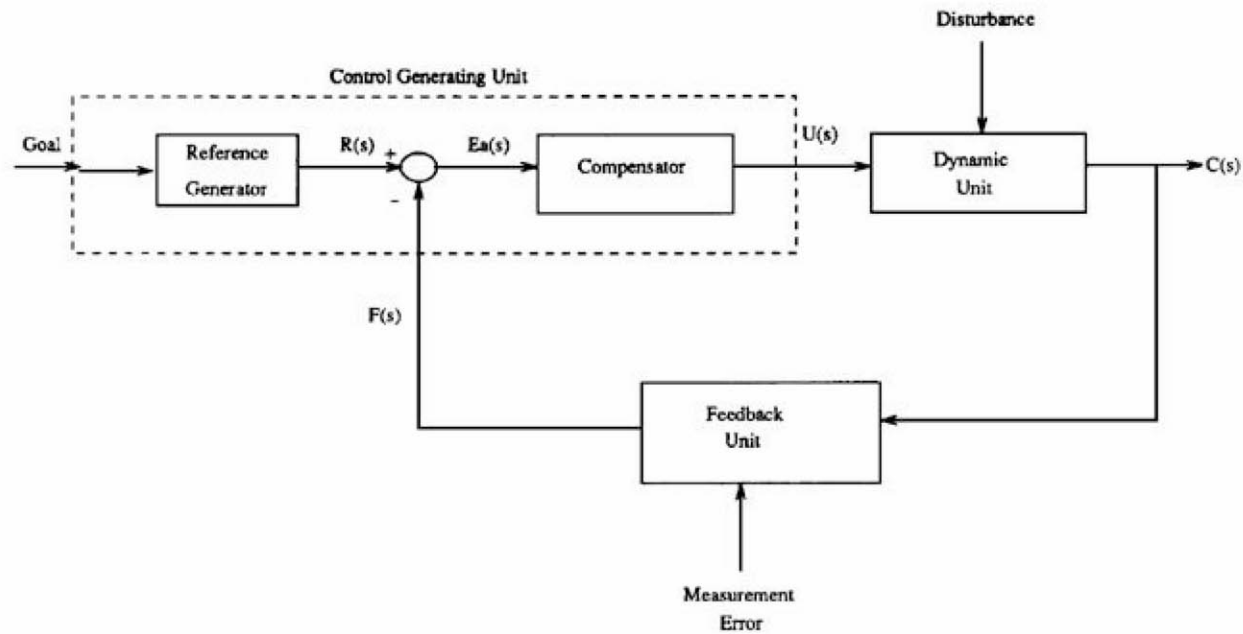
Feedback may bring certain disadvantages. Improperly applied feedback could cause instability. A loss of gain results from use of feedback and extra amplifier stages are needed. Additional high

precision components are required to provide the feedback signals.

## 152.2 Fundamentals of Feedback for Time-invariant Linear Systems

When all elements in the feedback loop of Fig. 152.1 can be described by linear, constant coefficient differential equations—Laplace transforms and transfer functions are convenient. This section deals with such systems. Digital or sampled data systems, described by difference equations, can be dealt with in similar ways using Z-transforms. When systems are high-order, time-variable, or nonlinear, the differential or difference equations must be dealt with directly, usually in state-variable format [Brogan, 1991].

**Figure 152.1** Block diagram of a generic feedback control system.



Let the forward loop transfer function from the **actuating signal** to the **dynamic unit** output be

$$C(s)/E_a(s) = KG(s) = Kg_n(s)/g_d(s) \quad (152.1)$$

where  $g_n(s)$  and  $g_d(s)$  are numerator and denominator polynomials. Let the transfer function of the feedback or **measurement unit** be

$$F(s)/C(s) = H(s) = h_n(s)/h_d(s) \quad (152.2)$$

The **open-loop** transfer function is defined as  $KG(s)H(s)$ . The actuating signal is  $E_a(s) = R(s) - F(s) = R(s) - H(s)C(s)$ . The term **error signal** will be reserved for the

difference between the input and the output:

$$E(s) = R(s) - C(s) \quad (152.3)$$

If  $H(s) = 1$ , then  $E(s) = E_a(s)$ . Using  $C(s) = KG(s)E_a(s)$ , algebraic manipulations give the **closed-loop** transfer functions from  $R(s)$  to  $C(s)$ ,  $E_a(s)$ , and  $E(s)$ , respectively, as

$$\begin{aligned} C(s)/R(s) &= KG(s)/[1 + KG(s)H(s)] \\ &= Kg_n(s)h_d(s)/[g_d(s)h_d(s) + Kg_n(s)h_n(s)] \end{aligned} \quad (152.4)$$

$$E_a(s)/R(s) = 1/[1 + KG(s)H(s)] \quad (152.5)$$

$$E(s)/R(s) = \{1 - KG(s)[1 - H(s)]\}/[1 + KG(s)H(s)] \quad (152.6)$$

External disturbances are ignored to this point. Knowledge of closed-loop pole locations allows prediction of stability, response time, and frequency of oscillation. Since Eqs. (152.4), (152.5), and (152.6) have the same denominators,  $c(t)$ ,  $e_a(t)$ , and  $e(t)$  all have the same types of terms in their time responses, as determined by the closed-loop poles, or roots of the characteristic equation

$$1 + KG(s)H(s) = 0 \quad \text{or} \quad g_d(s)h_d(s) + Kg_n(s)h_n(s) = 0 \quad (152.7)$$

The poles can be manipulated by choice of the gain  $K$ . This explains the ability of closed-loop systems to improve transient response compared with open-loop systems. Examples are provided later.

An open-loop system consisting of just the forward path  $KG(s)$  has the error

$$E(s) = R(s) - C(s) = [1 - KG(s)]R(s) \quad (152.8)$$

The equivalent closed-loop error is obtained from Eq. (152.6). The final value theorem of Laplace transforms allows evaluation of the steady state error value in either case. Traditionally, these errors are evaluated using step, ramp, and parabolic test input functions. The resulting errors are zero, finite, or infinite depending on the system type, that is, the number of poles of  $G(s)H(s)$  at the origin. From Eq. (152.8), an open-loop system will have a finite steady state error only if  $G(s)$  is type 0, and a zero error requires  $K = 1/G(0)$ . The feedback system steady state error is reduced by increasing the loop gain  $K$ , provided that all poles remain in the left-half plane. The position, velocity, and acceleration error constants (alternately step, ramp, and parabolic error coefficients) [D'Azzo and Houpis, 1988] are useful in evaluating these errors:

$$K_p = \lim_{s \rightarrow 0} [KG(s)H(s)] \quad (152.9)$$

$$K_v = \lim_{s \rightarrow 0} [sKG(s)H(s)] \quad (152.10)$$

$$K_a = \lim_{s \rightarrow 0} [s^2KG(s)H(s)] \quad (152.11)$$

The denominator in Eqs. (152.4), (152.5), and (152.6) is called the **return difference**  $R_d(s) = 1 + KG(s)H(s)$ . The return difference plays a major role in parameter sensitivity analysis. For example, sensitivity to perturbations in the forward transfer function  $G(s)$  is characterized by  $S_G = (\delta C/C)/(\delta G/G) = (\delta C/\delta G)(G/C)$ . For the open-loop system of Eq. (152.1)  $S_G = 1$ , meaning that the effect of a parameter error is reflected full strength into the output  $C(s)$ . The closed-loop system of Eq. (152.4) yields  $S_G = 1/R_d(s)$ . This sensitivity is reduced as the return difference is increased. The sensitivity to errors in  $H(s)$  is  $S_H = -KGH/(1 + KGH)$ . For high gain,  $S_H = -1$ , and  $C(s)/R(s) \rightarrow 1/H(s)$ . This means that performance depends almost exclusively on the feedback components. Using feedback with inaccurate components,  $H(s)$ , is not productive!

A feedback control system's ability to reduce the effect of external disturbances depends on the location where the disturbance enters the loop. Figure 152.2 shows a system with five separate disturbances. The return difference for this system is

$$R_d(s) = 1 + G_1(s)G_2(s)H_1(s)H_2(s) \quad (152.12)$$

For this linear system, superposition gives

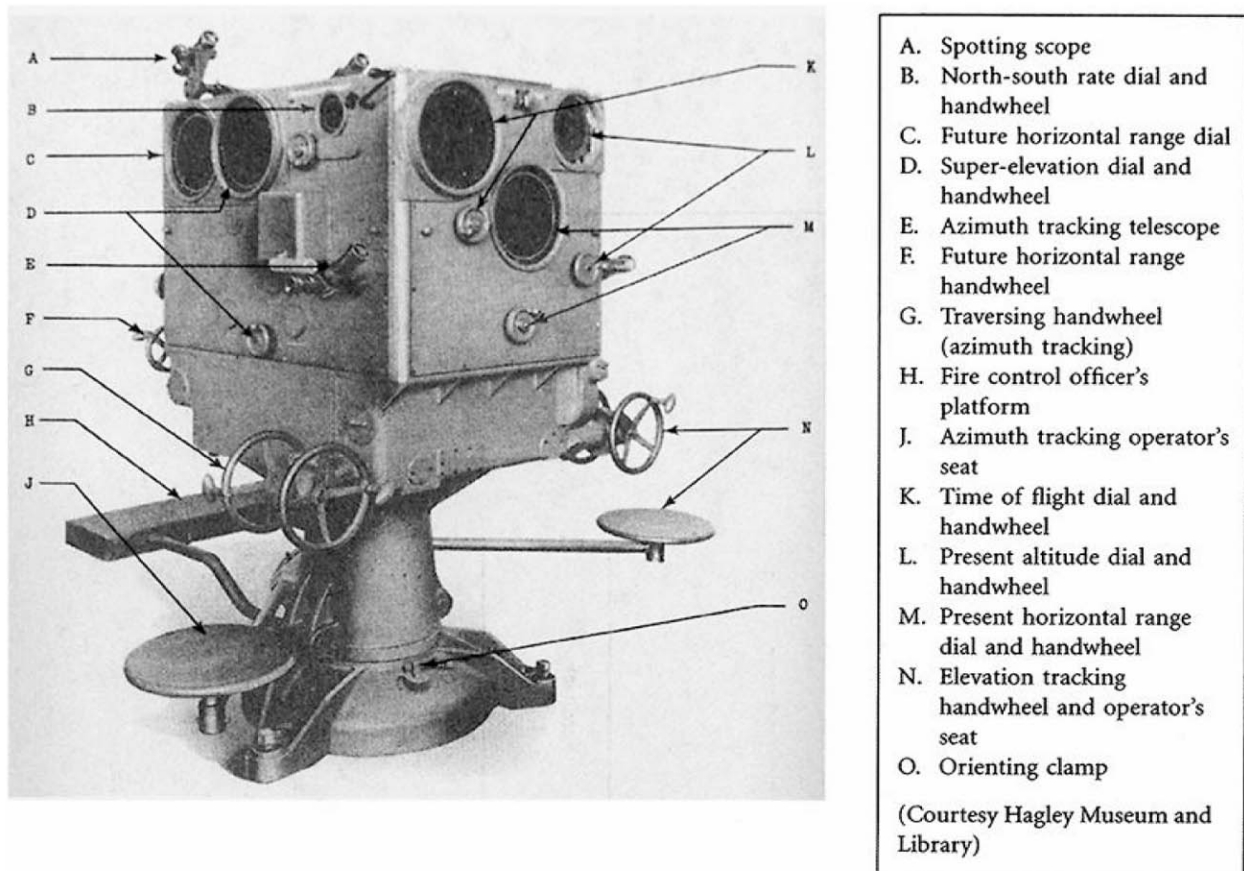
$$C(s) = [G_1G_2(R + D_1) + G_2D_2 + D_3 - H_1H_2G_1G_2D_4 - H_2G_1G_2D_5]/R_d(s) \quad (152.13)$$

When the return difference is much larger than unity,

$$C(s) \approx [R + D_1 + D_2/G_1 + D_3/(G_1G_2) - D_4H_1H_2 - D_5H_2]/(H_1H_2) \quad (152.14)$$

The desired output is caused by the reference input  $R$ . Contributions due to all disturbances  $D_i$  are undesirable. Since  $D_1$  enters the loop at the same point as  $R$ , nothing can be done with loop parameters to separate the two effects. The effect of  $D_2$  can be reduced if  $G_1$  has a high gain. The effect of  $D_3$  is reduced by increasing either  $G_1$  or  $G_2$ . Disturbance  $D_4$  has a smaller effect if  $H_1H_2$  is small. Finally,  $D_5$  effects are reduced if  $H_2$  is small. In all cases, the effect of a disturbance is reduced if the gain between the point of disturbance insertion and the output is small compared with the return difference. For a more complete look at disturbance rejection, see D'Azzo and Houpis [1988].

**Figure 1** The Sperry T-6 Director



## EARLY DISTRIBUTED CONTROL: ANTI-AIRCRAFT COMPUTERS AT SPERRY IN THE 1930S

### D. Mindell

*Massachusetts Institute of Technology*

From 1925 to 1940, the Sperry Gyroscope Company developed gun directors for controlling anti-aircraft guns. These devices were mechanical analog computers that connected four 3-inch anti-aircraft guns into an integrated system. Two tracking telescopes and a stereo rangefinder provide azimuth, elevation, and range data. The computer receives these data; incorporates manual adjustments for wind velocity, wind direction, muzzle velocity, air density, and other factors; and outputs three variables: azimuth, elevation, and a setting for the fuse. Manually set before loading, the fuse setting determines the time after firing at which the shell will explode (corresponding to slant range of the predicted position of the target). Shells are not intended to hit the target plane directly but rather to explode near it, scattering fragments to destroy it. The computer thus performs two main calculations: prediction and ballistics. Prediction corresponds to calculating the "lead" of the target, finding its position at some future time. Ballistic calculations correspond to looking up solutions in a "firing table," which determines the necessary gun elevation for a desired range. Here the firing table was mechanized by a "ballistic cam" or pin follower. The prediction and ballistics solution are coupled, since prediction requires a "time of flight" for the shell, which depends on the ballistics. A feedback loop thus connects the two calculations.



Figure 1 shows the Sperry T-6 director, built in 1930. A square box about four feet on a side, it was mounted on a pedestal on which it could rotate. The crew consisted of nine men. Three would sit on seats, and one or two would stand on a step mounted to the machine, revolving with the unit as the azimuth tracker followed the target. The remainder of the crew stood on a fixed platform; they would have had to shuffle around as the unit rotated. This was probably not a problem, as the rotation angles were small for any given engagement. The director's pedestal mounted on a trailer, on which data transmission cables and the rangefinder could be packed for transportation. Nearly all the work for the crew was in a "follow-the-pointer" mode: each man concentrated on an instrument with two indicating dials, one the actual and one the desired value for a particular parameter. With a hand crank he adjusted the parameter to match the two dials.

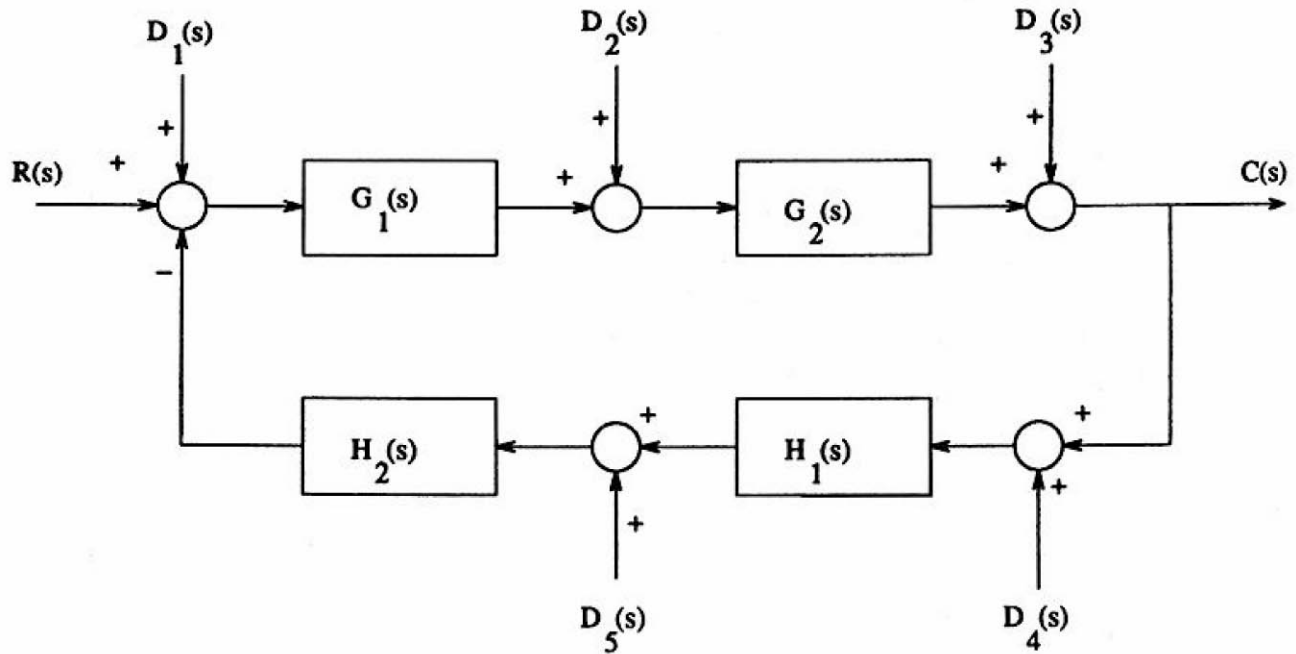
Even though the T-6 director took only three inputs, it still required nine operators. Mostly, these men connected one stage of the calculation to the next, "feeding back" data into the computer in a "follow-the-pointer mode." Later, many of these operators were replaced by servomechanisms, acting as impedance amplifiers to couple stages of the mechanical computer. But in the 1930s, servos could not perform this function as well as people. The data were noisy, and even an unskilled human eye could eliminate complications due to erroneous or corrupted data. Noisy data did more than corrupt firing solutions. The mechanisms themselves were rather delicate, and erroneous input data, especially if they indicated conditions that were not physically possible, could lock up or damage the mechanisms. The operators performed as integrators in both senses of the term: They integrated different elements into a system, and they integrated mathematically, acting as low-pass filters to reduce noise.

The Sperry directors of the 1930s were transitional, experimental systems. Exactly for that reason, however, they allow us to peer inside the process of automation, to examine the displacement of human operators by servomechanisms while the process was still under way. Sperry Company only gradually became comfortable with the automatic communication of data between subsystems. The company bragged about the low skill levels required of the operators of the machine, but in 1930 it was unwilling to remove them completely from the process. Men were the glue that held integrated systems together.

**Figure 2** The Sperry T-6 Director mounted on a trailer with operators. Note power supply at left and cables to other system elements. (Courtesy Hagley Museum and Library)



**Figure 152.2** Feedback system with multiple disturbances.



## 152.3 Illustrative Applications

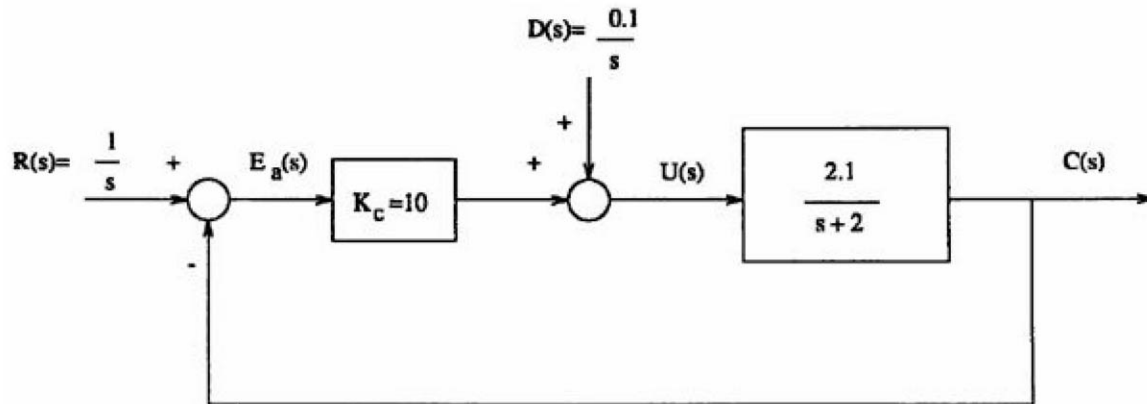
Two examples illustrate some major attributes of feedback. Suppose the speed of an automobile is ideally related to the depression of the throttle pedal by  $G(s) = 2/(s + 2)$ . A unit step input to an open-loop system leads to output  $c(t) = 1 - e^{-2t}$ . If the model is exact and no external disturbances exist, then open-loop control may suffice. However, assume that the numerator of  $G$  is actually 2.1 (a parameter error). Also, suppose that a 10% disturbance is acting so that  $R(s) + D(s) = 1.1/s$ . The open-loop response  $c(t) = 1.155(1 - e^{-2t})$  now has a 15.5% error. Figure 152.3 shows a feedback system (perhaps an automatic cruise control unit, or the driver responding to the speedometer) with the same parameter error and disturbance input. The control gain  $K_c = 10$  is arbitrarily chosen for illustration only. This type 0 system has  $K_p = 10.5$ . The steady state error (ignoring the disturbance) is  $e_{ss} = 1/(1 + K_p) = 0.08696$ . The closed-loop characteristic equation is  $(s + 23) = 0$ , indicating a much faster response. The actual output is  $c(t) = 0.92217(1 - e^{-23t})$ . The steady state error is  $-0.07783$ , roughly half the open-loop error. Figure 152.4 compares the open- and closed-loop responses. The plant input command also should be considered. In the open-loop configuration, it is a constant 1.1. In the closed-loop system, it is

$$u(t) = 0.878 + 9.22e^{-23t} \quad (152.15)$$

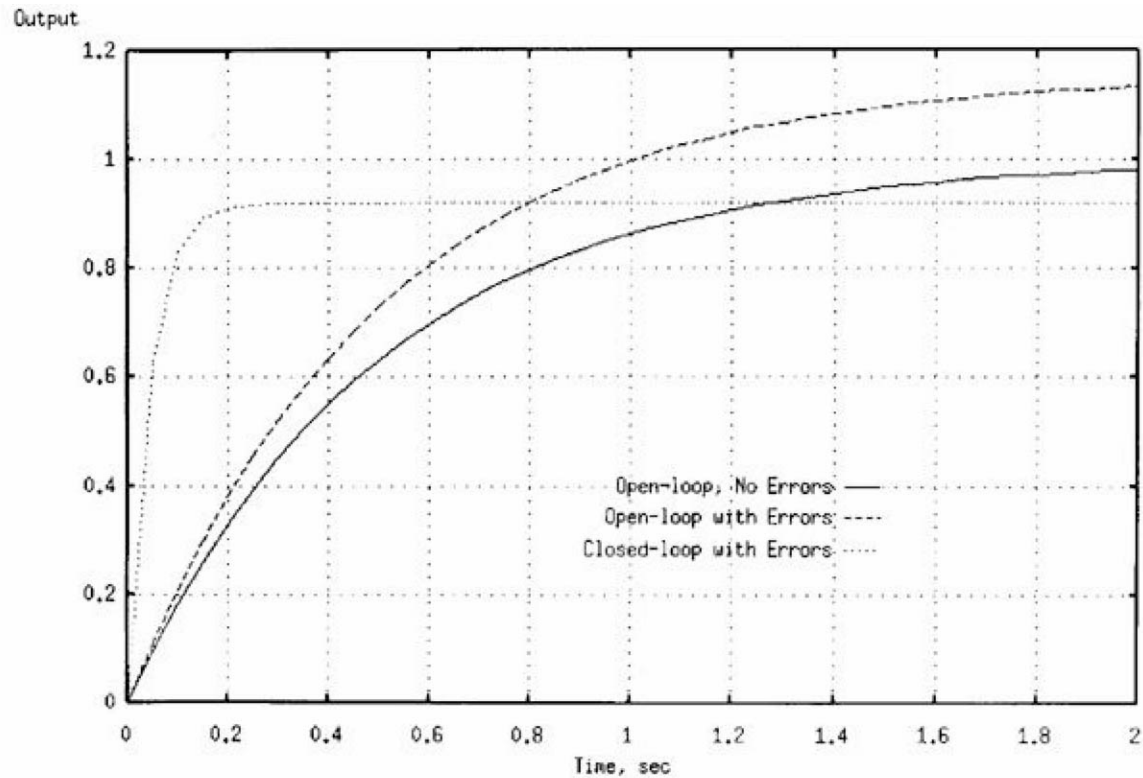
Initially, this is 9.18 times larger than the open-loop control signal, but smaller in steady state. This large initial spike causes the faster closed-loop response. Physical components may saturate, or in

other ways prevent, the achievement of performance predicted by linear mathematical models.

**Figure 152.3** Feedback system with parameter error and disturbance.



**Figure 152.4** Comparison of open- and closed-loop systems with disturbance and parameter error.



A second example is an unstable plant described by

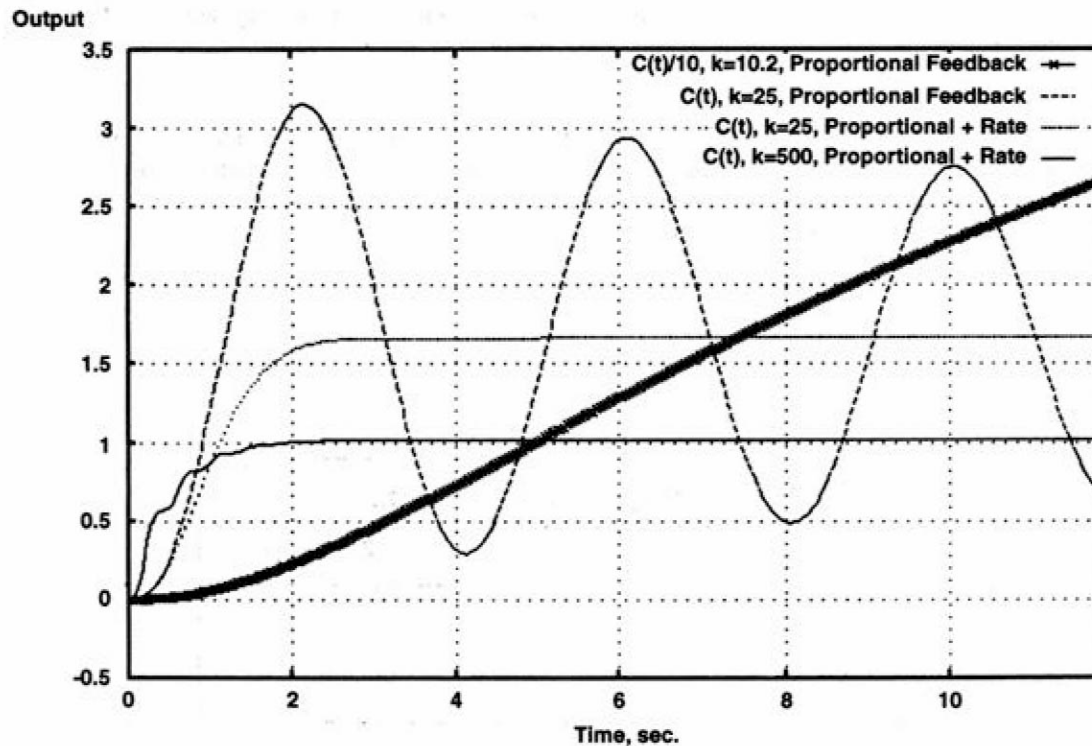
$$G(s) = K/[(s - 1)(s + 2)(s + 5)] \quad (152.16)$$

With unity feedback system, Routh's criterion [Dorf, 1989] shows that this system is stable for  $10 < K < 28$ . The steady state error decreases with increasing  $K$  values in the stable range. For  $K = 10.2$  the output smoothly approaches a final value of 51, giving a steady state error of 50. As  $K$  increases, the rise time and steady state error decrease, but overshoot and oscillations increase. Simple adjustment of  $K$  cannot produce a satisfactory response. A rate feedback term  $K_r s$  is added next, giving  $H(s) = K_r s + 1$ . The value  $K_r = 0.5$  is selected arbitrarily to cancel the pole at  $s = -2$ . Routh's criterion now shows a stable system  $K > 10$ . Figure 152.5 compares output using proportional-plus-rate feedback with proportional feedback alone. Rate feedback has reduced response time and steady state error, eliminated the overshoot, and squelched the oscillations. If a total gain  $K = 500$  is divided between the plant with a factor 25 and the controller with a factor of 20, the input to the plant  $G(s)$  is

$$u(t) = -4.08 + e^{-2t} [20.4 \cos(15.5t) + 2.629 \sin(15.5t)] \quad (152.17)$$

This large amplitude oscillation might exceed the physical limits of the plant. This example shows that feeding back output derivatives, in addition to the output, provides stronger control over the process. Extensions of this idea lead to **state feedback**, a topic of Chapter 159. Also see Brogan [1991]. A treatment of compensator design is also presented in Chapter 155.

**Figure 152.5** Comparison of proportional feedback and proportional-plus-rate feedback.



## 152.4 Feedback in Nonlinear Systems

---

Few physical systems are truly linear over broad ranges of possible inputs. Many can be treated (approximately) as linear over a limited range. This is attractive because linear systems analysis methods are more fully developed. Feedback can cause some nonlinear open-loop systems to have linear closed-loop behavior. This is sometimes called dynamic linearization [Brogan, 1991]. Other systems are intentionally made nonlinear (e.g., by using on-off controllers) to improve performance. Adaptive and self-learning control systems use feedback and are inherently nonlinear. Feedback also plays an essential role in the training of artificial neural networks through back propagation, and in Kalman filtering and recursive least squares estimation. A predicted signal is compared with a measured signal. The difference is then used to adjust values of parameters or states to reduce future errors.

## 152.5 Summary of the Typical Steps in the Design of Feedback Systems

---

The development of a feedback system is often an iterative process. Typically the following sequence of questions must be addressed:

1. What are the goals for this system?
2. What are the appropriate measures of system performance in attempting to reach these goals?
3. Are there portions of the system that are fixed by other considerations? What are the mathematical models of these units?
4. How can the feedback system be easily analyzed in terms of the goals and performance measures?
5. How can the system be modified to improve performance?
6. What are the economic factors related to the product?

A certain degree of robustness can be provided by feedback. That is, precise feedback sensors add expense but can sometimes make up for deficiencies in a cheaper forward path plant and thus might lead to a lower total cost. Systematic methods of analysis and design of feedback systems are presented in **Chapters 153–159**.

### Defining Terms

**Actuating signal:** The difference between the reference input and the feedback signal.

**Closed-loop system:** System which contains one or more feedback signals that modify the input to the dynamic unit.

**Dynamic unit:** The principal plant or process being controlled.

**Error signal:** The difference between the reference input and the dynamic system's output.

**Measurement unit:** Device that produces the feedback signal.

**Open-loop system:** A system whose input is independent of its output.

**Poles:** Values of complex variable  $s$  that are roots of a transfer function denominator.

**Return difference:** The difference between a unit signal inserted at an "opened" point in a feedback system and the signal which returns around the loop to the other side of the break.

**Zeros:** Values of complex variable  $s$  that are roots of a transfer function numerator.

## References

- Brogan, W. L. 1991. *Modern Control Theory*, 3rd ed. Prentice Hall, Englewood Cliffs, NJ.
- D'Azzo, J. J. and Houpis, C. H. 1988. *Linear Control System Analysis and Design*, 3rd ed. McGraw-Hill, New York.
- DiStefano, J. J., Stubberud, A. R., and Williams, I. J. 1967. *Theory and Problems of Feedback and Control Systems*. Schaum, New York.
- Dorf, R. C. 1989. *Modern Control Systems*, 5th ed. Addison-Wesley, Reading, MA.
- Franklin, G. F., Powell, J. D., and Emami-Naeini, A. 1986. *Feedback Control of Dynamic Systems*, Addison-Wesley, Reading, MA.

## Further Information

- Automatica*. A journal of IFAC, the International Federation of Automatic Control. Published bimonthly by Pergamon Press, Ltd., Oxford, England.
- IEEE Transactions on Automatic Control*. Published monthly by the Institute of Electrical and Electronic Engineers.
- IEEE Transactions on Control Systems Technology*. Published quarterly by the Institute of Electrical and Electronic Engineers.
- International Journal of Control*. Published monthly by Taylor and Francis, London, England.
- Proceedings of Conference on Decision and Control*. This annual conference sponsored by the IEEE Control Systems Society presents latest developments.
- Proceedings of the American Control Conference*. This annual conference is jointly sponsored by seven societies.

Brogan, W. L. "Root Locus"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

[153.1 The Root Locus Concept](#)[153.2 Root Locus Details](#)[153.3 Generation of Root Locus Plots](#)[153.4 Examples](#)[153.5 Summary and Conclusions](#)**William L. Brogan***University of Nevada, Las Vegas*

Evans [1950] introduced the root locus method of control systems analysis. The temporal behavior of a closed-loop system depends upon the location of the closed loop poles, or the roots of

$$1 + KG(s)H(s) = 0 \quad (153.1)$$

In a single loop control system,  $KG(s)$  is the forward path transfer function and  $H(s)$  is the feedback path transfer function. Multiple loop systems have more complicated characteristic equations that can be found using Mason's gain formula [D'Azzo and Houpis, 1988]. The closed-loop transfer function for a single loop feedback system is given in **Chapter 152** [Eq. (152.4)]. The denominator of that equation is the left side of Eq. (153.1). Root locus is a method for finding the  $n$  closed-loop poles [i.e., the roots of Eq. (153.1)] as the parameter  $K$  is varied. Partial fraction expansion and Laplace transform inversion give the control system time response, with one term due to each closed-loop pole. Major characteristics of the temporal behavior are determined solely by the pole locations. This explains the importance of root locus in control system design. Root locus procedures apply equally well to systems described by Z-transforms. The interpretation of root locations in the  $s$  and  $Z$  planes differ, but the process of plotting the roots is the same. This discussion is presented in terms of the Laplace  $s$  domain and is restricted to positive values of  $K$ .

## 153.1 The Root Locus Concept

---

Rearrangement of Eq. (153.1) gives  $KG(s)H(s) = -1$ , indicating that the complex number  $KG(s)H(s)$  must satisfy two conditions at any point  $s$  that is a root of Eq. (153.1):

The angle  $\phi$  must be an odd multiple of  $180^\circ$ ,  $\phi = \pm(2k + 1)180^\circ$ .

The magnitude  $|KG(s)H(s)|$  must be unity.



The root locus is the set of points in the complex plane that satisfies the angle condition. Let the numerator and denominator of  $KG(s)H(s)$  be of degrees  $m$  and  $n$ , respectively. Then, Eq. (153.1) has  $n$  roots, and the root locus has  $n$  branches. Evans developed a set of rules for sketching the paths that these  $n$  roots trace out as  $K$  is varied. Rough sketches are easy to draw and provide great insight into system behavior. The computer is used when complete and accurate plots are needed. Root locus uses the angle condition to find the root locations as functions of  $K$ . Then, the magnitude condition is used to find the gain  $K = 1/|G(s_1)H(s_1)|$  at a point  $s_1$  on the locus. Major root locus construction rules are now listed.

## 153.2 Root Locus Details

---

The numerator and denominator of the open-loop transfer function  $G(s)H(s)$  must be factored before the root locus method can be applied. The basis for root locus is the geometric interpretation of complex numbers represented as vectors. Let the numerator and denominator factors of  $G(s)H(s)$  be  $s + a_i$ , with the constants  $a_i$  either real or complex. Point  $s_1$  can be represented by the vector from the origin to  $s_1$ . The vector from  $-a_i$  to the origin is  $a_i$ . Addition gives  $s_1 + a_i$ , the vector from the pole or zero at  $-a_i$  to  $s_1$ . Each such vector is expressed in polar form, with magnitude  $\rho_i$  and angle  $\beta_i$ . The total angle of  $G(s_1)H(s_1)$  is the sum of  $\beta_i$  angles from the numerator terms, minus the  $\beta_i$  angles from the denominator terms. This total must equal  $(2k + 1)180^\circ$  (negative  $K$  would add  $180^\circ$  and require a change to some multiple of  $360^\circ$  to find the negative  $K$  locus). The angles can be measured with a protractor at a test point  $s_1$  and added.

The following rules simplify root locus plotting.

1. *Number of branches.* There are  $n$  branches of the root locus, where  $n$  is the number of open-loop poles  $G(s)H(s)$ .

2. *Loci starting points.* One branch of the positive gain root locus "begins" at each open-loop pole [poles of  $G(s)H(s)$ ]. These are the  $K = 0$  points.

3. *Locus behavior for large gain.* As  $K \rightarrow \infty$ , the product  $KGH$  must remain finite, which requires  $GH$  to be zero. This is true at the  $m$  open-loop zeros. As a result, as  $K \rightarrow \infty$ , one locus path approaches each of the  $m$  open-loop zeros. The remaining  $n - m$  paths go to  $\infty$ . When  $s$  is very large,  $KG(s)H(s) \rightarrow K/s^{n-m}$ .

4. *Asymptotic angles.* The angle condition at a test point  $s_1$  on an infinite circle shows that the angle to  $s_1$  must be  $\varphi_k = (2k + 1)180^\circ / (n - m)$  for some integer  $k$ .

5. *Loci ending points.* The  $K = \infty$  points are where the root loci branches "end." One branch of the root locus ends at each finite zero of  $G(s)H(s)$ . The remaining  $n - m$  branches go to infinity, with asymptotic angles  $\varphi_k$ .

6. *Centroid or center of gravity.* The vectors from the open-loop poles and zeros to an infinite point  $s_1$  on the locus appear to come from a common point  $\sigma$  on the real axis, called the centroid or center of gravity. Note that  $\sigma = \{\sum[\text{finite poles of } G(s)H(s)] - \sum[\text{finite zeros of } G(s)H(s)]\} / (n - m)$ . The  $n - m$  asymptotic rays emanating from the centroid  $\sigma$  are not part of the root locus. A root locus branch approaches each ray as  $K \rightarrow \infty$ .

7. *Real axis symmetry.* The root locus is mirror image symmetric with respect to the real axis because complex roots occur in conjugate pairs.

8. *Root loci on the real axis.* The angle condition shows that (for  $K \geq 0$ ) all sections of the real axis lying to the left of an odd number of open-loop poles or zeros are part of the root loci. There will be a closed-loop pole at each point in each such interval for some unique value of  $K$ . These real axis segments might be the composite of several nonoverlapping branches of the locus if breakaway points exist.

9. *Breakaway points of the root loci.* Breakaway points indicate multiple roots. The gain at such a point is called a critical gain. As  $K$  increases toward a critical value, two or more root paths converge or "break in" to the point from different directions. A further gain increase causes the multiple paths to "break away" from each other. The angle possibilities are given in Rule 11. At a critical  $s$ ,  $K$  has a relative maximum or minimum such that

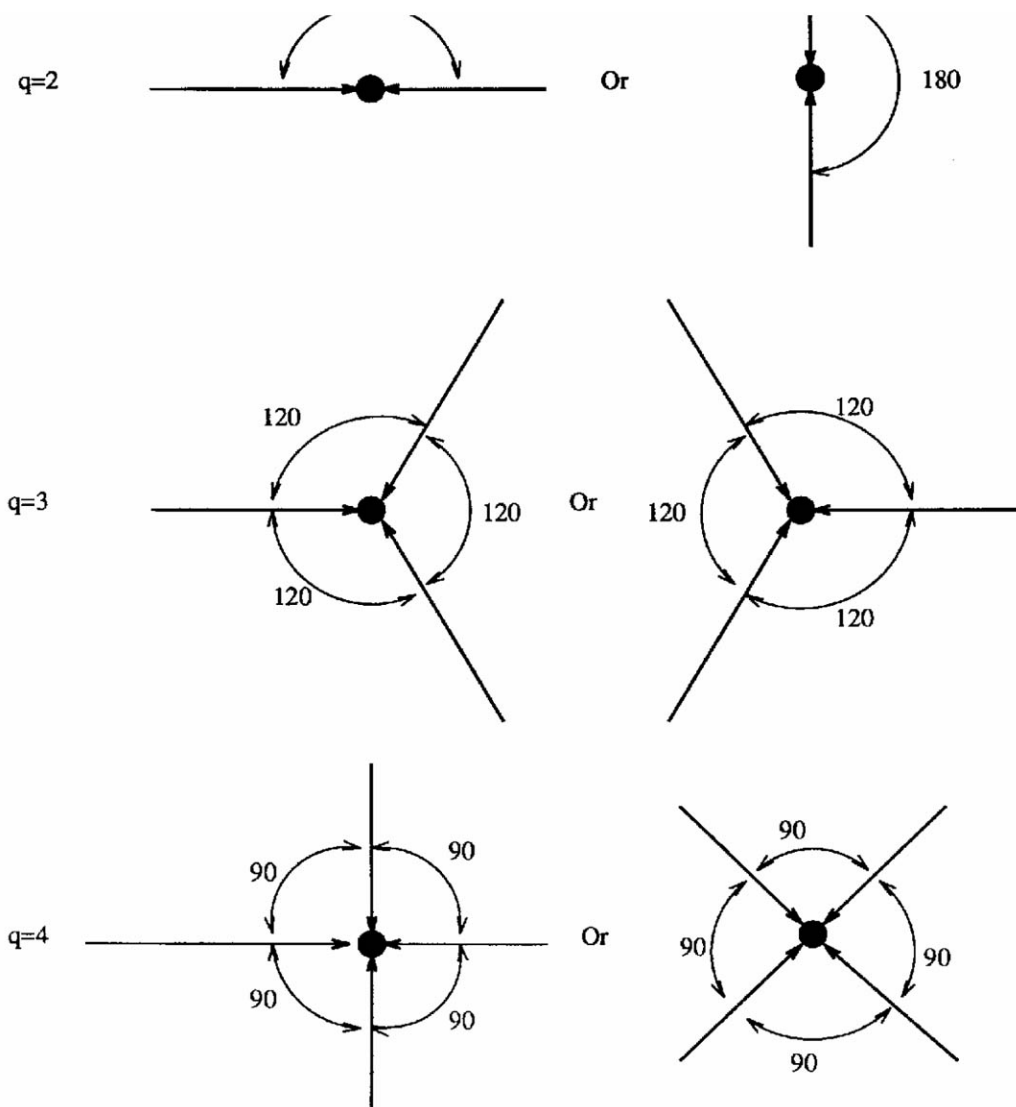
$$d[G(s)H(s)]/ds = 0 \quad \text{or} \quad d[1/G(s)H(s)]/ds = 0.$$

The solution points can be either real (real axis break in or breakaway points) or complex (points where two branches intersect).

10. *Angles of arrival or departure for complex zeros or poles.* Consider a test point  $s_1$  on an infinitesimal circle centered at a complex open-loop pole or zero. Vectors from all other poles or zeros to this point have known angles. The angle criterion gives the angle of the infinitesimal vector from the complex pole or zero being tested. This is the arrival (zero) or departure angle (pole).

11. *Path angles at breakaway points.* If  $s_1$  is the location of a  $q$ th multiple root, then  $q$  branches of the locus break in at that point along paths uniformly spaced with angular separation  $360^\circ/q$ . As  $K$  increases beyond the critical value, the  $q$  branches break away along different paths also separated by  $360^\circ/q$ . [Figure 153.1](#) shows the only arrangements possible for real axis break in points with  $q = 2, 3$ , and  $4$ . One option in [Fig. 153.1](#) applies to the break in angles and the other (with reversal of the arrows) applies to the breakaway. A multiple open-loop pole is a special case ( $K = 0$ ) source of multiple paths. A multiple open-loop zero is a special case ( $K = \infty$ ) sink for multiple paths.

**Figure 153.1** Possible break in angles.



12. *Imaginary axis cross-over.* Routh's criterion [Dorf, 1989] can be used to find the  $j\omega$  axis locus cross-over points and the corresponding gain values.

13. *Conservation of sum of open- and closed-loop poles.* If  $n \geq m + 2$ , the sum of the closed-loop poles is constant for all  $K$  [D'Azzo and Houpis, 1988]. The sum is known from the open-loop poles.

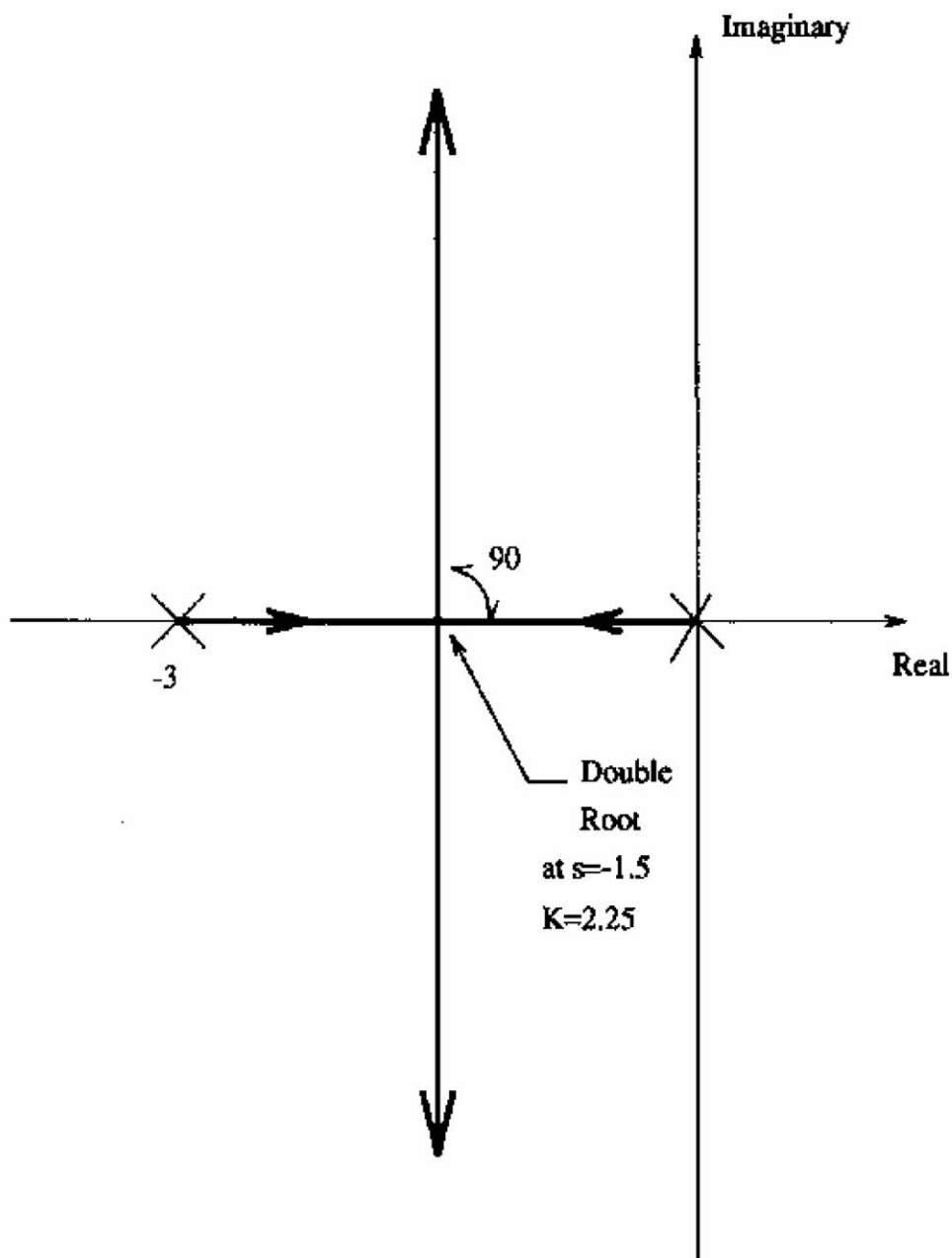
## 153.3 Generation of Root Locus Plots

An approximate root locus plot can be sketched using just the easier-to-apply rules (1–8, 10, 11). A helpful analogy treats the fixed open-loop zeros as negatively charged particles and the closed-loop poles as positive charges. The attraction between unlike charges (repulsion between like charges) often helps visualize the shape of a root locus. An accurate root locus plot requires determination of break away and imaginary axis cross-over points (using Rules 9 and 12). Additional points must be found that satisfy the angle condition. This can be done by trial-and-error using a protractor or a spirule. This tedious effort is generally avoided by using a computer. Many root locus software packages exist, differing in the ease of use and quality of output.

## 153.4 Examples

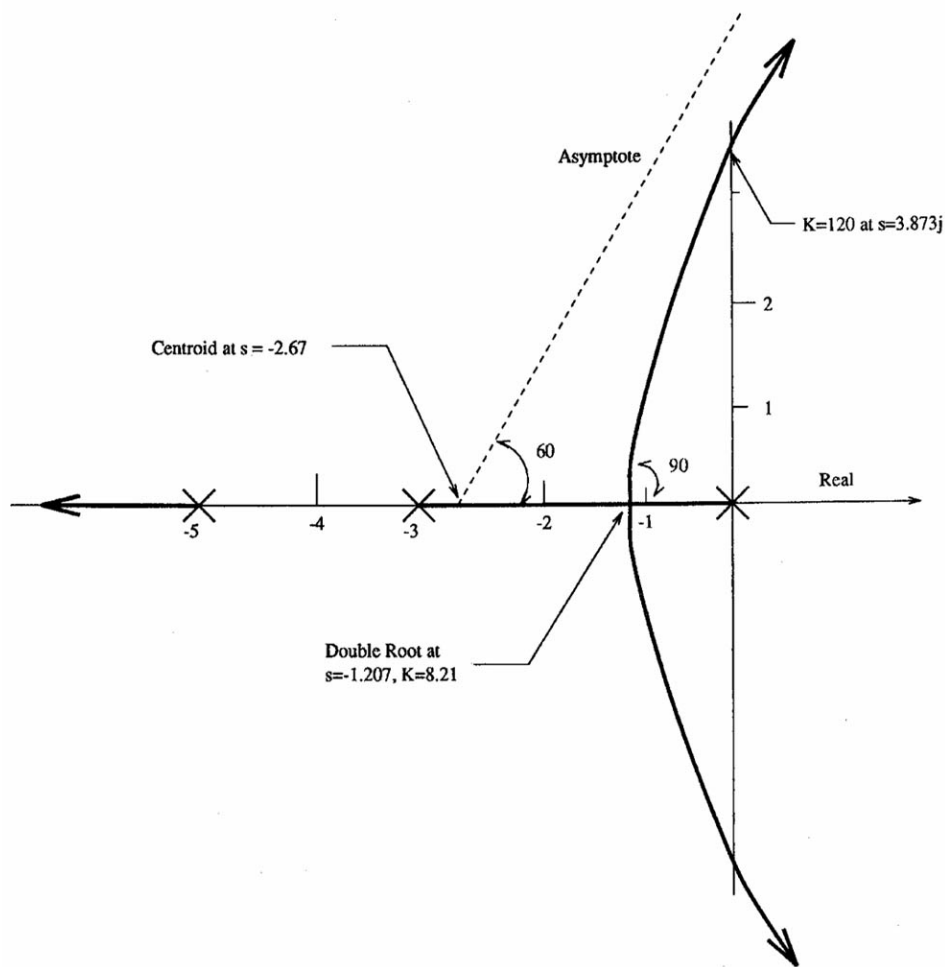
Figure 153.2 gives the root locus for  $KG(s)H(s) = K/[s(s+3)]$ . This second order system has  $n = 2$  branches and  $m = 0$  zeros. Both branches approach  $\infty$  as  $K$  increases, along asymptotes at  $\pm 90^\circ$  from the centroid at the midpoint,  $s = -1.5$  (Rules 1–8). Rule 9 shows a break away point, also at  $s = -1.5$ , meaning that a double real root occurs there. The magnitude condition gives  $K = \rho_1 \rho_2 = 2.25$ . For  $K < 2.25$ , there are two unequal real roots between  $s = 0$  and  $-3$ . For  $K > 2.25$ , there are two complex conjugate roots. The angles at which the branches break away from the real axis are  $\pm 90^\circ$  (Rule 11). In this case, the complex loci coincide with the asymptotic angles. Rule 10 does not apply. Rule 12 finds no imaginary axis cross-overs. The sum of the roots is  $-3$  for all  $K$  (Rule 13).

**Figure 153.2** Stable second order system showing break away angles at double root.

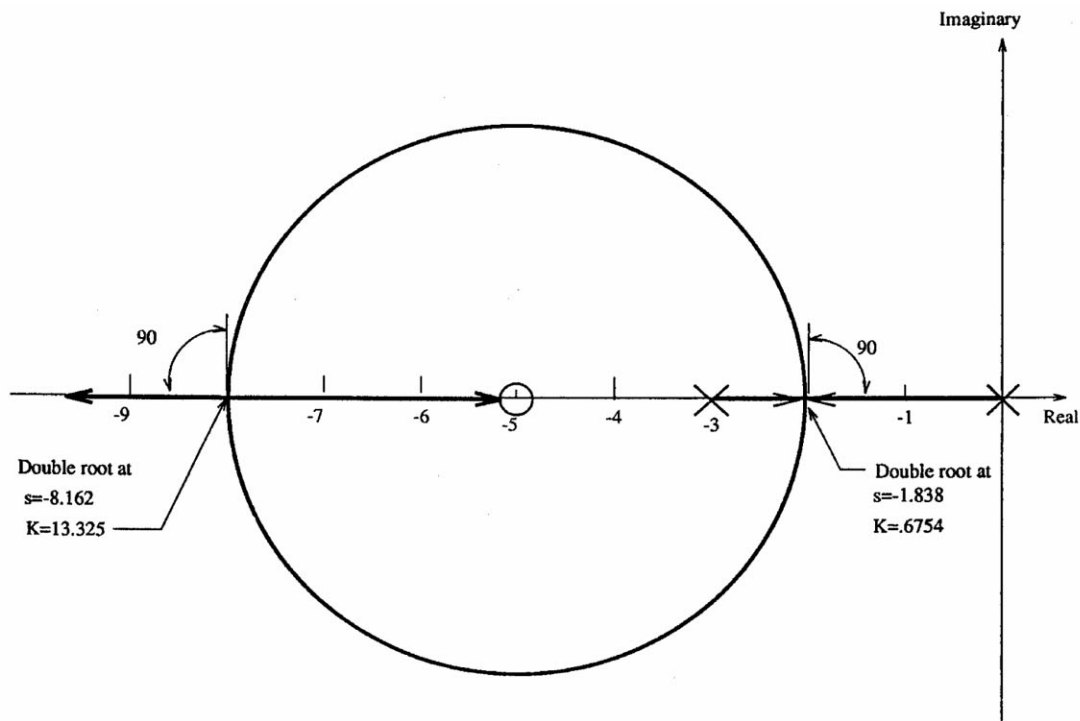


An open-loop pole is added at  $s = -5$ , giving the root locus of Fig. 153.3. The new asymptotes are  $\pm 60^\circ$  and  $180^\circ$ , meaning that the third branch lying on the real axis to the left of  $s = -5$  goes to  $\infty$ . The other two asymptotes emanate from the centroid at  $s = -2.67$ . The charged particle analogy predicts that the added pole will push the breakaway point to the right of its former midpoint. The exact breakaway point can be found from Rule 9 ( $s = -1.207, K = 8.21$ ). The angles at which the loci leave the real axis are still  $\pm 90^\circ$ . Rule 12 gives the imaginary axis cross-over points ( $s = \pm j3.873, K = 120$ ). The closed-loop system will have damped oscillations for  $8.21 < K < 120$  and will be unstable for  $K > 120$ . From Rule 12, if two complex roots are selected with real parts at  $s = -1$ , the third root location will be at  $s = -6$  because the pole sum is  $-8$ . Placing a zero instead of a pole at  $s = -5$  gives the root locus in Fig. 153.4. The charged particle analogy predicts the attraction toward the zero will cause a leftward shift of the loci of Fig. 153.2. There are two loci branches. One terminates at the zero. The other goes to  $\infty$  along the negative real axis (Rules 3, 4, 5, and 8). Rule 9 gives multiple solutions:  $s = -1.838, K = 0.6754$  (breakaway point) and  $s = -8.162, K = 13.325$  (break in point). The system has complex poles (will oscillate) for  $0.6754 < K < 13.325$ , double real poles with  $K$  at either limit, and real distinct poles for all other positive  $K$ . The break away and break in angles are both  $\pm 90^\circ$ . The complex part of the loci is a circle centered at the zero (positive charges in orbit about a negative charge).

**Figure 153.3** Third order root locus, no finite zeros.

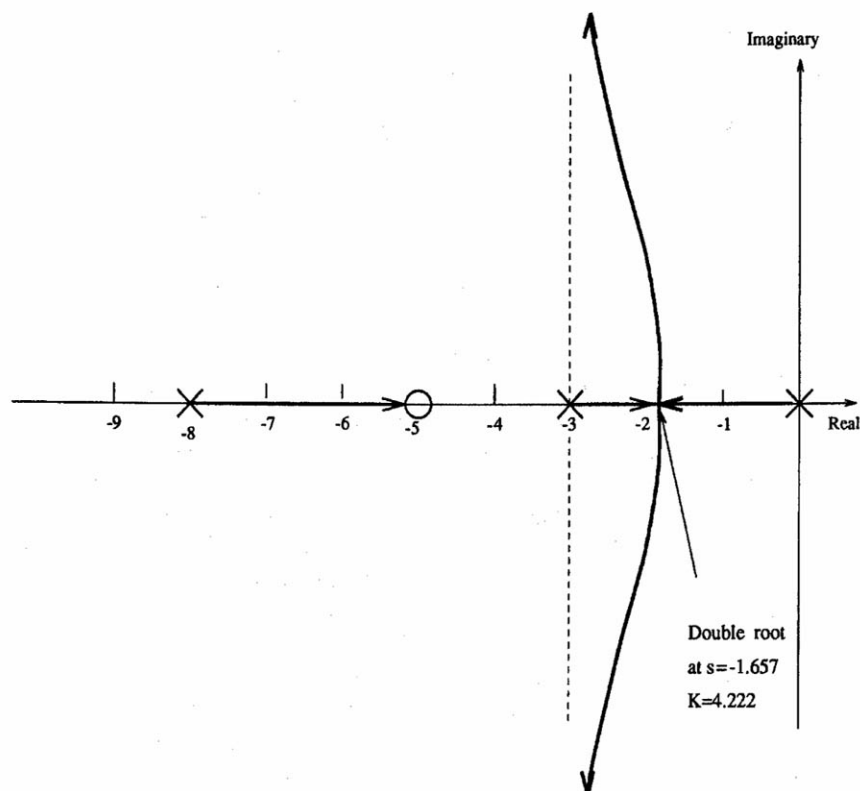


**Figure 153.4** Second order system with one zero. Double real roots for two different gains.

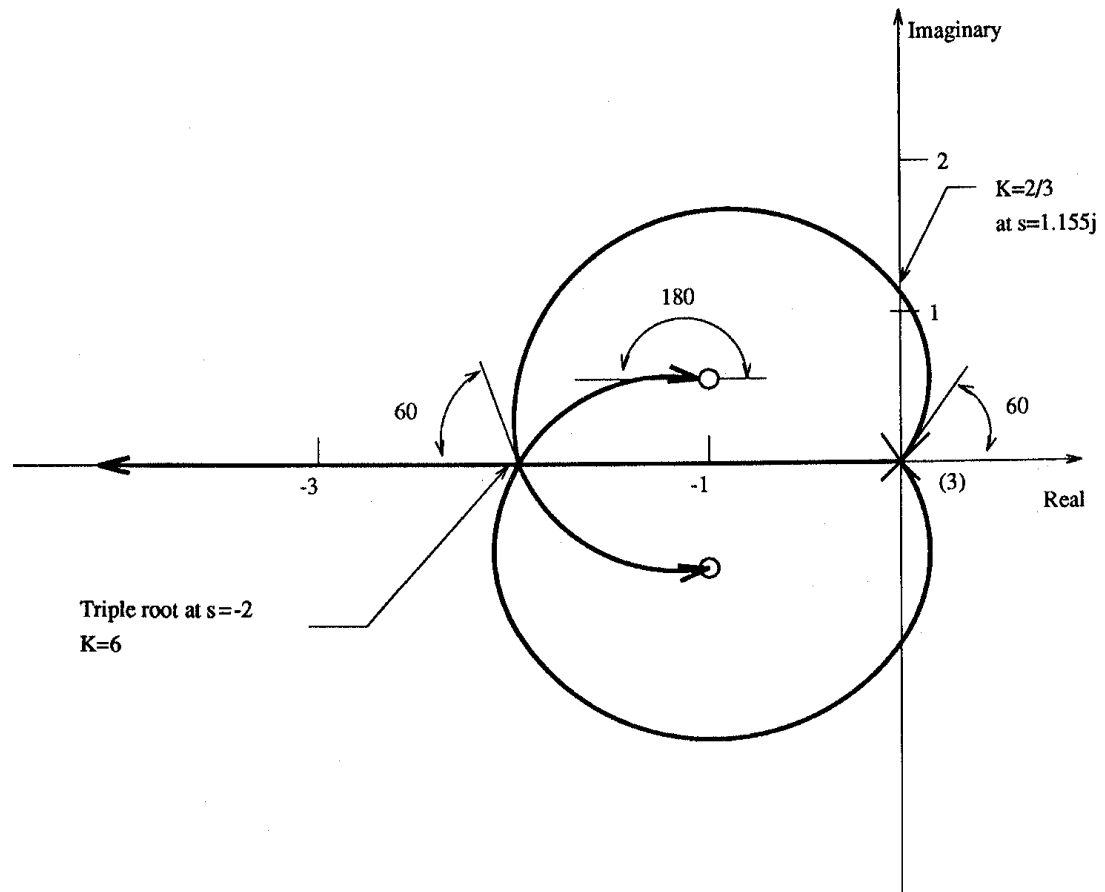


Adding a zero at  $s = -5$  and a pole at  $-8$  to the system of Fig. 153.2 gives the modified locus of Fig. 153.5. The asymptotes are  $\pm 90^\circ$ , with centroid at  $s = -3$ . The charged particle analogy correctly predicts a leftward deformation of the loci of Fig. 153.2. The zero at  $-5$  is closer and exerts stronger attraction than the repulsion exerted by the pole at  $-8$ . Rule 9 would give the exact breakaway point, but this requires factoring a cubic.

**Figure 153.5** Third order root locus, one finite zero.



**Figure 153.6** Conditionally stable third order system with complex zeros.

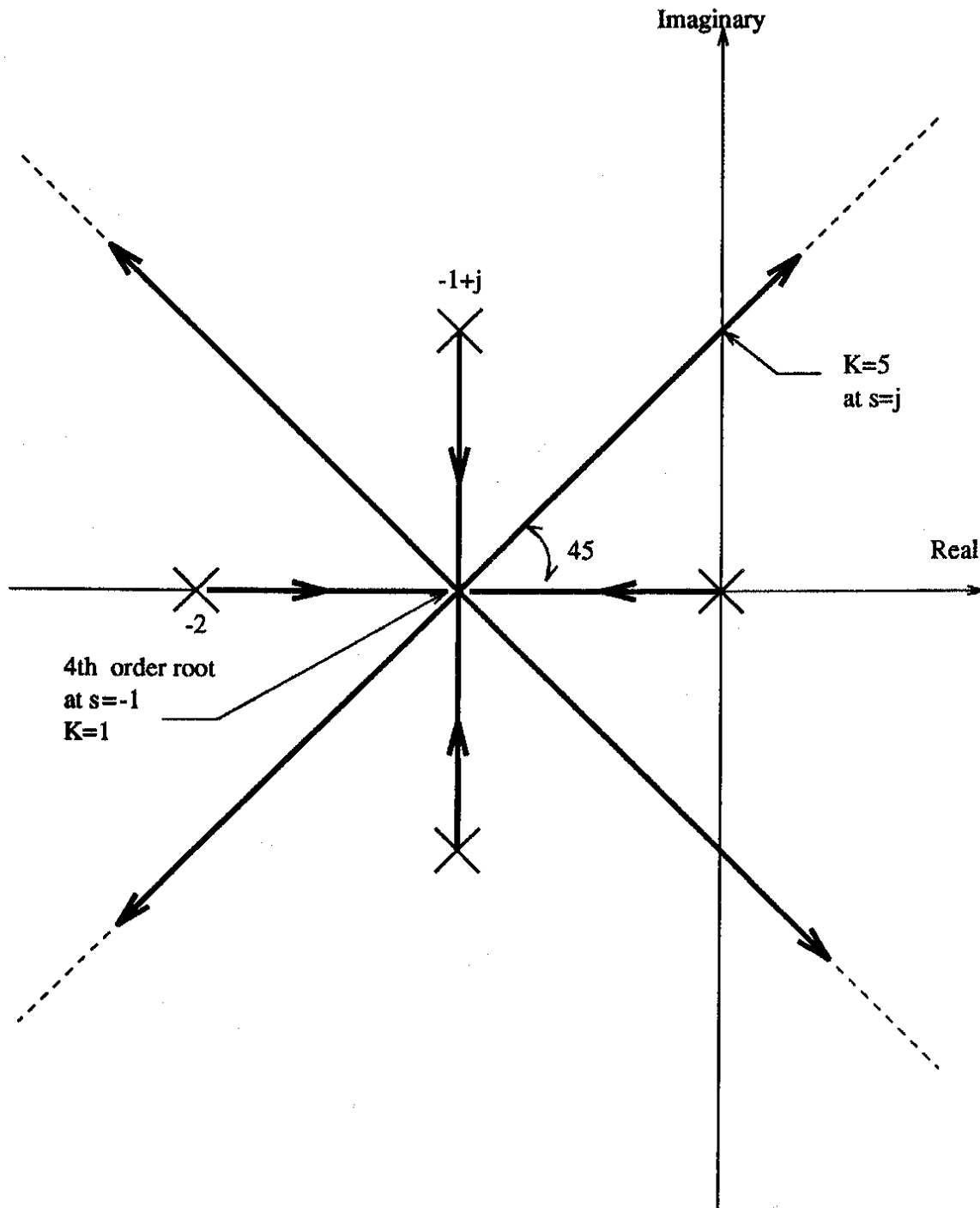


A system with  $KG(s)H(s) = K[s^2 + 2s + 4/3]/s^3$  is considered. The zeros are located at  $s = -1 \pm j/\sqrt{3}$ . There are  $n = 3$  paths, all starting at the triple pole at the origin and breaking away with angles  $\pm 60^\circ, 180^\circ$  (Rule 11). Rule 12 gives imaginary axis cross-overs at  $s = \pm 2j/\sqrt{3}$  with  $K = 2/3$ . One branch of the locus ends at each zero, and the angle of arrival (Rule 10) is  $180^\circ$ . The third branch goes to  $\infty$  along the negative real axis. This example illustrates a real axis break in and breakaway not at a pole or zero. Setting  $K = 6$  gives a triple root at  $s = -2$ . Rule 11, with knowledge that the real axis branch approaches from the right, shows that the break in angles are  $0^\circ$  and  $\pm 120^\circ$ . These paths continue and break away from the triple root point at angles  $180^\circ, \pm 60^\circ$ .

Figures 153.7, 153.8, and 153.9 illustrate root locus variations as the imaginary components of the complex open-loop poles are varied slightly. In each case, there are four branches, all ending at  $\infty$  along asymptotic angles  $\pm 45^\circ$  and  $\pm 135^\circ$  with a centroid at  $s = -1$ . The angle of departure from the positive complex pole is downward (i.e.,  $-90^\circ$ ). Figure 153.7 has a single break in point at  $s = -1$ , indicating a fourth order pole. The gain there,  $K = 1$ , is found by multiplying the four vector lengths to that point. Figure 153.1 provides angle information. The incoming angle choice is clear because two real axis branches approach the critical point. The outgoing paths coincide with the asymptotic rays. Similar considerations in Fig. 153.8 show a double real root at  $s = -1$  when  $K = 1.21$ , and a repeated pair of complex conjugate roots at  $s = -1 \pm 0.324j$  when  $K = 1.221$ .

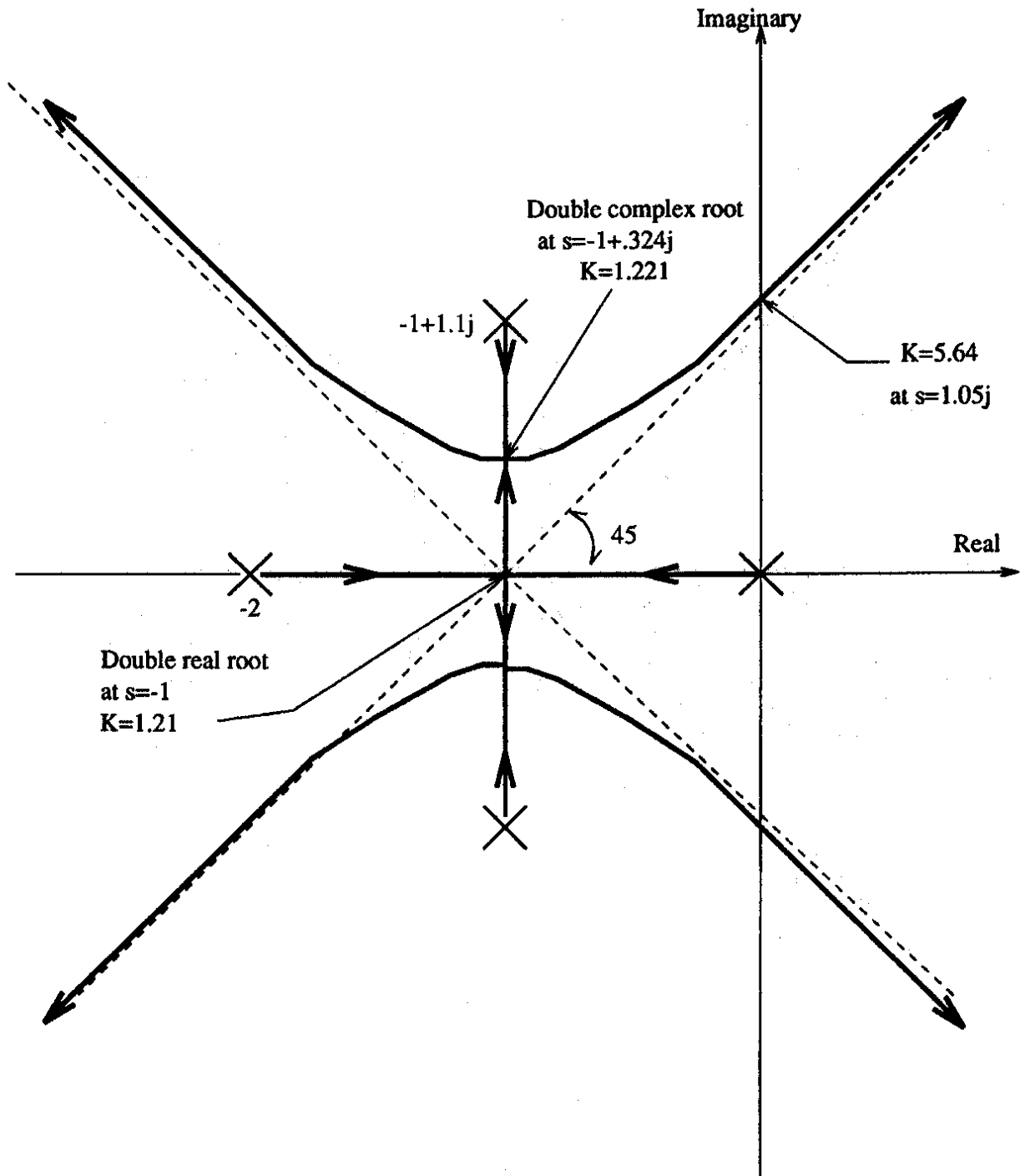
Figure 153.9 indicates a double root at  $s = -1$  when  $K = 0.81$ . When  $K = 0.819$ , there are double roots at  $s = -0.69178$  and  $-1.30882$ .

**Figure 153.7** Fourth order system with potential fourth order root.

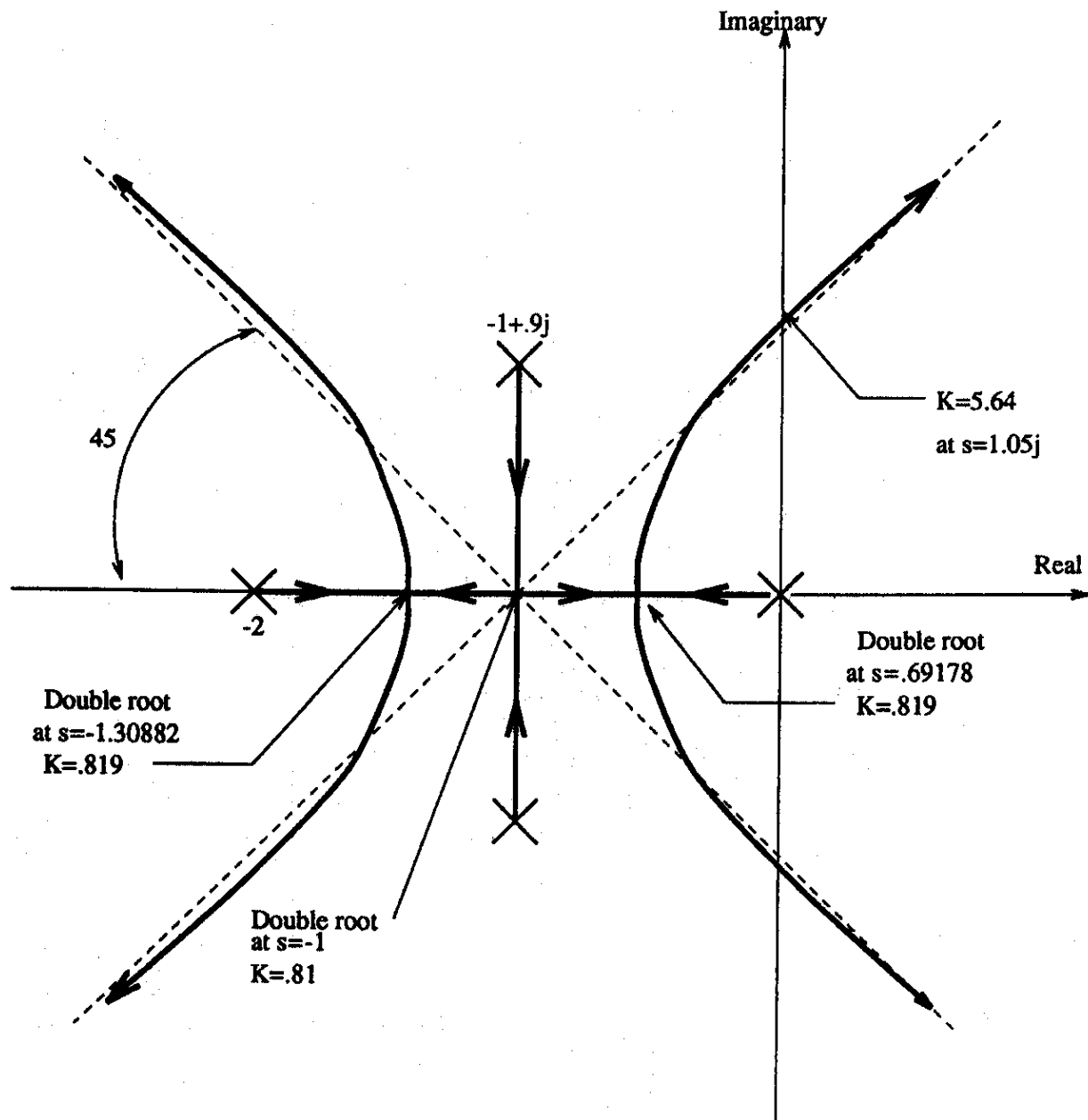




**Figure 153.8** Slightly modified system. Potential double complex roots.



**Figure 153.9** Modified system with three sets of potential double real roots.



## 153.5 Summary and Conclusions

Root locus design of feedback control systems reduces to (a) selecting closed-loop poles to give desired time response, (b) plotting the root locus, (c) finding the gain  $K$  if the desired poles are on the locus, and (d) reshaping the locus if it does not pass through the desired poles. This means adding suitable open-loop poles and zeros (compensation). Compensation is considered in

## Chapter 155.

### Defining Terms

**Angle condition:** Requirement that the phase angle be an odd multiple of  $180^\circ$ .

**Asymptotic angles:** Directions defining how loci approach infinity.

**Breakaway points:** Points where locus paths cross.

**Centroid:** Point from which the asymptotic angles are drawn.

**Magnitude condition:** The magnitude of  $K G(s) H(s)$  must be one.

**Poles:** Roots of the denominator of a transfer function.

### References

D'Azzo, J. J. and Houpis, C. H. 1988. *Linear Control System Analysis and Design*, 3rd ed. McGraw-Hill, New York.

Dorf, R. C. 1989. *Modern Control Systems*, 5th ed. Addison-Wesley, Reading, MA.

Evans, W. R. 1950. *Control System Synthesis by Root Locus Method*, AIEE Preprint 50–51.

### Further Information

*IEEE Transactions on Automatic Control*. Published monthly by the Institute of Electrical and Electronic Engineers.

*IEEE Transactions on Control Systems Technology*. Published quarterly by the Institute of Electrical and Electronic Engineers.

*IEEE Control Systems Magazine*. Published bimonthly by the Institute of Electrical and Electronic Engineers.

Nise, N. S. "Nyquist Criterion and Stability"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Nyquist Criterion and Stability

---

\* Adapted from Nise, N. S. 1995. *Control Systems Engineering*, 2nd ed. Copyright ©1995 by The Benjamin/Cummings Publishing Company, an imprint of Addison Wesley Longman, Inc. With permission. This material may not be reproduced in any format for any purpose without written permission of Addison Wesley Longman, Inc.

\* This material is used by permission of John Wiley & Sons, Inc.

### 154.1 Concept and Definition of Frequency Response

### 154.2 Plotting Frequency Response

### 154.3 Stability

### 154.4 Nyquist Criterion for Stability

### 154.5 Gain Design for Stability via the Nyquist Criterion

### 154.6 Stability via Separate Magnitude and Phase Plots (Bode Plots)

### Norman S. Nise

*California State Polytechnic University, Pomona*

Frequency response methods for the analysis and design of control systems were developed by H. Nyquist and H. W. Bode in the 1930s. These methods are older than, but not as intuitive as, the root locus, which was discovered by W. R. Evans in 1948. Frequency response yields a new vantage point from which to view feedback control systems. This technique possesses distinct advantages (1) when modeling transfer functions from physical data, (2) when designing lead compensators to meet a steady state error requirement and a transient response requirement, (3) when determining the stability of nonlinear systems, and (4) in settling ambiguities when sketching a root locus. This chapter introduces frequency response concepts and the determination of stability using the Nyquist criterion.

## 154.1 Concept and Definition of Frequency Response

---

In the steady state, sinusoidal inputs to a linear system generate sinusoidal responses of the same frequency. Even though these responses are of the same frequency as the input, they differ in amplitude and phase angle from the input. These differences are a function of frequency.

Sinusoids can be represented as complex numbers, or vectors, called **phasors**. The magnitude of the complex number is the amplitude of the sinusoid and the angle of the complex number is the phase angle of the sinusoid. Thus,  $M_1 \cos(\omega t + \phi_1)$  can be represented as  $M_1 \angle \phi_1$ , where the frequency,  $\omega$ , is implicit.

Since a system causes both the amplitude and phase angle of the input to be changed, we can

therefore think of the system itself represented by a complex number defined so that the product of the input phasor and the system function yields the phasor representation of the output.

Consider the system of Fig. 154.1 . Assume that the system is represented by the complex number,  $M(\omega)\angle\phi(\omega)$  . The output steady state sinusoid is found by multiplying the complex number representation of the input by the complex number representation of the system. Thus, the steady state output sinusoid is

$$M_o(\omega)\angle\phi_o(\omega) = M_i(\omega)M(\omega)\angle g(\phi_i(\omega) + \phi(\omega)) \quad (154.1)$$

From Eq. (154.1), we see that the system function is given by

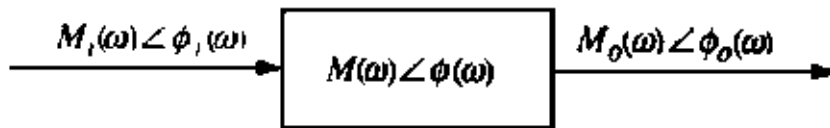
$$M(\omega) = \frac{M_o(\omega)}{M_i(\omega)} \quad (154.2)$$

and

$$\phi(\omega) = \phi_o(\omega) - \phi_i(\omega) \quad (154.3)$$

Equations (154.2) and (154.3) form the definition of frequency response. We call  $M(\omega)$  the **magnitude frequency response**, and  $\phi(\omega)$  the **phase frequency response**. The combination of the magnitude and phase frequency responses is called the **frequency response** and is  $M(\omega)\angle\phi(\omega) = G(j\omega)$  .

**Figure 154.1** Steady state sinusoidal frequency response function.



If we know the transfer function,  $G(s)$  , of a system, we can find  $G(j\omega)$  by using the relationship [Nilsson, 1990]

$$G(j\omega) = G(s)|_{s \rightarrow +j\omega} = M(\omega)\angle\phi(\omega) \quad (154.4)$$

## 154.2 Plotting Frequency Response

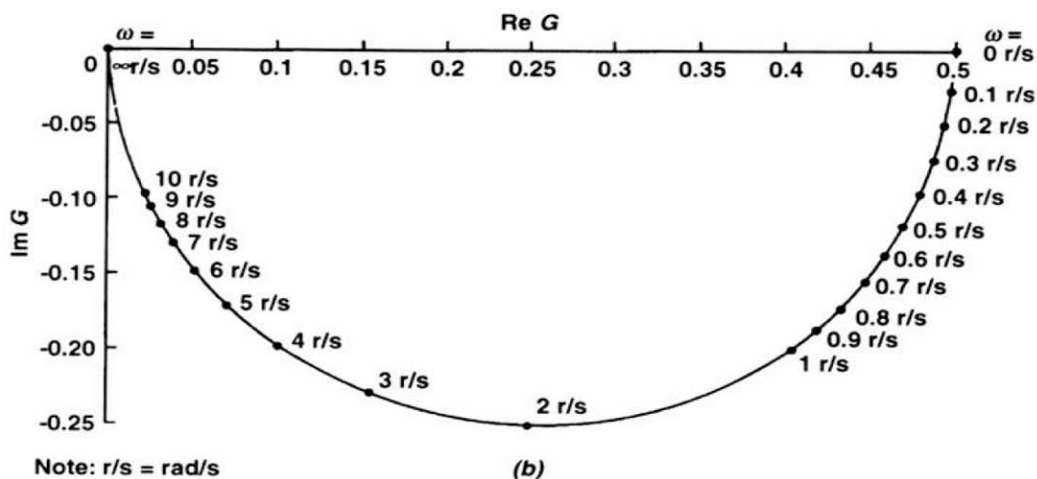
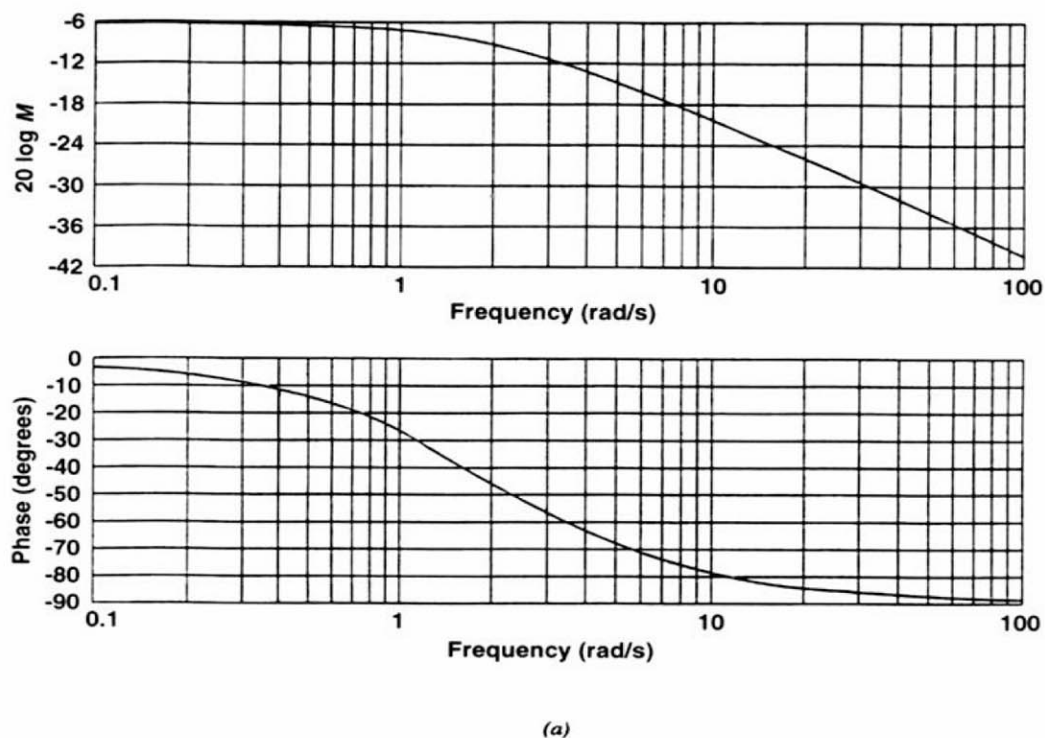
$G(j\omega) = M(\omega)\angle\phi(\omega)$  can be plotted in several ways. Two of these ways are (1) as a function of frequency with separate magnitude and phase plots, or (2) as a polar plot, where the phasor length is the magnitude and the phasor angle is the phase. When plotting separate magnitude and phase plots, the magnitude curve can be plotted in decibels (dB) vs.  $\log \omega$ , where  $\text{dB} = 20 \log M$  . The phase curve is plotted as phase angle vs.  $\log \omega$  . Plots that use dB and log frequency are called **Bode plots** . Bode plots can be easily drawn using asymptotic approximations [Nise, 1995].

As an example, find the analytical expression for the frequency response of the system,  $G(s) = 1/(s + 2)$ . Then, plot the separate magnitude and phase diagrams, as well as the polar plot.

First, substitute  $s = j\omega$  in the system function and obtain  $G(j\omega) = 1/(j\omega + 2) = (2 - j\omega)/(\omega^2 + 4)$ . The magnitude of this complex function,  $|G(j\omega)| = M(\omega) = 1/\sqrt{\omega^2 + 4}$ , is the magnitude frequency response. The phase angle of  $G(j\omega)$ ,  $\phi(\omega) = -\tan^{-1}(\omega/2)$ , is the phase frequency response.

$G(j\omega)$  can be plotted in two ways—Bode plots and a polar plot. The Bode plots are shown in Fig. 154.2(a), where the magnitude diagram is  $20 \log M(\omega) = 20 \log (1/\sqrt{\omega^2 + 4})$  vs.  $\log \omega$ , and the phase diagram is  $\phi(\omega) = -\tan^{-1}(\omega/2)$  vs.  $\log \omega$ . The polar plot, shown in Fig. 154.2(b), is a plot of  $M(\omega) \angle \phi(\omega) = 1/\sqrt{\omega^2 + 4} \angle -\tan^{-1}(\omega/2)$  for different  $\omega$ .

**Figure 154.2** Frequency response plots for  $G(s) = 1/(s + 2)$ : (a) Bode plots, (b) polar plot.



## 154.3 Stability

A linear, time-invariant system is **stable** if the natural response approaches zero as time approaches infinity. A linear, time-invariant system is unstable if the natural response grows without bound as time approaches infinity. Finally, a linear, time-invariant system is marginally stable if the natural response neither decays nor grows, but remains constant or oscillates as time approaches infinity. From the point of view of the transfer function, stable systems have closed-loop transfer functions with only left half-plane **poles**, where a pole is defined as a value of  $s$  that causes  $F(s)$  to be infinite, such as a root of the denominator of a transfer function. Unstable systems have closed-loop transfer functions with at least one right half-plane pole and/or poles of multiplicity greater than one on the imaginary axis. Marginally stable systems have closed-loop transfer functions with only imaginary axis poles of multiplicity one and left half-plane poles. Stability is the most important system specification. An unstable system cannot be designed for a specific transient response or steady state error requirement. Physically, instability can cause damage to a system, adjacent property, and human life. Many times, systems are designed with limit stops to prevent total runaway.

### CONTROL SYSTEMS AND THE NATIONAL DEFENSE RESEARCH COMMITTEE

*D. Mindell,*

*Massachusetts Institute of Technology*

With war approaching in 1940, Vannevar Bush, former MIT professor of electrical engineering and President of the Carnegie Institute of Washington, persuaded President Roosevelt to establish the National Defense Research Committee (NDRC) to bring the nation's scientific talent to bear on military problems. The NDRC and its successor and umbrella organization, the Office of Scientific Research and Development (OSRD), would become among the most successful wartime agencies, spending over \$500 million on scientific research during the war. They supervised work on the atomic bomb (which moved to the Army when it became the Manhattan Project), the development of microwave radar (at MIT's radiation lab), applications of penicillin, proximity fuses, operations research, and a host of other technologies that define our modern world.

In the field of control systems, the NDRC took prewar feedback and regulation techniques and transformed them into modern control engineering. The Fire Control Division of the NDRC, under the leadership of mathematician Warren Weaver and then under servo pioneer Harold Hazen, supervised a broad array of research into building automatic machines to shoot down attacking aircraft. A contract with the NDRC founded the MIT Servomechanisms Laboratory, which, under the leadership of Gordon Brown and Jay Forrester, would go on to build Whirlwind, the first digital computer for real-time control. Students of Brown, including Herbert Harriss and A. C. Hall, also began to combine the time-domain "transient analysis" developed for servomechanisms with the frequency-oriented Nyquist techniques developed for electronic amplifiers. The NDRC also funded Norbert Wiener, whose wartime work in antiaircraft control led to his famous 1948 book *Cybernetics*. Claude Shannon, founder of information theory, also worked on fire control.



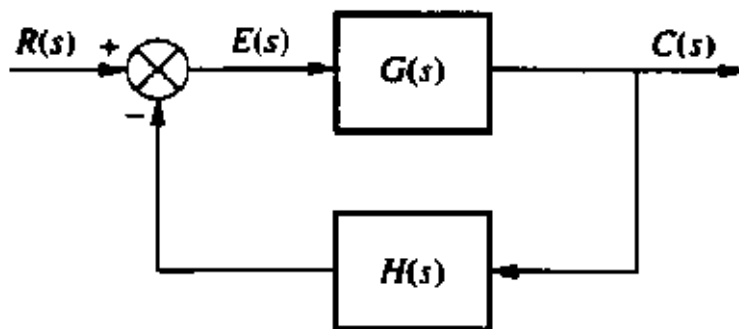
Along with H. Bode and R. C. Blackman at Bell Labs, Shannon studied noise in control systems as a problem in communications engineering, "since data smoothing is evidently a special case of the transmission, manipulation, and utilization of intelligence," thus sowing the seeds for information theory. Harold Hazen directed the NDRC to fund a broad program of "human factors" research, combining engineers and psychologists to understand the human dimension of control systems design. The Moore School of Electrical Engineering, under NDRC fire control contracts, did research in computing ballistics solutions for antiaircraft guns. When this group proposed their idea for an electronic numeric integrator and calculator, however, the NDRC turned it down as not being of immediate application to the war. Under a separate Army contract, this idea became ENIAC, the first electronic digital computer.

The NDRC project which achieved greatest success in the field prefigured the "cybernetic battlefield," so much a part of modern warfare. An electronic gun director developed by Bell Labs (the M-9) was integrated and combined with an automatic tracking radar built by the Radiation Lab (the SCR-584), electrohydraulic power controls built by Sperry, and the proximity fuse built by Johns Hopkins Applied Physics Laboratory. Together, this equipment formed the first automatic antiaircraft fire control system, precursor of Nike, ABM, and Patriot. It proved especially good at shooting down V-1 "buzz bombs" which attacked London in the summer of 1944. The straight and level flight path of the early cruise missiles made them particularly vulnerable to "straight-line" prediction methods employed by the Bell Labs computer. From July 18 to August 31, 1944, the system shot down 1286 V-1s, or 34% of the total attack. The servo engineering done at the Radiation Lab for the NDRC on the project, which coupled "conical scanning" radar to a servo-controlled antenna, led to the 1947 book by James, Nichols, and Phillips, *Theory of Servomechanisms*, which became the standard textbook in control theory after the war.

## 154.4 Nyquist Criterion for Stability

The **Nyquist criterion** relates the stability of a closed-loop system to the open-loop frequency response and open-loop pole location. Consider the system of [Fig. 154.3](#). The Nyquist criterion can tell us how many closed-loop poles are in the right half-plane.

**Figure 154.3** Closed-loop control system.

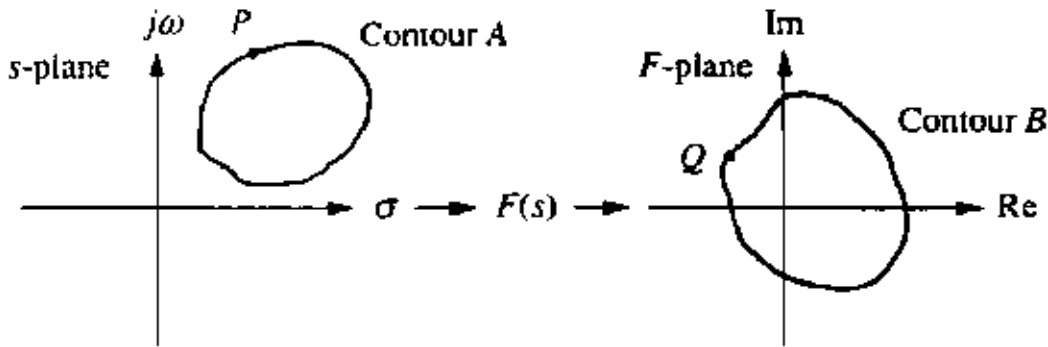


Before stating the Nyquist criterion, let us discuss the concept of *mapping*. If we take a complex number on the  $s$  plane and substitute it into a function,  $F(s)$ , another complex number will result. This process is called mapping. For example, substituting  $s = 4 + j3$  into the function  $(s^2 + 2s + 1)$  yields  $16 + j30$ . We say that  $4 + j3$  maps into  $16 + j30$  through the function  $(s^2 + 2s + 1)$ .

Now consider a collection of points, called a *contour*, shown in Fig. 154.4 as contour A. Also, assume

$$F(s) = \frac{(s - z_1)(s - z_2) \dots}{(s - p_1)(s - p_2) \dots} \quad (154.5)$$

**Figure 154.4** Mapping contour A through function  $F(s)$  to contour B.



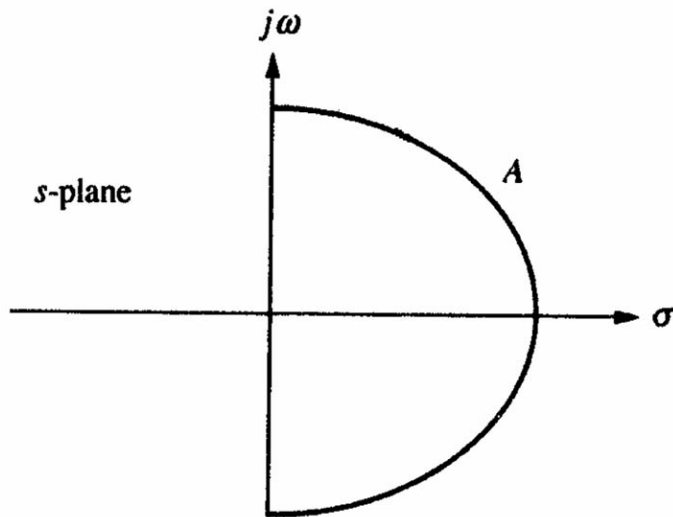
Contour A can be mapped through  $F(s)$  into contour B by substituting each point of contour A into the function  $F(s)$  and plotting the resulting complex numbers. For example, point P in Fig. 154.4 maps into point Q through the function  $F(s)$ . Let us now put these concepts together and state the Nyquist criterion for stability [see (Nise,1995) for derivation]:

Assume the system of Fig. 154.3, where the open-loop transfer function is  $G(s)H(s)$  and the closed-loop transfer function is  $T(s) = G(s)/[1 + G(s)H(s)]$ . If a contour A that encircles the entire right half-plane, as shown in Fig. 154.5, is mapped through  $G(s)H(s)$  into contour B, then  $Z = P - N$ . Z is the number of closed-loop poles [the poles of  $T(s)$ ] in the right half-plane, P is the number of open-loop poles [the poles of  $G(s)H(s)$ ] in the right half-plane, and N is the number of counterclockwise encirclements of  $-1$  of contour B. The mapping, contour B, is called the **Nyquist diagram**.

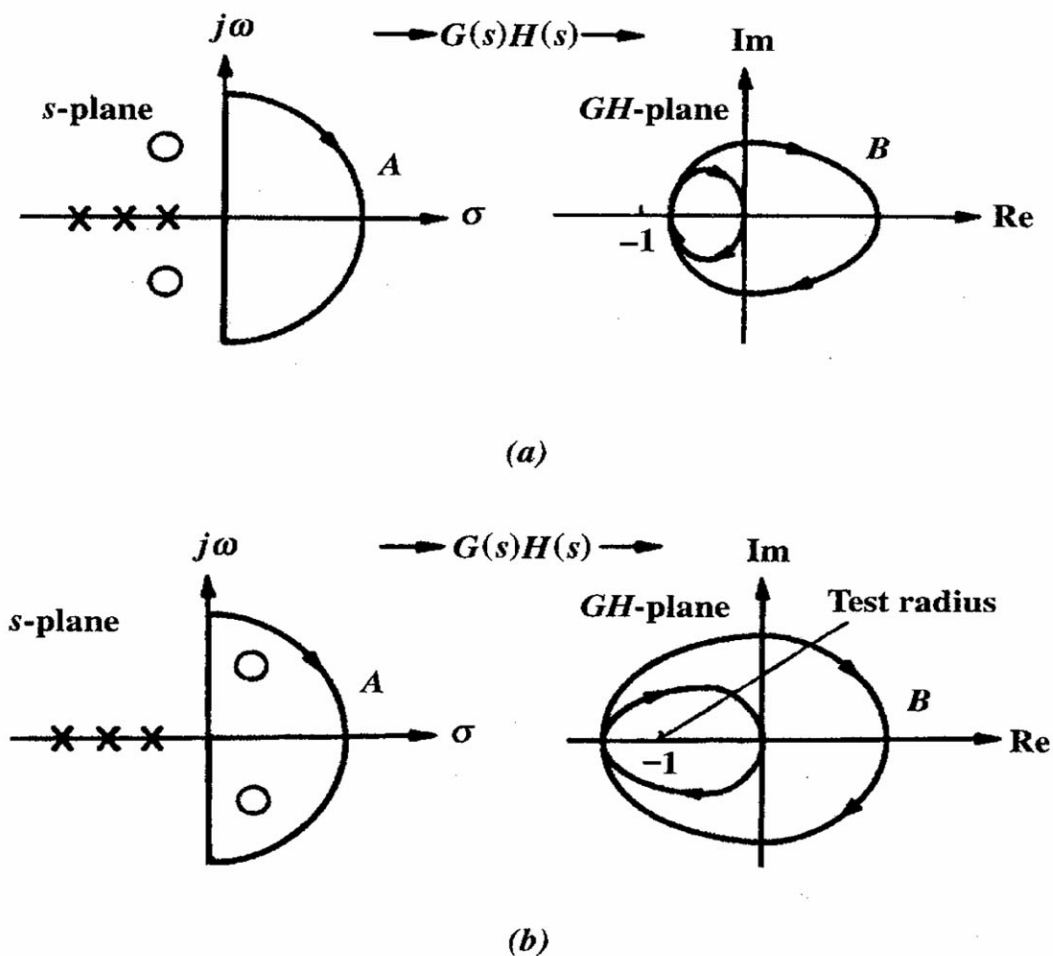
The Nyquist criterion is classified as a frequency response technique because, around contour A in Fig. 154.5, the mapping of the points on the  $j\omega$  axis through the function  $G(s)H(s)$  is the same as substituting  $s = j\omega$  into  $G(s)H(s)$  forming the frequency response function,  $G(j\omega)H(j\omega)$ . Thus, part of the Nyquist diagram is the polar plot of the frequency response of  $G(s)H(s)$ . Let us look at two examples that illustrate the application of the Nyquist criterion.

In Fig. 154.6(a), contour A maps through  $G(s)H(s)$  into a Nyquist diagram that does not encircle  $-1$ . Hence,  $P = 0$ ,  $N = 0$ , and  $Z = P - N = 0$ . Because  $Z = 0$  is the number of closed-loop poles inside contour A, which encircles the right half-plane, this system does not have any right half-plane poles and is stable.

**Figure 154.5** Contour enclosing right half-plane to determine stability.



**Figure 154.6** Examples of mapping for the Nyquist criterion: (a) contour A does not enclose closed-loop poles; (b) contour A encloses closed-loop poles.



○ = zeros of  $1 + G(s)H(s)$   
 = poles of closed-loop system  
 Location not known

× = poles of  $1 + G(s)H(s)$   
 = poles of  $G(s)H(s)$   
 Location is known

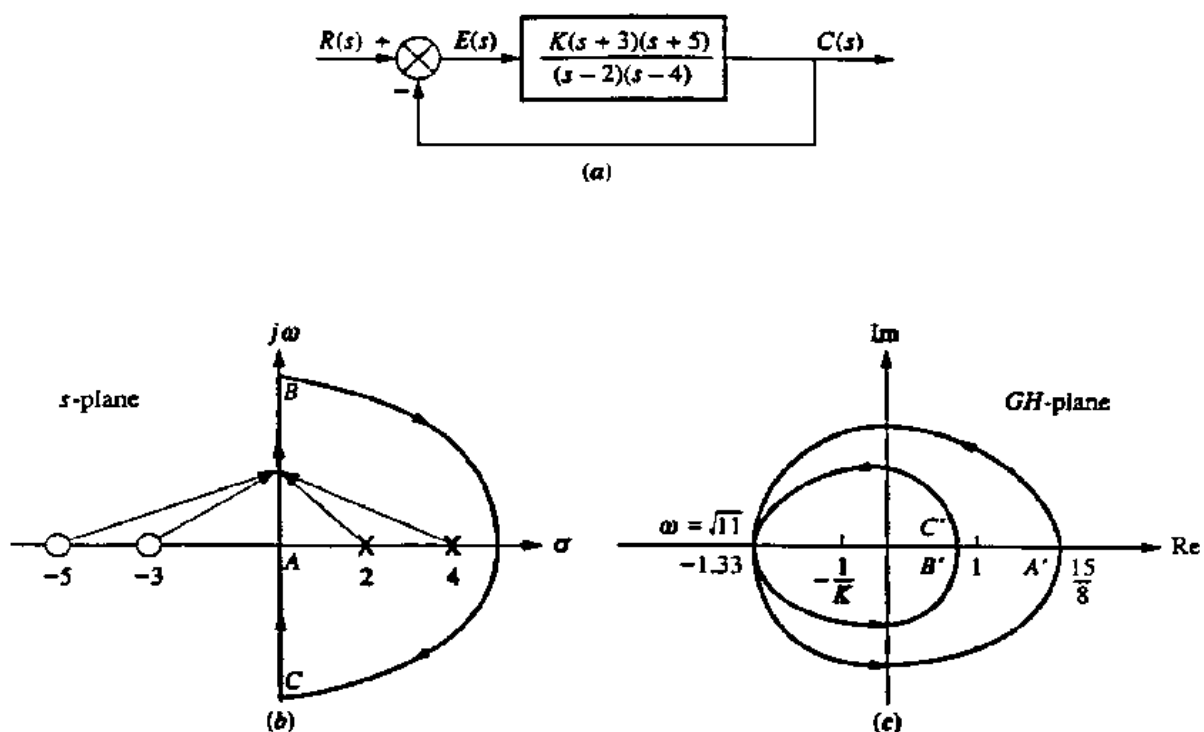
On the other hand, Fig. 154.6(b) shows a contour  $A$  that generates two clockwise encirclements of  $-1$  when mapped through the function  $G(s)H(s)$ . Thus,  $P = 0$ ,  $N = -2$ , and  $Z = P - N = 2$ . The system is unstable because it has two closed-loop poles in the right half-plane ( $Z = 2$ ). The two closed-loop poles are shown inside contour  $A$  in Fig. 154.6(b) as zeros of  $1 + G(s)H(s)$ . The reader should keep in mind that the existence of these poles is not known a priori.

In this example, notice that clockwise encirclements imply a negative value for  $N$ . The number of encirclements can be determined by drawing a test radius from  $-1$  in any convenient direction and counting the number of times the Nyquist diagram crosses the test radius. Counterclockwise crossings are positive and clockwise crossings are negative. For example, in Fig. 154.6(b), contour  $B$  crosses the test radius twice in a clockwise direction. Hence, there are  $-2$  encirclements of the point  $-1$ .

## 154.5 Gain Design for Stability via the Nyquist Criterion

We now use the Nyquist criterion to design a system's gain for stability. The general approach is to set the loop gain equal to unity and draw the Nyquist diagram. Since gain is simply a multiplying factor, the effect of the gain is to multiply the resultant by a constant anywhere along the Nyquist diagram. For example, consider Fig. 154.7, which summarizes the Nyquist approach for a system with variable gain  $K$ . As the gain is varied, we can visualize the Nyquist diagram [Fig. 154.7(c)] expanding (increased gain) or shrinking (decreased gain) like a balloon. This motion could move the Nyquist diagram past the  $-1$  point and change the stability picture. For this system, since  $P = 2$ , the critical point must be encircled two times in the counterclockwise direction by the Nyquist diagram to yield  $N = 2$  and a stable system. A reduction in gain would place the critical point outside the Nyquist diagram where  $N = 0$  yielding  $Z = 2$ , an unstable system.

**Figure 154.7** Feedback control system to demonstrate Nyquist stability: (a) system, (b) contour, (c) Nyquist diagram.



From another perspective, we can think of the Nyquist diagram as remaining stationary and the  $-1$  point moving along the real axis. In order to do this, we set the gain to unity and position the critical point at  $-1/K$  rather than  $-1$ . Thus, the critical point appears to move closer to the origin as  $K$  increases.

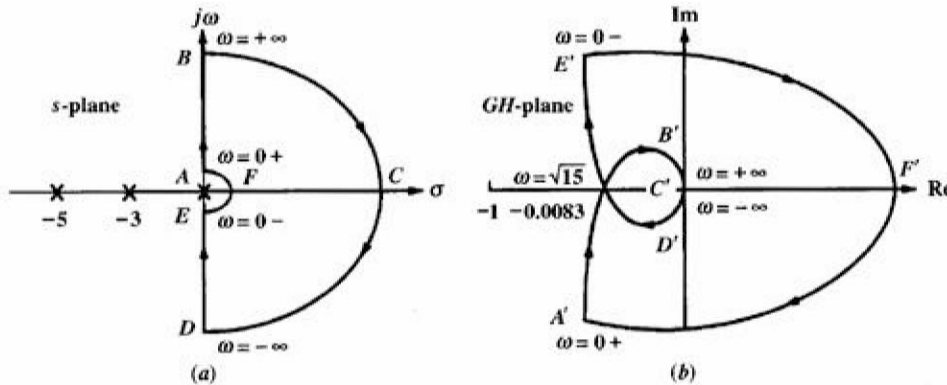
Finally, if the Nyquist diagram intersects the real axis at  $-1$ , then  $G(j\omega)H(j\omega) = -1$ . From the root locus, when  $G(s)H(s) = -1$ , the variable  $s$  is a closed-loop pole of the system. Thus, the frequency at which the Nyquist diagram intersects  $-1$  is the same frequency at which the root locus crosses the  $j\omega$  axis. Hence, the system is marginally stable if the Nyquist diagram intersects the real axis at  $-1$ .

In summary, if the open-loop system contains a variable gain  $K$ , set  $K = 1$  and sketch the Nyquist diagram. Consider the critical point to be at  $-1/K$  rather than at  $-1$ . Adjust the value of gain  $K$  to yield stability based upon the Nyquist criterion.

Let us look at an example. For a unity feedback system, where  $G(s) = K/[s(s+3)(s+5)]$ , find the range of gain  $K$  for stability, instability, and marginal stability. For marginal stability, also find the frequency of oscillation.

First, set  $K = 1$  and sketch the Nyquist diagram for the system using the contour shown in Fig. 154.8(a). Conceptually, the Nyquist diagram is plotted by substituting the points of the contour shown in Fig. 154.8(a) into  $G(s) = 1/[s(s+3)(s+5)]$ . The contour, as shown, must detour around open-loop imaginary axis poles in order to plot a continuous Nyquist diagram. The detour, however, is epsilon close to the open-loop poles to ensure that any closed-loop right half-plane poles close to the imaginary open-loop poles are still inside the contour and are counted.

**Figure 154.8** (a) Contour for example; (b) Nyquist diagram.



From A to B, we use

$$G(j\omega) = \frac{1}{s(s+3)(s+5)} g|_{s \rightarrow j\omega} = \frac{-8\omega^2 - j\omega(15 - \omega^2)}{64\omega^4 + \omega^2(15 - \omega^2)^2} \quad (154.6)$$

and let  $\omega$  vary from  $0+$  to  $\infty$ . The mapping in Fig. 154.8(b) goes from A' at  $\infty$  to B' at the origin.

Around the infinite circle from B through C to D, the mapping is found by replacing each complex factor in  $G(s)$  by its polar form. Thus,

$$G(s) = \frac{1}{(R_0 \angle \theta_0)(R_3 \angle \theta_3)(R_5 \angle \theta_5)} \quad (154.7)$$

The angles are the angles drawn from the respective poles to a point on the infinite circle. The  $R_i$ s are the vector lengths (in this case, infinite). Hence, all points on the infinite circle map into the origin.

The negative imaginary axis from  $D$  to  $E$  maps into a mirror image of the mapping from  $A$  to  $B$ , because  $G(j\omega)$  has an even real part and an odd imaginary part.

From  $E$  through  $F$  to  $A$ , we can again use Eq. (154.7).  $R_0$  is zero. Thus, the resultant magnitude is infinite. At  $E$ , the angles add to  $-90^\circ$ . Hence, the resultant is  $+90^\circ$ . Similar reasoning yields the mapping of  $F$  and  $A$  to  $F'$  and  $A'$ , respectively.

Finally, let us find the point where the Nyquist diagram intersects the negative real axis. We set the imaginary part of Eq. (154.6) equal to zero using  $\omega = \sqrt{15}$ . Then we substitute this value of  $\omega$  back into Eq. (154.6) and find that the real part equals  $-0.0083$ .

From the contour of Fig. 154.8(a),  $P = 0$  and, for stability,  $N$  then must be equal to zero. From Fig. 154.8(b), the system is stable if the critical point lies outside the contour ( $N = 0$ ) so that  $Z = P - N = 0$ . Thus,  $K$  can be increased by  $1/0.0083 = 120.48$  before the Nyquist diagram encircles  $-1$ . Hence, for stability,  $K < 120.48$ . For marginal stability,  $K = 120.48$ . At this gain, the Nyquist diagram intersects  $-1$ , and the frequency of oscillation is  $\sqrt{15}$  rad/s.

Now that we have used the Nyquist diagram to determine stability, we can develop a simplified approach that uses only the mapping of the positive  $j\omega$  axis.

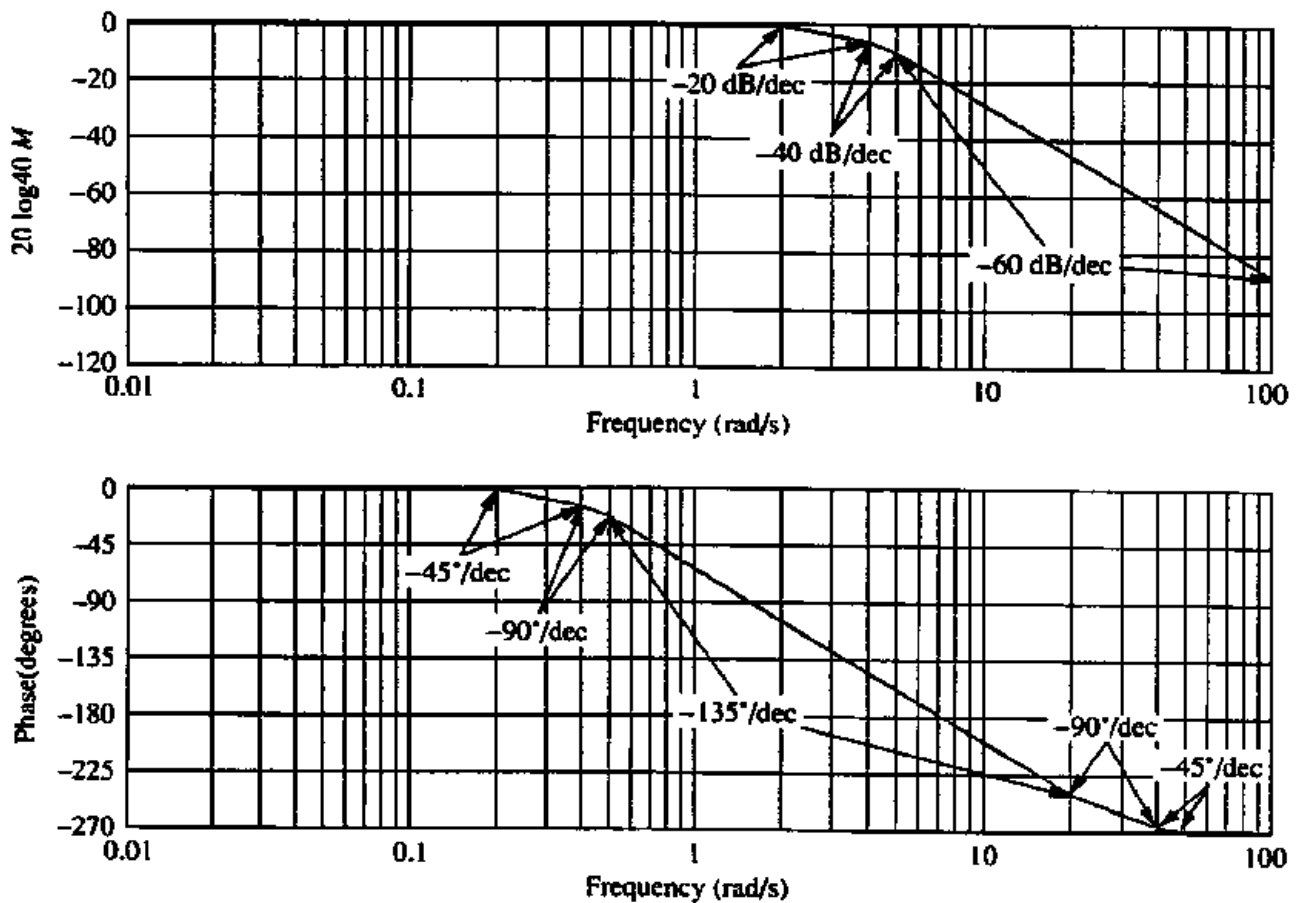
## 154.6 Stability via Separate Magnitude and Phase Plots(Bode Plots)

In many cases, stability can be determined from just a mapping of the positive  $j\omega$  axis, or, in other words, stability can be determined from the frequency response of the open-loop transfer function. Consider a stable open-loop system, such as that shown in Fig. 154.6(a). Since the open-loop system does not have right half-plane poles,  $P = 0$ . For the closed-loop system to be stable,  $N = 0$  yields  $Z = P - N = 0$ .  $N = 0$  (i.e., no encirclements of  $-1$ ) implies that the Nyquist diagram passes to the right of  $-1$ . Stated another way, the gain of the open-loop transfer function is less than unity at  $180^\circ$ . Thus, if we use Bode plots, we can determine stability from the open-loop frequency response by looking at the magnitude response curve where the phase is  $180^\circ$ . If the gain is less than unity at this frequency, the closed-loop system is stable.

Let us look at an example and determine the range of gain for stability of a system by implementing the Nyquist stability criterion using asymptotic log-magnitude and phase plots. From the log-magnitude plot, we will determine the value of gain that ensures that the magnitude is less than 0 dB (unity gain) at the frequency where the phase is  $\pm 180^\circ$ .

Shown in Fig. 154.9 are the asymptotic approximations to the frequency response of  $G(s) = K/[(s+2)(s+4)(s+5)]$ , scaled to  $K = 40$ . Since this system has all of its open-loop poles in the left half-plane, the open-loop system is stable. Hence, the closed-loop system will be stable if the open-loop frequency response has a gain less than unity when the phase is  $180^\circ$ . Accordingly, we see that at a frequency of approximately 7 rad/s, when the phase plot is  $180^\circ$ , the magnitude plot is  $-20$  dB. Therefore, an increase in gain of  $+20$  dB is possible before the system becomes unstable. Since the gain plot was scaled for a gain of 40,  $+20$  dB (a gain of 10) represents the required increase in gain above 40. Hence, the gain for instability is  $40 \times 10 = 400$ . The final result is  $0 < K < 400$  for stability. This result, obtained by approximating the frequency response by Bode asymptotes, can be compared to the result obtained from the actual frequency response, which yields a gain of 378 at a frequency of 6.16 rad/s.

**Figure 154.9** Log-magnitude and phase plots.



In this chapter, we discussed the concept of frequency response, including the Nyquist criterion and its application to the stability of linear feedback control systems. The concept of frequency response can be further applied to transient response analysis and design. Frequency response techniques can also be applied to digital and nonlinear control systems.

## Defining Terms

**Code plot:** A sinusoidal frequency response plot where the magnitude response is plotted separately from the phase response. The magnitude plot is decibels vs. log frequency, and the phase plot is phase vs. log frequency. In control systems, the Bode plot is usually made for the open-loop transfer function.

**Frequency response:** The combination of magnitude and phase frequency responses expressed as separate magnitude and phase responses, or as a complex function of frequency.

**Magnitude frequency response:** The ratio of the magnitude of the steady state sinusoidal response to the magnitude of the sinusoidal input as a function of frequency. The ratio can be expressed in dB.

**Nyquist criterion for stability:** Assume a feedback system where the open-loop transfer function is  $G(s)H(s)$  and the closed-loop transfer function is  $T(s) = G(s)/[1 + G(s)H(s)]$ . If a contour  $A$  that encircles the entire right half-plane is mapped through  $G(s)H(s)$  into contour



$B$ , then  $Z = P - N$ .  $Z$  is the number of closed-loop poles [the poles of  $T(s)$ ] in the right half-plane,  $P$  is the number of open-loop poles [the poles of  $G(s)H(s)$ ] in the right half-plane, and  $N$  is the number of counterclockwise encirclements of  $-1$  of contour  $B$ . The mapping, contour  $B$ , is called the Nyquist diagram.

**Nyquist diagram:** A mapping of a closed-contour on the  $s$  plane that encloses the right half-plane.

**Phase frequency response:** The difference between the phase angle of the steady state sinusoidal response and the phase angle of the input sinusoid as a function of frequency.

**Phasor:** A complex number, or vector, representing a sinusoid.

**Pole:** A value of  $s$  that causes  $F(s)$  to be infinite, such as a root of the denominator of a transfer function.

**Stability of a linear, time-invariant system:** That characteristic of a linear, time-invariant system defined by a natural response that decays to zero as time approaches infinity.

## References

- Nilsson, J. W. 1990. *Electric Circuits*, 3rd ed. Addison Wesley, Reading, MA.  
Nise, N. S. 1995. *Control Systems Engineering*, 2nd ed. Benjamin/Cummings, Redwood City, CA.

## Further Information

- Bode, H. W. 1945. *Network Analysis and Feedback Amplifier Design*. Van Nostrand, Princeton, NJ.  
Dorf, R. C. and Bishop, R. H. 1995. *Modern Control Systems*, 7th ed. Addison Wesley, Reading, MA.  
Kuo, B. C. 1995. *Automatic Control Systems*, 7th ed. Prentice-Hall, Englewood Cliffs, NJ.  
Kuo, F. F. 1966. *Network Analysis and Synthesis*. John Wiley & Sons, New York.  
Nyquist, H. 1932. Regeneration Theory. *Bell Syst. Tech. J.*, January, pp. 126–147.



Raven, F. H. "System Compensation"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

155.1 Correlation between Transient and Frequency Response

155.2 Determining  $K$  to Yield a Desired  $M_p$

155.3 Gain Margin and Phase Margin

155.4 Series Compensation

Lead Compensation • Lag Compensation • Lag-lead Compensation

155.5 Internal Feedback

155.6 Compensation on the  $S$  Plane

**Francis H. Raven**

*University of Notre Dame*

The frequency response of a control system may be determined experimentally. Since this is the frequency response for the actual system, correlation criteria which relate transient response to frequency response may then be used to ascertain the transient behavior of the system. Various system **compensation** techniques are available for changing the frequency response so as to improve the transient behavior of the system.

## 155.1 Correlation between Transient and Frequency Response

Figure 155.1 shows a second order, type 1 system. For this system, the **natural frequency** is  $\omega_n = \sqrt{K_1/\tau}$  and the **damping ratio** is  $\zeta = 1/(2\sqrt{K_1\tau})$ . The transient behavior is completely described by  $\zeta$  and  $\omega_n$  [Raven, 1995]. For a sinusoidal input  $r(t) = r_o \sin \omega t$ , after the initial transients have died out, the response will be  $c(t) = c_o \sin(\omega t + \phi)$ . The ratio of the amplitude of the output sinusoidal  $c_o$  to the input sinusoidal  $r_o$  is  $M = c_o/r_o$ . The value of  $\omega$  at which  $M$  is a maximum is

$$\omega_p = \omega_n \sqrt{1 - \zeta^2} \quad 0 \leq \zeta \leq 0.707 \quad (155.1)$$

where  $\omega_p$  is the value of  $\omega$  at which  $M$  attains its peak or maximum value. The corresponding peak or maximum value of  $M$  which is designated  $M_p$  is

$$M_p = \frac{1}{2\zeta\sqrt{1 - \zeta^2}} \quad 0 \leq \zeta \leq 0.707 \quad (155.2)$$

The preceding result has significance only for  $0 \leq \zeta \leq 0.707$  , in which case  $M_p \geq 1$  . Solving Eq. (155.2) for the damping ratio  $\zeta$  gives

$$\zeta = [(1 - \sqrt{1 - 1/M_p^2}) / 2]^{1/2} \quad M_p \geq 1 \quad (155.3)$$

The transient behavior of higher order, type 1 systems is closely approximated by the preceding correlation criteria. For the case in which  $M_p < 1$  , the transient response is described by a first order, type 1 system whose time constant  $\tau_c$  is

$$\tau_c = 1/\omega_c \quad M_p < 1 \quad (155.4)$$

where  $\omega_c$  is the value of  $\omega$  at which the  $G(j\omega)H(j\omega)$  plot crosses the unit circle. That is,  $|G(j\omega)H(j\omega)| = 1$  .

**Figure 155.1** Second order, type 1 system.

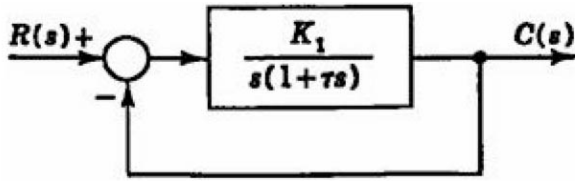


Figure 155.2 shows a second order, type 0 system. For this system, the natural frequency is  $\omega_n = \sqrt{(1 + K_o)/b}$  , and the damping ratio is  $\zeta = a/(2\sqrt{(1 + K_o)b})$  . The value of  $\omega$  at which  $M$  is a maximum is the same as that given by Eq. (155.1). The corresponding peak or maximum value of  $M$  is

$$\frac{M_p}{K_o/(1 + K_o)} = \frac{1}{2\zeta\sqrt{1 - \zeta^2}} \quad 0 \leq \zeta \leq 0.707 \quad (155.5)$$

For large values of  $K_o$  , then  $K_o/(1 + K_o) \approx 1$  , and this criterion becomes the same as that for a type 1 system. Solving Eq. (155.5) for the damping ratio  $\zeta$  gives

$$\zeta = \{ [1 - \sqrt{1 - (K/M_p)^2}] / 2 \}^{1/2} \quad M_p/K \geq 1 \quad (155.6)$$

where

$$K = K_o/(1 + K_o)$$

For the case in which  $M_p/K = M_p/[K_o/(1 + K_o)] < 1$  , the transient response is described by a first order, type 0 system whose time constant  $\tau_c$  is

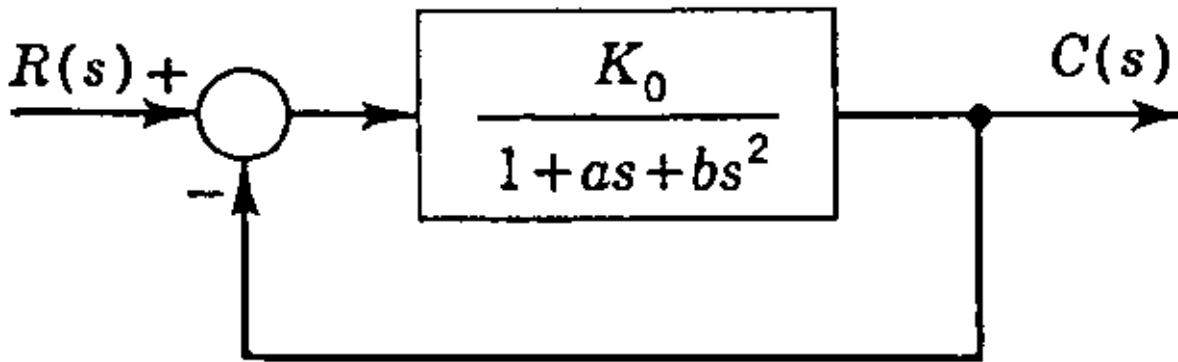
$$\tau_c = 1/\omega_c \quad (155.7)$$

where  $\omega_c$  is the value of  $\omega$  at which the  $G(j\omega)H(j\omega)$  plot crosses the circle whose radius is

$$r = \frac{K_o}{\sqrt{1 + (1 + K_o)^2}} \quad (155.8)$$

For large values of  $K_o$ , the radius  $r$  approaches the unit circle and the preceding criterion becomes the same as that for a type 1 system.

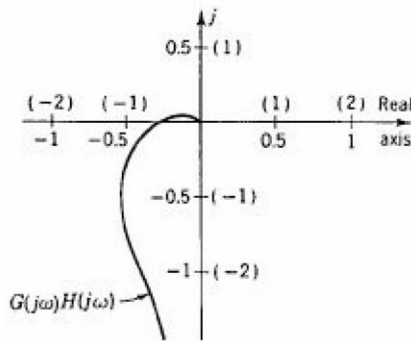
**Figure 155.2** Second order, type 0 system.



## 155.2 Determining $K$ to Yield a Desired $M_p$

Figure 155.3 shows a typical polar plot of  $G(j\omega)H(j\omega)$ . If the gain  $K$  of the system is doubled, the value  $G(j\omega)H(j\omega)$  is doubled at every point. Multiplying the old scale by a factor of 2 yields the polar plot for the system in which the gain  $K$  has been doubled. Values of this new scale are shown in parentheses. Changing the gain  $K$  does not affect the shape of the polar plot.

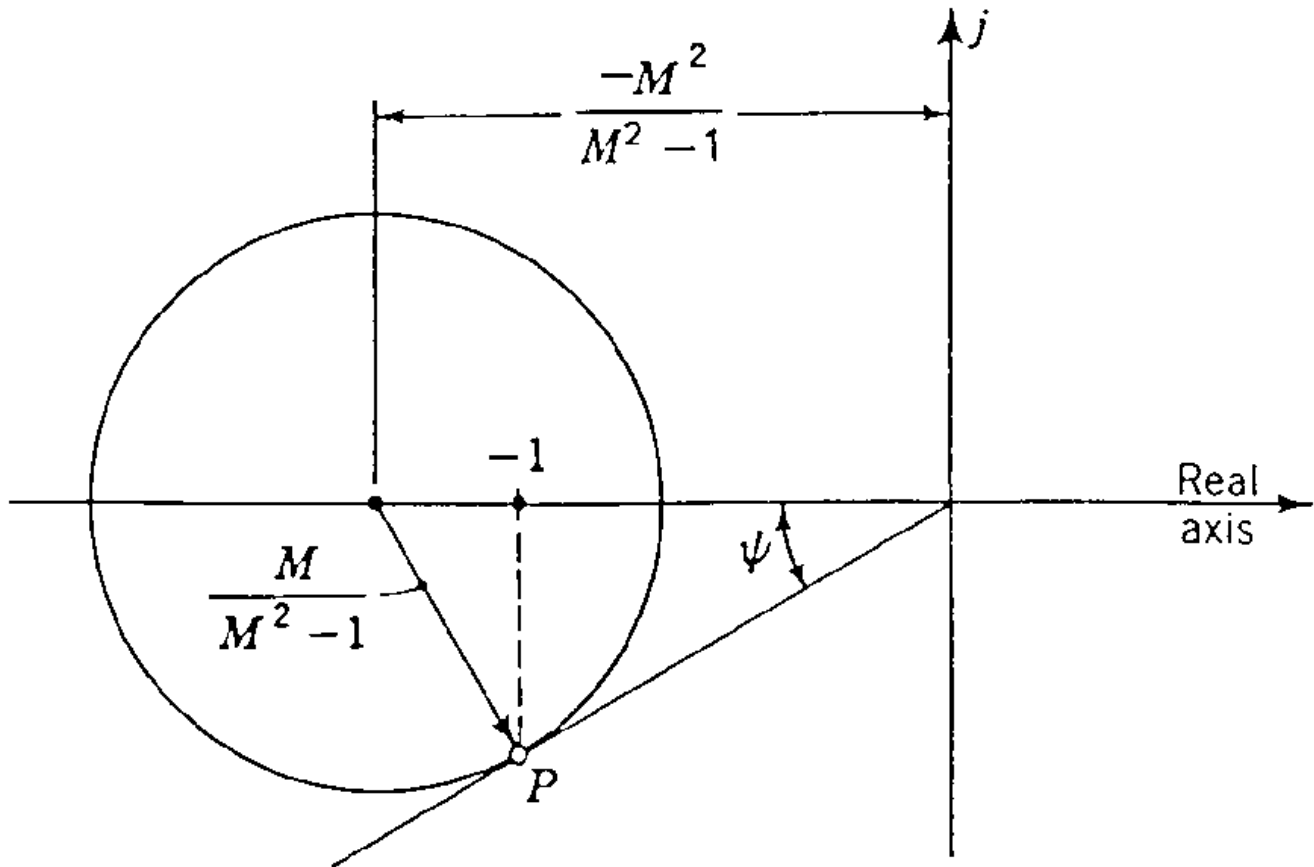
**Figure 155.3** Typical polar plot.



An  $M$  circle is shown in Fig. 155.4. The center of the circle is located on the real axis at  $-M^2/(M^2 - 1)$ , and the radius is  $M/(M^2 - 1)$ . The line drawn from the origin, tangent to the  $M$  circle at the point  $P$ , has an included angle of  $\psi$ . The value of  $\sin \psi$  is

$$\sin \psi = 1/M \quad (155.9)$$

**Figure 155.4** Tangent to an  $M$  circle.



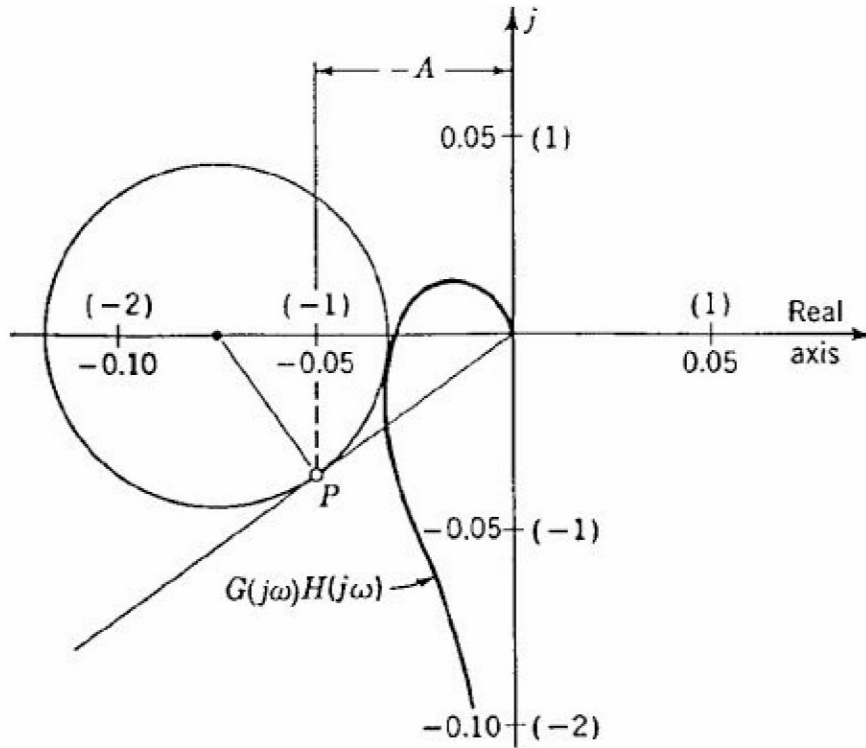
A characteristic feature of the point of tangency  $P$  is that a line drawn from the point  $P$  perpendicular to the negative axis intersects this axis at the  $-1$  point.

The procedure for determining the gain  $K$  so that the  $G(j\omega)H(j\omega)$  plot will have a desired value of  $M_p$  is as follows.

1. Draw the polar plot for  $G(j\omega)H(j\omega)/K$ .
2. Draw the tangent line to the desired  $M_p$  circle [ $\psi = \sin^{-1}(1/M_p)$ ].
3. Draw the circle with center on the negative real axis that is tangent to both the  $G(j\omega)H(j\omega)/K$  plot and the tangent line, as is shown in Fig. 155.5.
4. Erect the perpendicular to the negative real axis from point  $P$ . This perpendicular intersects the negative real axis at the point  $-A = -0.05$ .

5. In order that the circle drawn in step 3 is the desired  $M_p$  circle, this point should be  $-1$  rather than  $-A$ . The required gain is that value of  $K$  which changes the scale so that this does become the  $-1$  point. Thus  $K(-A) = -1$  or  $K = 1/A$ .

**Figure 155.5** Determination of  $K$  to yield a desired  $M_p$ .



As is illustrated in Fig. 155.5, the perpendicular drawn from point  $P$  to the negative real axis intersects the axis at the value  $-A = -0.05$ . Multiplication of the scale by a factor of 20, as is shown in Fig. 155.5 by the numbers in parentheses, converts this point to the  $-1$  point. Thus, the required value of the gain  $K$  such that the polar plot  $G(j\omega)H(j\omega)$  is tangent to the desired  $M_p$  circle is 20.

The plot  $G(j\omega)H(j\omega)/K$  is the plot for the case in which  $K$  is one. By constructing the plot for  $G(j\omega)H(j\omega)$  rather than  $G(j\omega)H(j\omega)/K$ , the resulting value of  $1/A$  represents the required factor  $K_c$  by which the gain should be changed such that the resulting polar plot will be tangent to the desired  $M_p$  circle. That is,

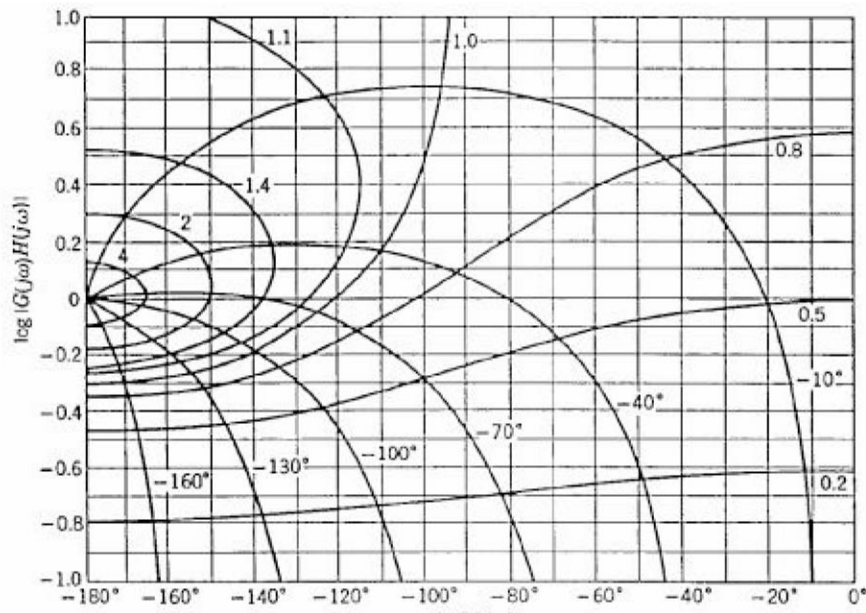
$$\text{New gain} = K_c (\text{original gain}) \quad (155.10)$$

Another method for representing frequency response information is the log-modulus or Nichols plot [Nichols, 1947]. A log-modulus plot is a plot of  $\log |G(j\omega)H(j\omega)|$  versus  $\angle G(j\omega)H(j\omega)$ . Lines of constant  $M$  and constant  $\alpha$ , where

$$M = |C(j\omega)/R(j\omega)| \quad \text{and} \quad \alpha = \angle C(j\omega)/R(j\omega) \quad (155.11)$$

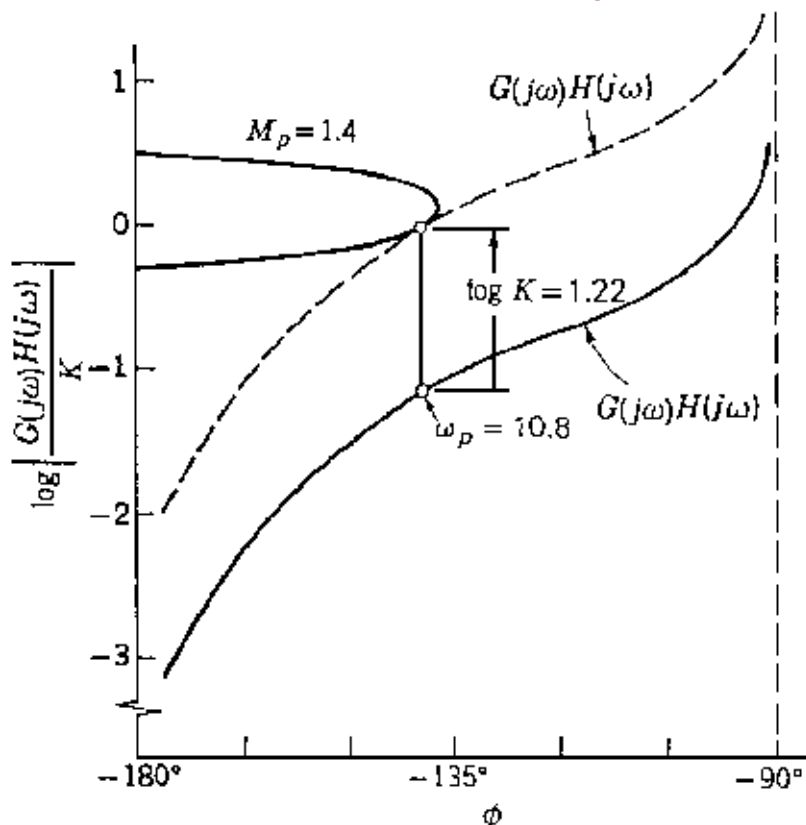
are circles on the polar plot, become contours when drawn on the log-modulus plot. These  $M$  and  $\alpha$  contours are shown in Fig. 155.6.

**Figure 155.6** Log-modulus representation for lines of constant  $M$  and lines of constant  $\alpha$ .



Changing the gain  $K$  does not affect the phase angle, but merely moves the log-modulus plot  $G(j\omega)H(j\omega)/K$  for a system vertically up for  $K > 1$  and vertically down for  $K < 1$ . The  $M_p$  contour shown in Fig. 155.7 is the desired value of  $M_p$ . The solid curve is the log-modulus plot of  $G(j\omega)H(j\omega)/K$  for the system. The vertical distance that this curve must be raised such that it is tangent to the desired  $M_p$  contour is  $\log K$ . The dashed curve is the resultant frequency response  $G(j\omega)H(j\omega)$ . The frequency at the point of tangency is  $\omega_p$ .

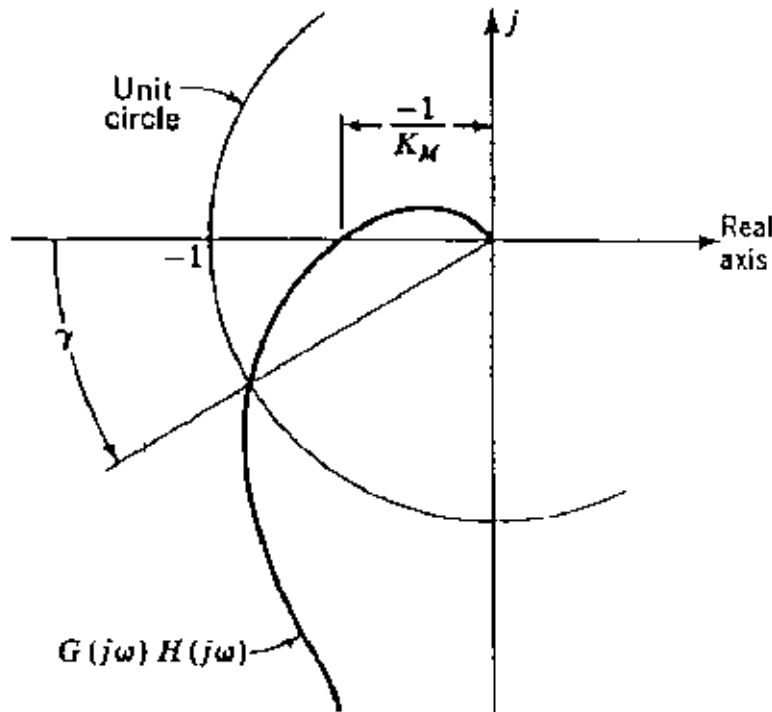
**Figure 155.7** Determination of  $K$  for a desired  $M_p$ .



## 155.3 Gain Margin and Phase Margin

The  $-1$  point of the  $G(s)H(s)$  map has great significance with regard to the stability of a system. Figure 155.8 shows a typical  $G(j\omega)H(j\omega)$  plot in the vicinity of the  $-1$  point. If the gain were multiplied by an amount  $K_M$ , called the **gain margin**, the  $G(j\omega)H(j\omega)$  plot would go through the  $-1$  point. Thus, the gain margin is an indication of how much the gain can be increased before the  $G(j\omega)H(j\omega)$  plot goes through the critical point.

**Figure 155.8** Gain margin  $K_M$  and phase margin  $\gamma$  on the polar plot.



The angle  $\gamma$  in Fig. 155.8 is the angle measured from the negative real axis to the radial line through the point where the polar plot crosses the unit circle. If the angle  $\gamma$  were zero, the polar plot would go through the  $-1$  point. The angle  $\gamma$ , called the **phase margin**, is a measure of the closeness of the polar plot to the critical point. A positive phase margin indicates a stable system, as does a gain margin greater than one.

For the system shown in Fig. 155.1, the relationship between the damping ratio  $\zeta$  and the phase margin  $\gamma$  is

$$\zeta = \frac{\tan \gamma \sqrt{\cos \gamma}}{2} \quad (155.12)$$

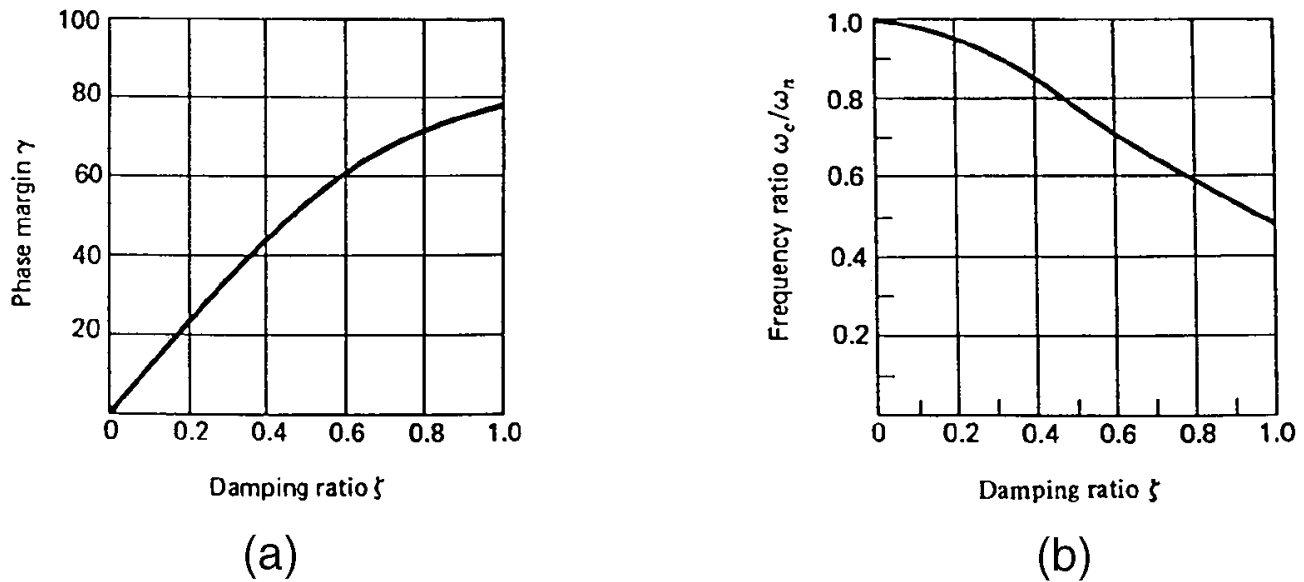


A plot of this relationship is shown in Fig. 155.9(a). The frequency  $\omega_c$  at which the open-loop frequency response crosses the unit circle is

$$\omega_c/\omega_n = [\sqrt{4\zeta^4 + 1} - 2\zeta^2]^{1/2} \quad (155.13)$$

This relationship is shown in Fig. 155.9(b). By knowing the phase margin  $\gamma$ , the damping ratio  $\zeta$  may be determined from Fig. 155.9(a). From Fig. 155.9(b), the ratio  $\omega_c/\omega_n$  can be found. Thus, by knowing  $\omega_c$ , the natural frequency  $\omega_n$  can now be calculated. This method for ascertaining the damping ratio  $\zeta$  and natural frequency from the phase margin  $\gamma$  and the frequency  $\omega_c$  at which the open-loop frequency response crosses the unit circle yields good approximations for systems other than that shown in Fig. 155.1.

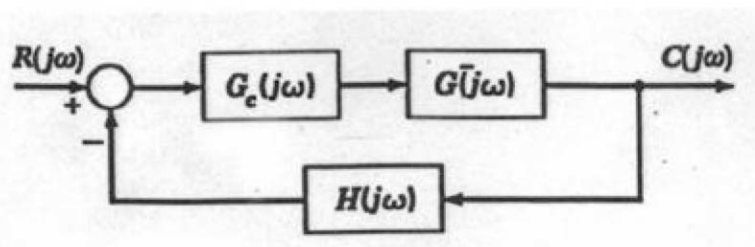
**Figure 155.9** Correlation between  $\zeta$ ,  $\omega_n$ , and phase margin  $\gamma$ : (a) plot of  $\gamma$  versus  $\zeta$ , and (b) plot of  $(\omega_c/\omega_n)$  versus  $\zeta$ .



## 155.4 Series Compensation

A change in the gain  $K$  changes the scale factor of the polar plot, but does not affect the basic shape of the plot. In the design of control systems, it is often necessary to change the shape of the polar plot in order to achieve the desired dynamic performance. A common way of doing this is to insert elements in series with the feedforward portion of the control, as is illustrated by the elements  $G_c(j\omega)$  shown in Fig. 155.10. This method of compensating the performance of a control system is called **cascade** or **series compensation**.

**Figure 155.10** Series compensator  $G_c(j\omega)$ .



## Lead Compensation

The frequency response of a phase lead compensator is

$$\frac{1 + j\tau_1\omega}{1 + j\tau_2\omega} \quad \tau_1 > \tau_2 \quad (155.14)$$

The asymptotic approximation to the Bode diagram [Bode, 1945] for  $(1 + j\tau_1\omega)/(1 + j\tau_2\omega)$  is shown in Fig. 155.11. The frequency at which the maximum phase lead occurs is at

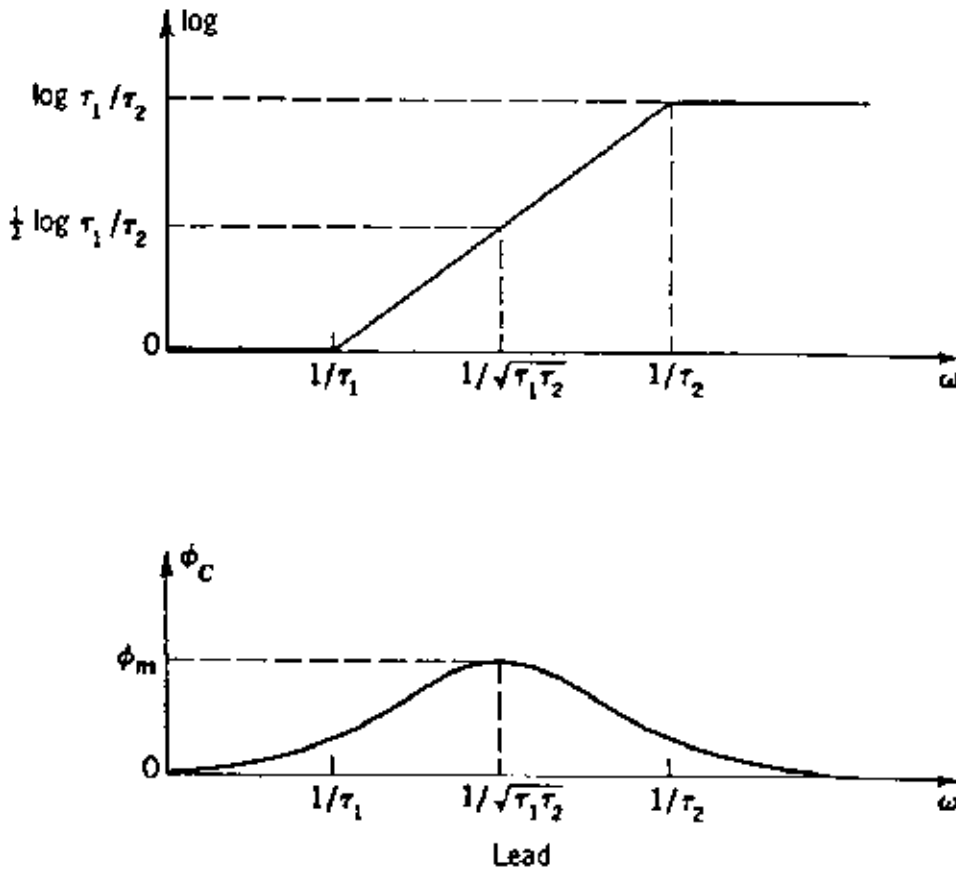
$$\omega_m = 1/\sqrt{\tau_1\tau_2} \quad (155.15)$$

The value of the maximum phase lead is

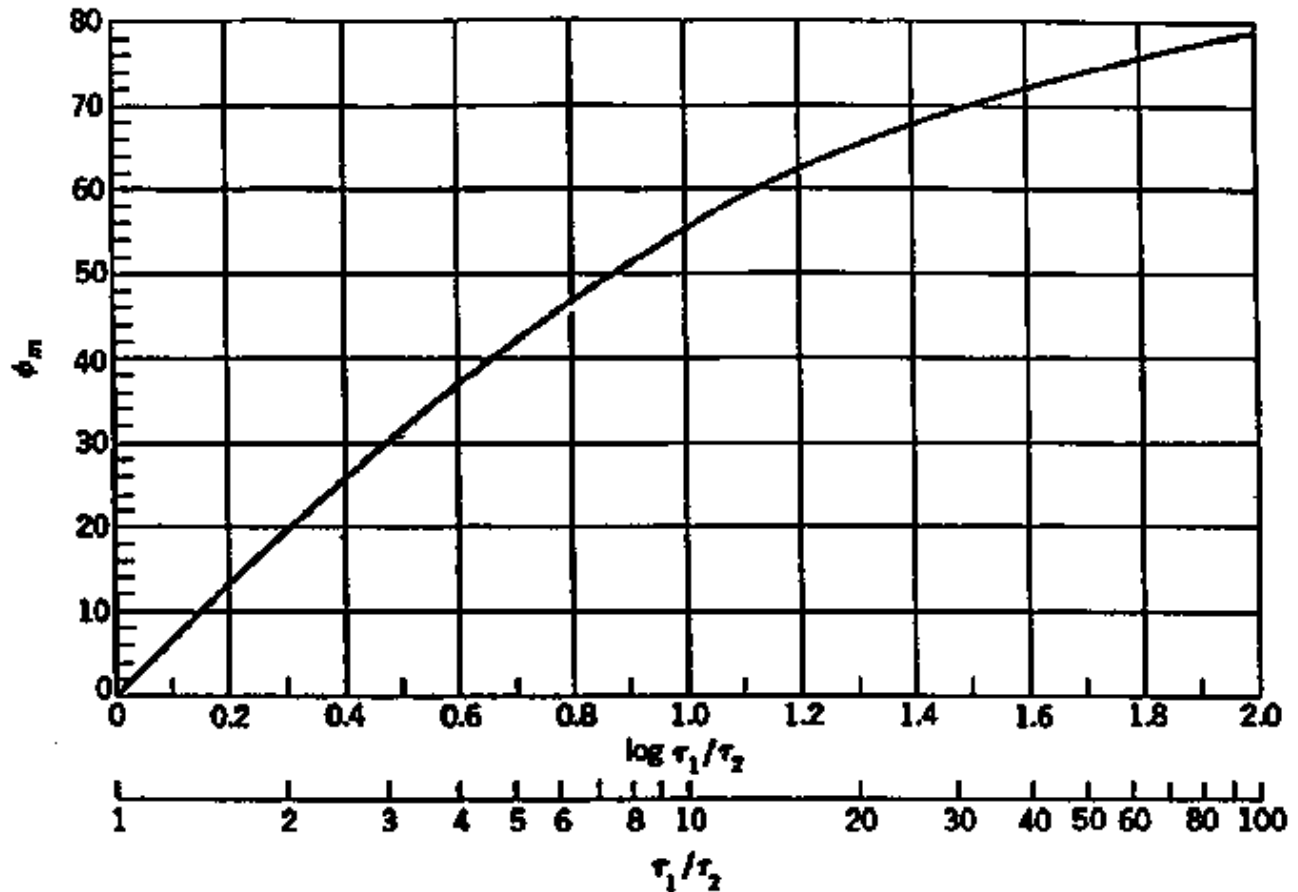
$$\phi_m = \tan^{-1} \frac{(\tau_1/\tau_2) - 1}{2\sqrt{\tau_1/\tau_2}} \quad (155.16)$$

Figure 155.12 shows a plot of  $\phi_m$  versus both  $\log \tau_1/\tau_2$  and  $\tau_1/\tau_2$ .

**Figure 155.11** Bode diagram for a phase lead compensator.

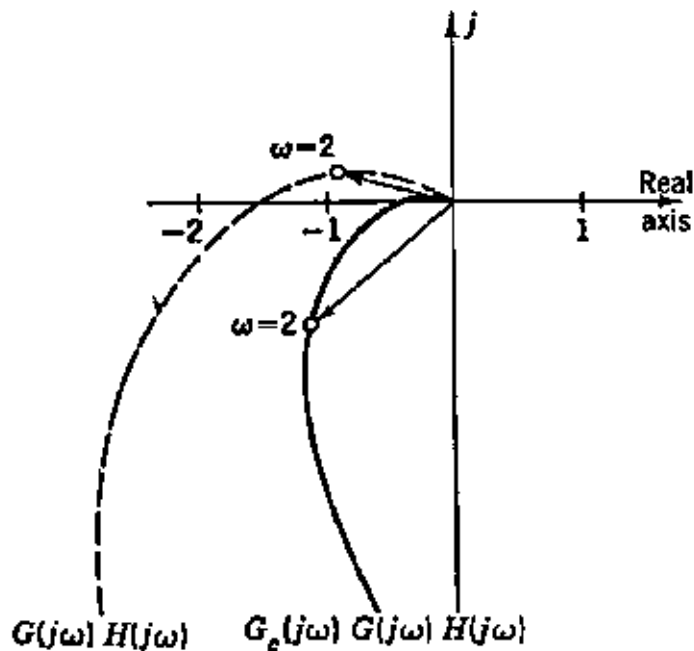


**Figure 155.12** Lead compensator characteristics: maximum phase shift  $\phi_m$  versus  $\log \tau_1/\tau_2$ .



The effect of using a phase lead compensator in series with the feedforward portion of a control system is illustrated in Fig. 155.13. The dashed curve is the frequency response  $G(j\omega)H(j\omega)$  for the uncompensated control system. This system is unstable. The addition of the lead compensation  $G_c(j\omega)$  to reshape the high frequency portion of the polar plot is shown by the solid line curve. Note that lead compensation rotates a typical vector such as that for  $\omega = 2$  in a counterclockwise direction away from the  $-1$  point. Because of the counterclockwise rotation, lead compensation has the very desirable effect of increasing the natural frequency, which increases the speed of response of the system.

**Figure 155.13** Use of phase lead to reshape a polar plot.



To select a lead compensator, it is necessary to specify the values of both  $\tau_1$  and  $\tau_2$ . Because of the two unknowns  $\tau_1$  and  $\tau_2$ , the selection of a lead compensator to achieve desired design specifications is basically a trial-and-error process. However, a systematic procedure which rapidly converges is described in the following steps.

1. Determine the phase margin for the uncompensated system.
2. Select a value for  $\phi_m$  which is the difference between the desired phase margin and the value obtained in step 1, plus a small additional amount, such as  $5^\circ$ .
3. Determine the ratio  $\tau_1/\tau_2$  from Fig. 155.12.
4. Determine the frequency where the log magnitude for the uncompensated system is  $-0.5 \log(\tau_1/\tau_2)$ . Use this frequency for  $\omega_m$ .
5. Because the phase lead compensator provides a gain of  $0.5 \log(\tau_1/\tau_2)$  at  $\omega_m = 1/\sqrt{\tau_1\tau_2}$ , this will be the frequency where the compensated system crosses the unit circle. Determine the resulting phase margin for the compensated system.

If the phase margin is too small, increase the value  $\tau_1/\tau_2$  and, if it is too large, decrease the ratio  $\tau_1/\tau_2$  and then repeat the steps.

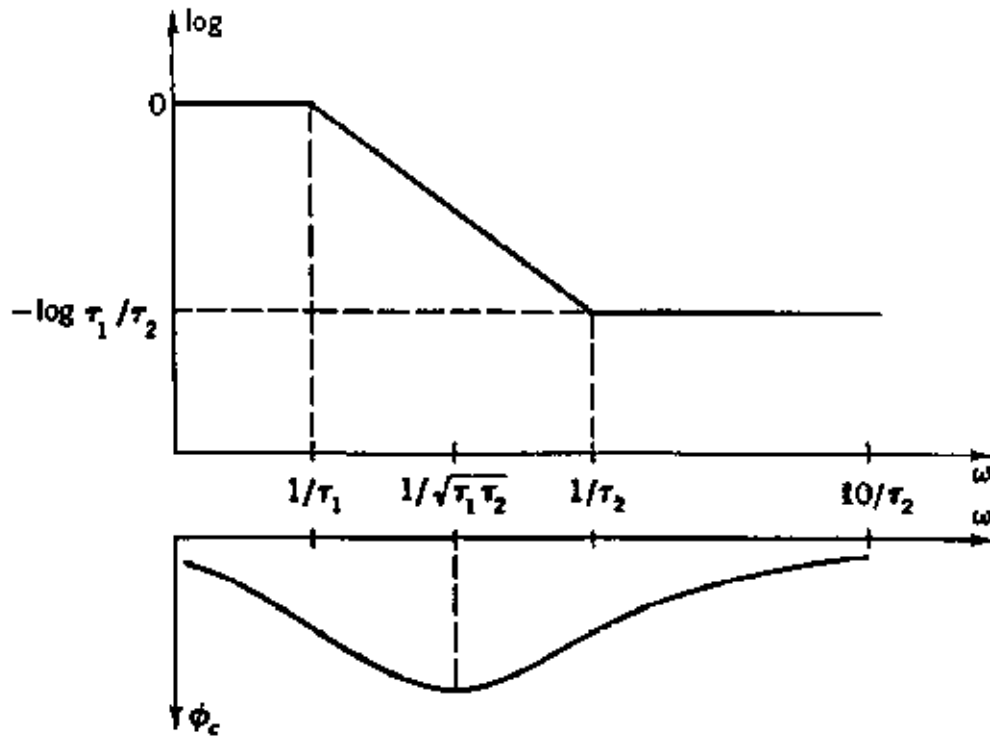
## Lag Compensation

The frequency response for a phase lag compensator is

$$\frac{1 + j\tau_2\omega}{1 + j\tau_1\omega} \quad \tau_1 > \tau_2 \quad (155.17)$$

The Bode diagram for  $(1 + j\tau_2\omega)/(1 + j\tau_1\omega)$  is shown in Fig. 155.14. Note that when  $\omega = 10/\tau_2$ , the negative phase shift  $\phi_c$  is very small. The negative phase shift associated with lag compensation is usually undesirable. The effectiveness of lag compensation is attributed to the attenuation ( $-\log \tau_1/\tau_2$ ) that occurs at high frequencies.

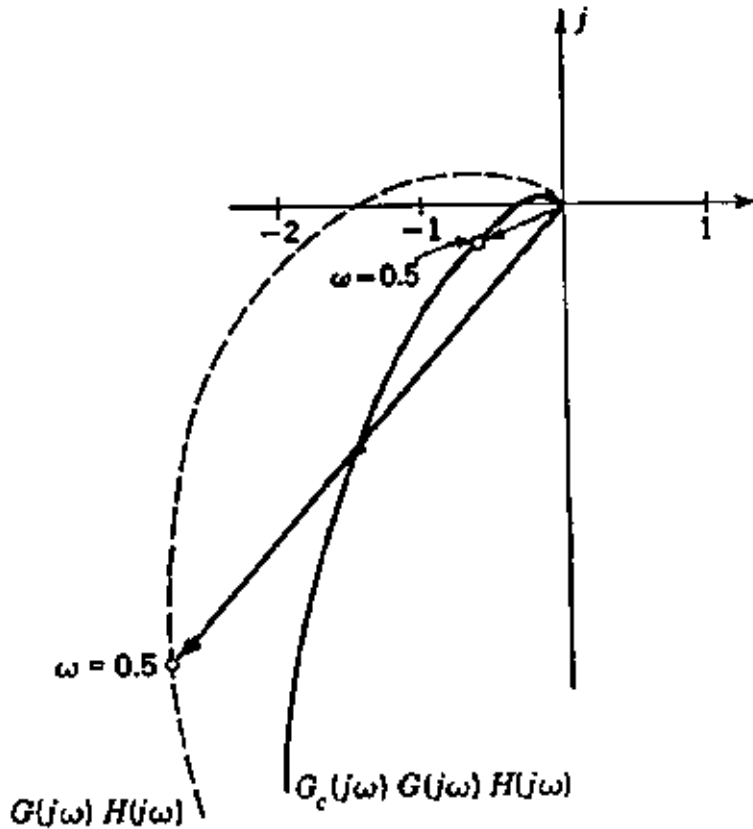
**Figure 155.14** Bode diagram for a lag compensator.



The effect of using a phase lag compensator in series with the feedforward portion of a control system is illustrated in Fig. 155.15. The dashed curve is the frequency response of the uncompensated system. Note that the effect of lag compensation is to shorten a typical vector, such as that for  $\omega = 0.5$ , and to rotate it slightly in a counterclockwise direction. A procedure for determining the lag compensator for obtaining a desired dynamic performance is described in the following steps.

1. Add  $5^\circ$  to the desired phase margin and then subtract  $180^\circ$  from this result.
2. For the uncompensated system, determine the value of the  $\log |G(j\omega)H(j\omega)|$  at the angle determined in step 1.
3. Set this value of  $\log |G(j\omega)H(j\omega)|$  equal to  $\log \tau_1/\tau_2$ . When the lag compensator is added to the uncompensated system, this will be the point where the resultant system crosses the unit circle (i.e.,  $\log |G(j\omega)H(j\omega)| - \log \tau_1/\tau_2 = 0$ ). This will be the  $10/\tau_2$  frequency.

**Figure 155.15** Use of phase lag to reshape a polar plot.



## Lag-lead Compensation

A lag-lead compensator is a series combination of a lag and a lead network. The general transfer function for **lag-lead compensation** is

$$\frac{1 + jc\tau_2\omega}{1 + jc\tau_1\omega} \frac{1 + j\tau_1\omega}{1 + j\tau_2\omega} \quad \tau_1 > \tau_2 \quad (155.18)$$

Rather than using a lag and a lead compensator in series, it is possible to use a single compensator. The Bode diagram for a typical lag-lead compensator is shown in [Fig. 155.16 \[Palm, 1986\]](#).

Because  $c\tau_1 > c\tau_2 > \tau_1 > \tau_2$ , the lag compensation takes place before and at lower frequency than the lead compensation. A factor of five provides a reasonable separation between the lag and lead compensations:

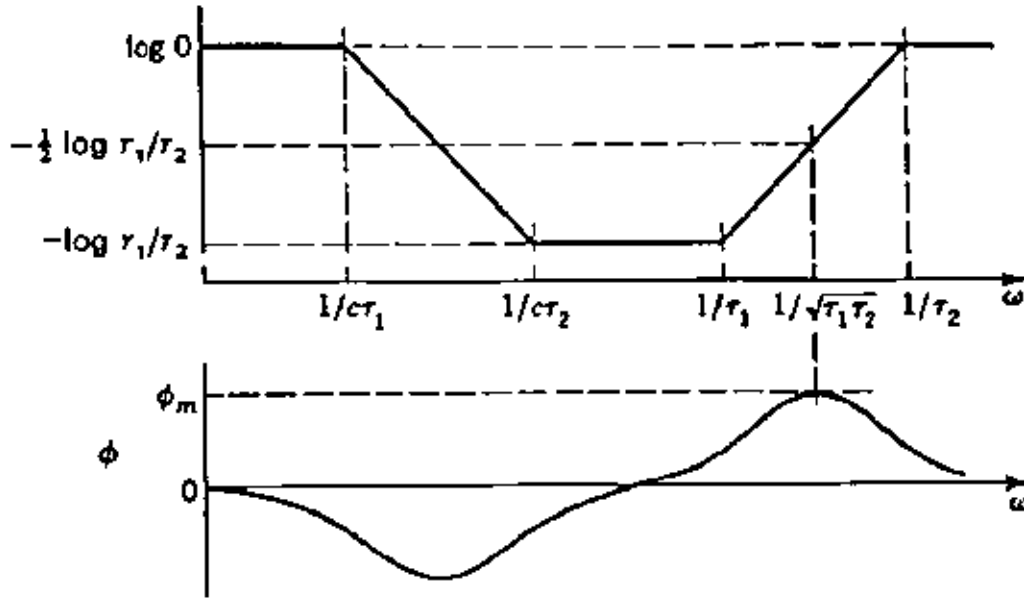
$$\frac{1}{\tau_1} = 5 \frac{1}{c\tau_2}$$

or

$$c = 5 \frac{\tau_1}{\tau_2} \quad (155.19)$$

The maximum phase shift  $\phi_m$  occurs at  $\omega = 1/\sqrt{\tau_1 \tau_2}$ , and the corresponding gain is  $-0.5 \log \tau_1/\tau_2$ . This is the same as for a lead compensator, except that the sign of the gain  $0.5 \log \tau_1/\tau_2$  is negative. This feature makes the lag-lead compensator considerably more effective than the lead compensator only.

**Figure 155.16** Bode diagram for a lag-lead compensator.

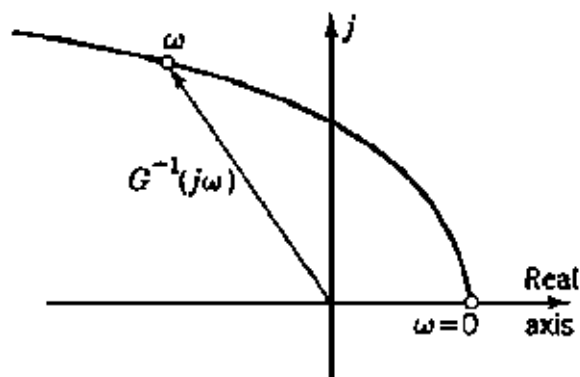


## 155.5 Internal Feedback

Another method commonly used to alter frequency response characteristics is that of providing a separate **internal feedback** path about certain components [D'Souza, 1988]. A plot of the function  $G^{-1}(j\omega) = 1/G(j\omega)$  is called an inverse polar plot. Figure 155.17 shows a typical inverse polar plot for the function  $G^{-1}(j\omega)$ . At any frequency  $\omega$ , the vector from the origin to a point on the plot defines the vector  $G^{-1}(j\omega)$  for that frequency. The length of the vector is  $|G^{-1}(j\omega)| = 1/|G(j\omega)|$ , and the angle is

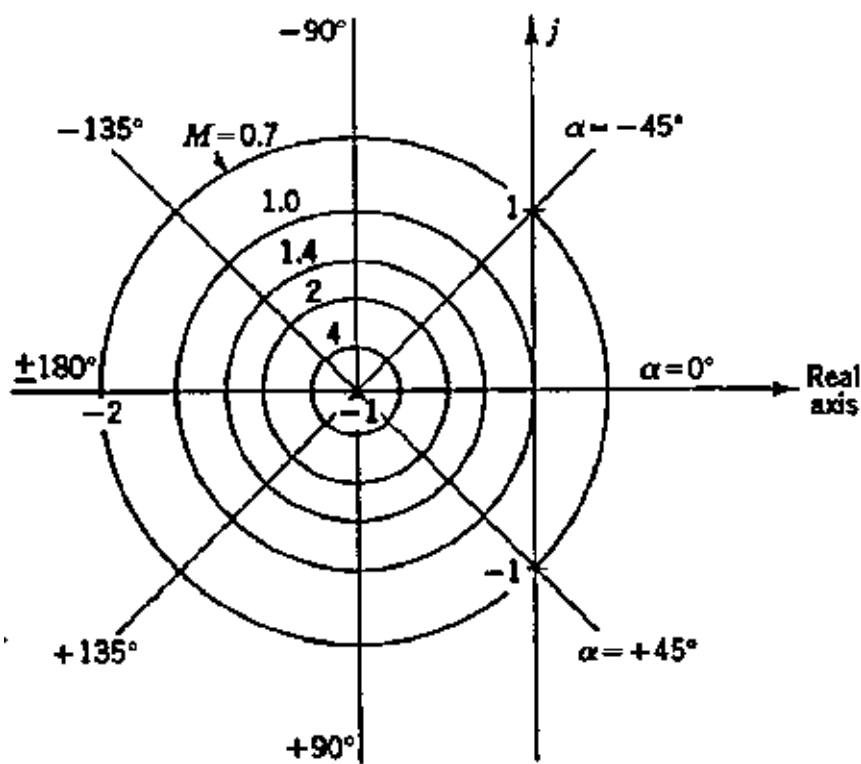
$$\angle G^{-1}(j\omega) = \angle \frac{1}{G(j\omega)} = -\angle G(j\omega) \quad (155.20)$$

**Figure 155.17** Typical inverse polar plot  $G^{-1}(j\omega)$ .



On the inverse plane, lines of constant  $M$  are circles of radius  $1/M$ . The center of these concentric  $M$  circles is at the point  $x = -1$  and  $y = 0$  (e.g., the  $-1$  point). A plot of the  $M$  circles on the inverse plane is shown in Fig. 155.18. Because the reciprocal of  $-1$  is still  $-1$ , this point has the same significance for an inverse polar plot as for a direct polar plot. The lines of constant  $\alpha = \angle[C(j\omega)/R(j\omega)] = -\angle[R(j\omega)/C(j\omega)]$  are radial straight lines (rays) which pass through the  $-1$  point.

**Figure 155.18**  $M$  circles and  $\alpha$  rays on the inverse plane.



As is illustrated in Fig. 155.19, the angle  $\psi$  of a radial line drawn from the origin and tangent to any  $M$  circle is



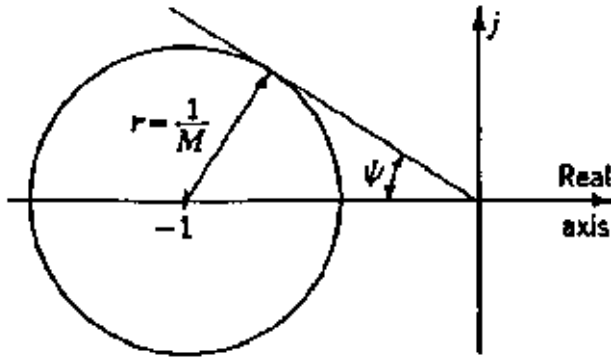
$$\sin \psi = 1/M \quad (155.21)$$

The general procedure for determining the required gain  $K$  to yield a desired  $M_p$  is as follows.

1. Plot the inverse function  $KG^{-1}(j\omega)H^{-1}(j\omega)$  .
2. Construct the tangent line at the angle  $\psi = \sin^{-1}(1/M_p)$  .
3. Construct the circle which is tangent to both the  $KG^{-1}(j\omega)H^{-1}(j\omega)$  plot and the tangent line.
4. The center of the circle is at the point  $-A$  . The desired gain is  $K = A$  .

When the function  $G^{-1}(j\omega)H^{-1}(j\omega)$  is plotted rather than  $KG^{-1}(j\omega)H^{-1}(j\omega)$  , then  $A$  is equal to the factor  $K_c$  by which the gain should be changed to yield the desired  $M_p$  .

**Figure 155.19** Tangent line to an  $M$  circle.

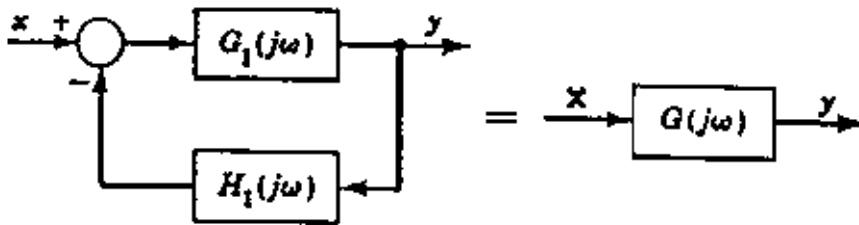


The major advantage of using the inverse plane is realized for systems with internal feedback. For the system of Fig. 155.20,

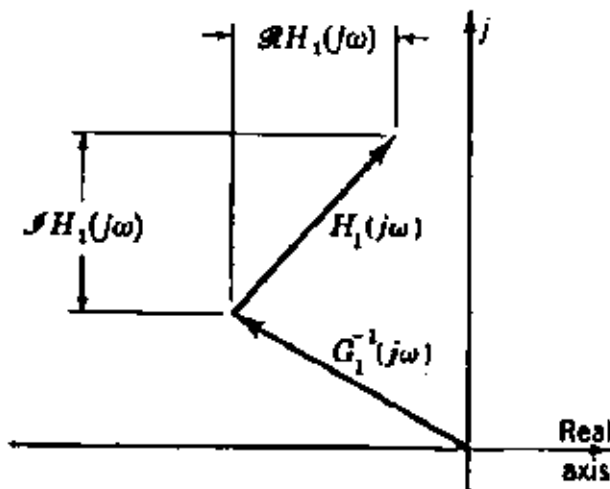
$$G^{-1}(j\omega) = \frac{1 + G_1(j\omega)H_1(j\omega)}{G_1(j\omega)} = G_1^{-1}(j\omega) + H_1(j\omega) \quad (155.22)$$

As is illustrated in Fig. 155.21, the vectors  $G_1^{-1}(j\omega)$  and  $H_1(j\omega)$  may be added as vector quantities to yield  $G^{-1}(j\omega)$  . For a given  $G_1^{-1}(j\omega)$  plot,  $H_1(j\omega)$  may be determined to move any given point to a desired location, such as a point of tangency with an  $M_p$  circle, or to yield a desired phase margin.

**Figure 155.20** Internal feedback  $H_1(s)$  placed about  $G_1(s)$ .



**Figure 155.21** Vector addition of  $G_1^{-1}(j\omega)$  and  $H_1(j\omega)$  to yield  $G^{-1}(j\omega)$ .



If  $H_1(s) = \alpha$  a constant, then the  $G^{-1}(j\omega)$  plot is shifted to the right by the distance  $\alpha$ . If  $H_1(s) = \beta s$ , then  $H_1(j\omega) = j\beta\omega$ , and the  $G^{-1}(j\omega)$  plot is shifted upwards vertically in proportion to the frequency  $\omega$ . If  $H(s) = \alpha + \beta s$ , then the  $G^{-1}(j\omega)$  plot is shifted both to the right and upwards vertically.

## 155.6 Compensation on the $S$ Plane

The transfer function for a phase lead compensator may be written in the form

$$\frac{1 + \tau_1 s}{1 + \tau_2 s} = \frac{s + 1/\tau_1}{s + 1/\tau_2} = \frac{s - z}{s - p} \quad \tau_1 > \tau_2 \quad (155.23)$$

where  $z = -1/\tau_1$  is a zero, and  $p = -1/\tau_2$  is a pole. The root locus plot for the system whose characteristic equation is  $s(s - r_1)(s - r_2) + K = 0$  is shown in Fig. 155.22(a). Assume that the root locus plot goes through the root location for the dominant roots (point  $a \pm jb$ ) shown in Fig. 155.22(b). The zero is drawn directly under the point  $a + jb$ . Application of the angle condition gives

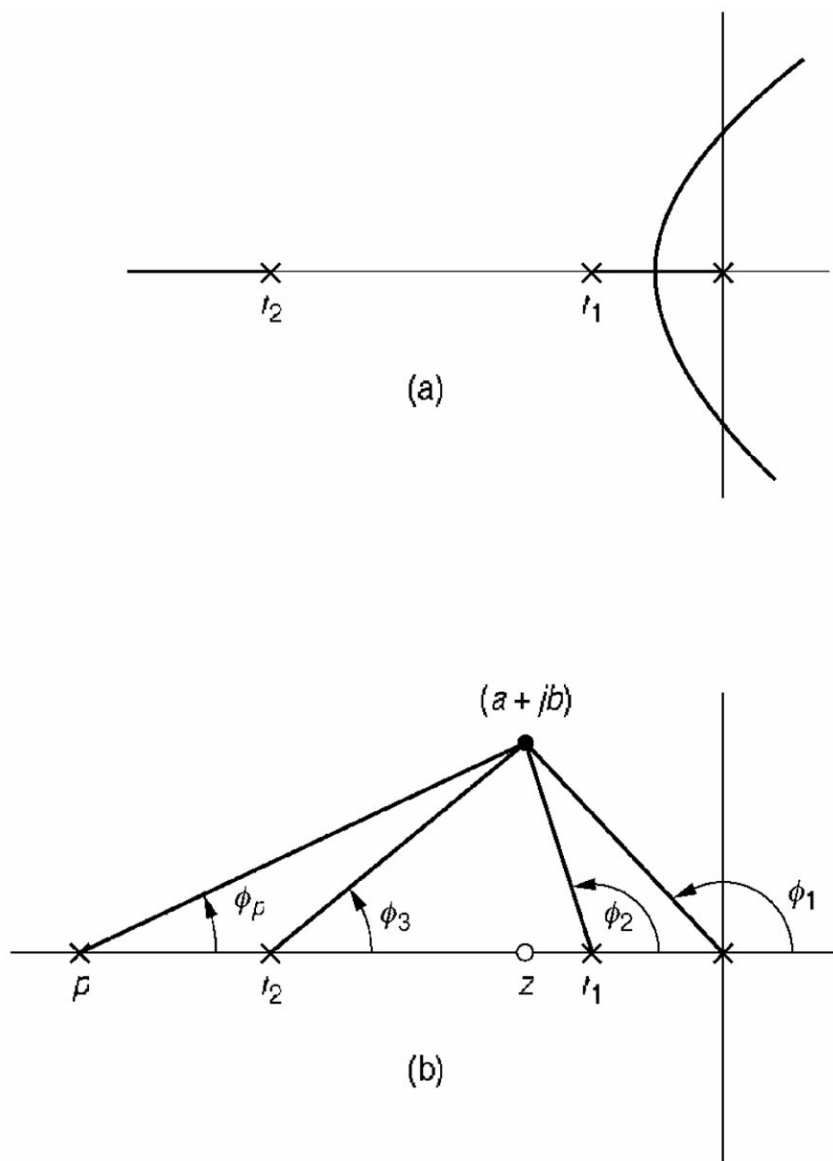
$$\phi_1 + \phi_2 + \phi_3 + \phi_p - 90^\circ = 180^\circ$$

Hence,

$$\phi_p = 270^\circ - (\phi_1 + \phi_2 + \phi_3)$$

The angle  $\phi_p$  determines the location of the pole. Similarly, phase lag and lag-lead compensators [Dorf, 1989] may be designed on the  $s$  plane.

**Figure 155.22** (a) Root locus plot for  $s(s - r_1)(s - r_2) + K = 0$ , and (b) addition of phase lead compensation.



## HAROLD HAZEN AND THE THEORY OF SERVOMECHANISMS

Before the 1930s, feedback controls mostly concerned regulators and governors—devices to keep motion steady and stable. That changed in 1934, when Harold Hazen published "Theory of Servo Mechanisms" in the *Journal of the Franklin Institute*. This paper provided a consistent taxonomy of feedback devices (relay, pulsed, and continuous) and laid out a "transient analysis" approach for designing servos with specified response. It also noted that the theory of servomechanisms can be applied to the speed control of steam turbines and water wheels, the stabilization of ships by gyroscopes, the operation of gyrocompass repeaters, the automatic stabilization and guiding of aircraft, and "in fact the automatic recording or control of almost any measurable or measurable and controllable physical quantity." Hazen shifted the emphasis of control engineering from static to dynamic behavior, from regulation to control, and from diverse approaches in different applications to a single approach based on a unified view of dynamic systems.

## Defining Terms

**Cascade compensation:** The insertion of elements in series with the feedforward portion of a control system in order to obtain the desired performance.

**Compensation:** The method of improving the performance of a control system by inserting an additional component within the structure of the system.

**Damping ratio:** A dimensionless number which is a measure of the amount of damping in a system.

**Gain margin:** If the gain of a system were changed by this factor, called the gain margin, the  $G(j\omega)H(j\omega)$  plot would go through the -1 point.

**Internal feedback:** The method of system compensation in which an internal feedback path is provided about certain components.

**Lag compensation:** A component that provides significant attenuation over the frequency range of interest.

**Lag-lead compensation:** A component that provides both significant positive phase angle and attenuation over the frequency range of interest.

**Lead compensation:** A component that provides significant positive phase angle over the frequency range of interest.

**Natural frequency:** The frequency at which a system would oscillate if there was no damping.

**Phase margin:** The angle measured from the negative real axis to the radial line where the polar plot crosses the unit circle.

**Series compensation:** The insertion of elements in series with the feedforward portion of a control system in order to obtain the desired performance.

## References

- Bode, H. W. 1945. *Network Analysis and Feedback Amplifier Design*. D. Van Nostrand, Princeton, NJ.
- Dorf, R. C. 1995. *Modern Control Systems*, 7th ed. Addison-Wesley, Reading, MA.
- D'Souza, A. F. 1988. *Design of Control Systems*. Prentice Hall, Englewood Cliffs, NJ.
- Nichols, J. H. and Phillips, R. S. 1947. *Theory of Servomechanisms*. McGraw-Hill, New York.
- Palm III, W. J. 1986. *Control Systems Engineering*. John Wiley & Sons, New York.
- Raven, F. H. 1995. *Automatic Control Engineering*, 5th ed. McGraw-Hill, New York.

Marlin, T. E “Process Control”  
*The Engineering Handbook.*  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

### 156.1 Control Performance and Design Decisions

Single-Variable Control • Multiple Variables with One Dependent Controlled Variable • Multiple Input-Output Control

**Thomas E. Marlin**

*McMaster University*

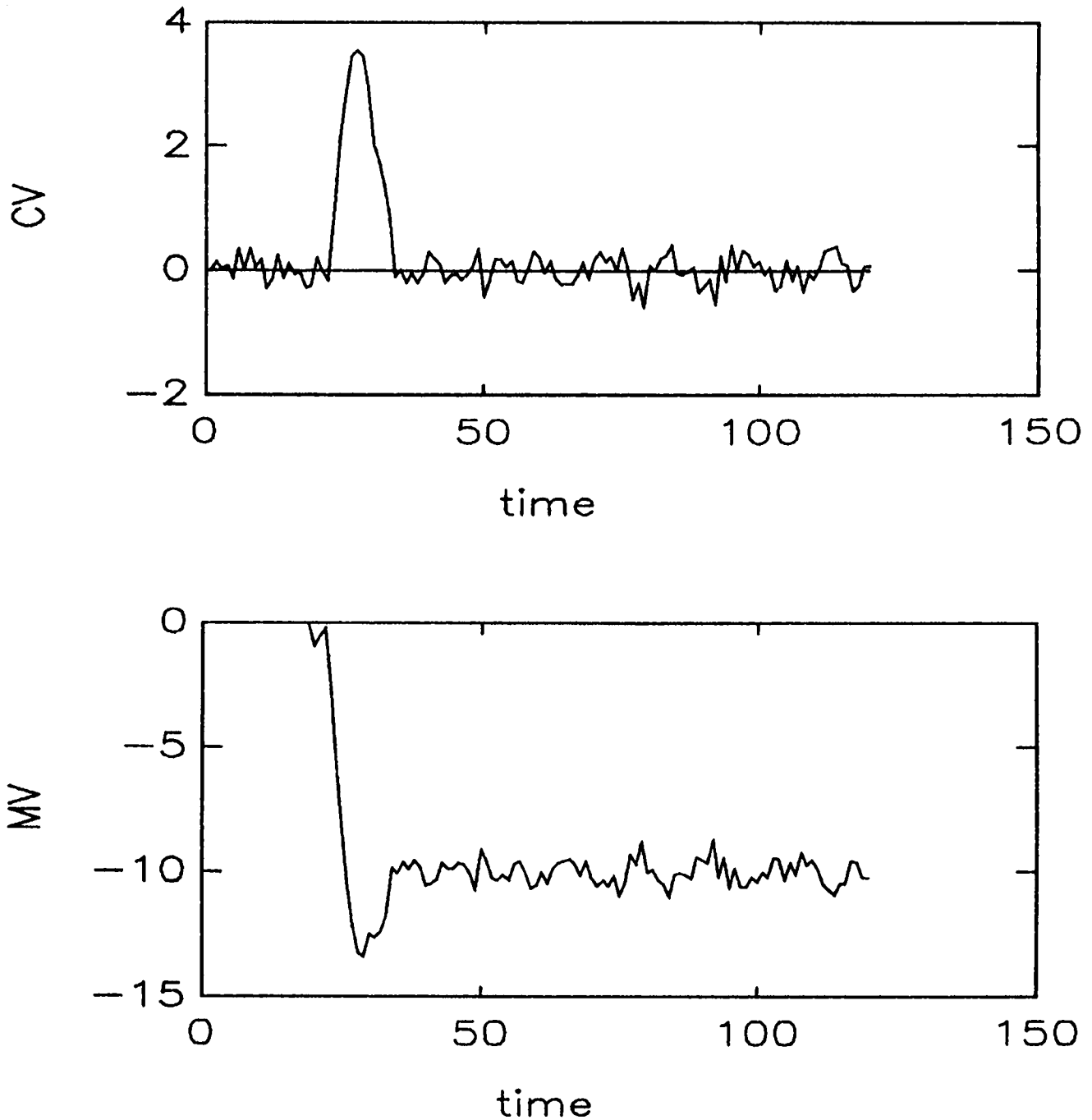
Process control addresses the application of automatic control theory to the process industries. These industries typically involve the continuous processing of fluids and slurries in chemical reactors, physical separation units, combustion processes, and heat exchangers in industries such as chemicals, petroleum, pulp and paper, food, steel, and electrical power generation. These industries share some general characteristics that influence the control application. First, the process design and operating conditions impose fundamental limits on the best possible control performance, so that the engineer should strive to design processes and select variables that are inherently easiest to control. Second, the control structure—that is, the measured and manipulated variables—must be selected from the large number of candidates. Third, the control algorithms must be matched to the control structure selected. Fourth, the processes tend to be highly nonlinear and difficult to model accurately using first principles because of the complex physiochemical systems; these complex and time-varying dynamics require all control designs and applications to explicitly consider robustness to ensure stability and performance over the expected range of conditions.

### 156.1 Control Performance and Design Decisions

---

Generally, the goal of process control is to reduce the variability of key process variables by adjusting selected manipulated variables, thereby moving variability to less important areas of the process. In evaluating control performance, many factors are relevant, as shown in [Fig. 156.1](#). First, the controlled variable should have a small deviation from its set point (or reference value); this can be measured by the integral of the error squared,  $\int (SP - CV)^2 dt$ , for short data sets and by the variance of the error for long data sets. In addition, the behavior of the manipulated variable is crucial, as large variance of this variable could lead to propagation of disturbances or to damage of process equipment. Most importantly, the stability and performance should remain acceptable for expected changes in the process dynamics; these changes are due to unavoidable errors in empirical identification and to changes in operating conditions such as production rate. Thus, the application of process control must satisfy a complex, multiobjective performance measure.

**Figure 156.1** Example control loop dynamic response.



Process control design requires many decisions to meet the demands of the process industries; these typically are (1) selecting sensors that reflect process performance, (2) selecting final elements with good dynamic responses and sufficient range, (3) ensuring that the feedback process



dynamics are favorable to good control performance, (4) selecting a structure for linking controlled and manipulated variables, and (5) selecting algorithms and tuning parameters to give robust performance. These decisions will be reviewed for single-variable and multivariable control.

## Single-Variable Control

Many complex control systems involve multiple, single-loop controllers; thus, the principles of single-loop control must be mastered as a basis for realistic, multivariable control systems. In these systems one measured variable is regulated by adjusting one manipulated variable. In most cases the algorithm used is the proportional-integral-derivative (PID) controller, which is given for a continuous system:

$$MV = K_c \left( E + \frac{1}{T_I} \int_0^t E \, dt' + T_D \frac{dE}{dt} \right) \quad (156.1)$$

with the error,  $E = SP - CV$ . Both the process and the controller appear in the closed-loop system and influence the control performance. To consider the behavior of the system, see the simple block diagram in [Fig. 156.2](#). In this example the dynamics between the manipulated variable and the outlet controlled variable have been approximated by a first order with dead time linear model, as can many overdamped systems of higher order. The transfer functions for the process (including sensor and final element dynamics) and the controller are given as follows:

$$\begin{aligned} \text{Feedback process:} \quad \frac{CV(s)}{MV(s)} &= \frac{K_p e^{-\theta s}}{\tau s + 1}; \\ \text{Disturbance:} \quad \frac{CV(s)}{D(s)} &= \frac{K_D e^{-\theta_D s}}{\tau_D s + 1} \end{aligned} \quad (156.2)$$

$$\text{Controller:} \quad \frac{MV(s)}{E(s)} = K_c \left( 1 + \frac{1}{T_I s} + T_D s \right) \quad (156.3)$$

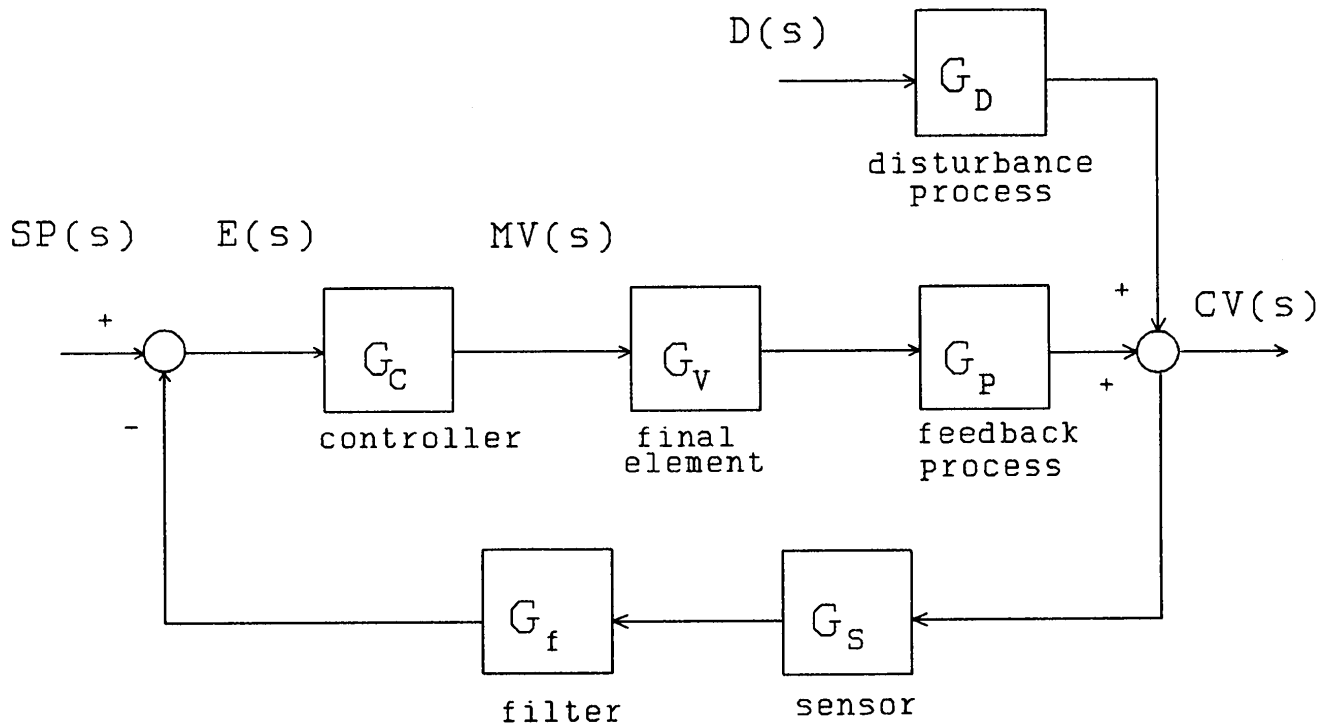
The transfer function describing the relationship between the disturbance and the controlled variable is

$$\frac{CV(s)}{D(s)} = \frac{G_D(s)}{1 + G_p(s)G_v(s)G_c(s)G_f(s)G_s(s)} \quad (156.4)$$

The stability of the closed-loop system depends on the terms in the characteristic equation, the denominator of the closed-loop transfer function; generally, the slower the dynamics of the elements in the feedback process, the less aggressive the feedback controller for stable behavior. Many methods for tuning the PID controller have been proposed—for example, the

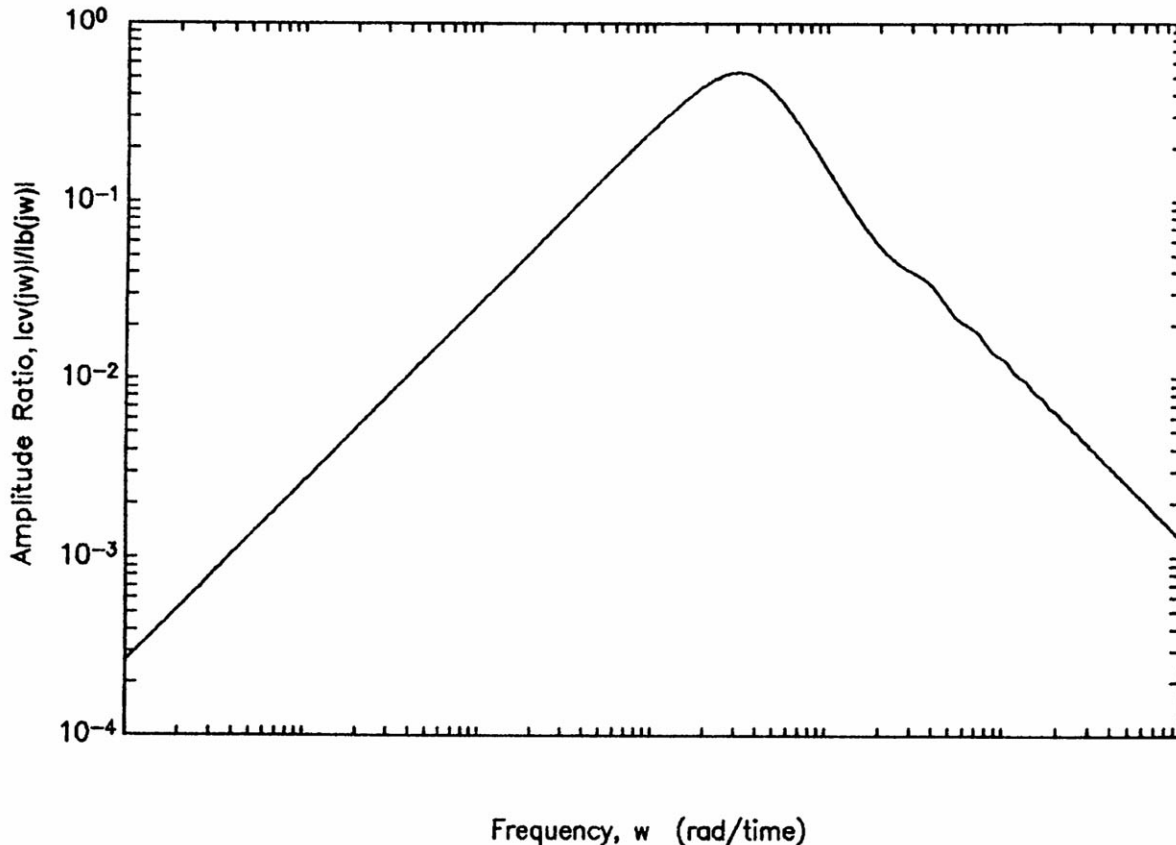
Ziegler-Nichols method [1942]—although many suffer from poor robustness. Methods described in Fertik [1975] and Marlin [1994] provide better robustness.

**Figure 156.2** Single-loop block diagram.



The performance of the system depends on all terms in Eq. (156.4), and important insights can be obtained by considering the frequency response of  $|CV(j\omega)|/|D(j\omega)|$  from Eq. (156.4), shown in Fig. 156.3. For disturbances with very low frequencies, feedback control is much faster than the disturbance, and feedback is thus quite effective. For very high frequencies the disturbance time constant attenuates the effect of the disturbance on the controlled variable, and the performance is quite good, although not due to feedback compensation. In some intermediate range of frequencies, resonance occurs and feedback control performance is not generally very good. The location of the resonance peak depends on the feedback dynamics and the feedback controller.

**Figure 156.3** Closed-loop frequency response.



A summary of the effects of various elements in the feedback loop on the performance of single-loop control is given in [Table 156.1](#). This table provides the basis for design decisions to improve control. The process provides the most important limits to feedback control performance. It is important to note that large dead times and time constants in the feedback loop are always detrimental, whereas large time constants in the disturbance path can improve control performance.

## Multiple Variables with One Dependent Controlled Variable

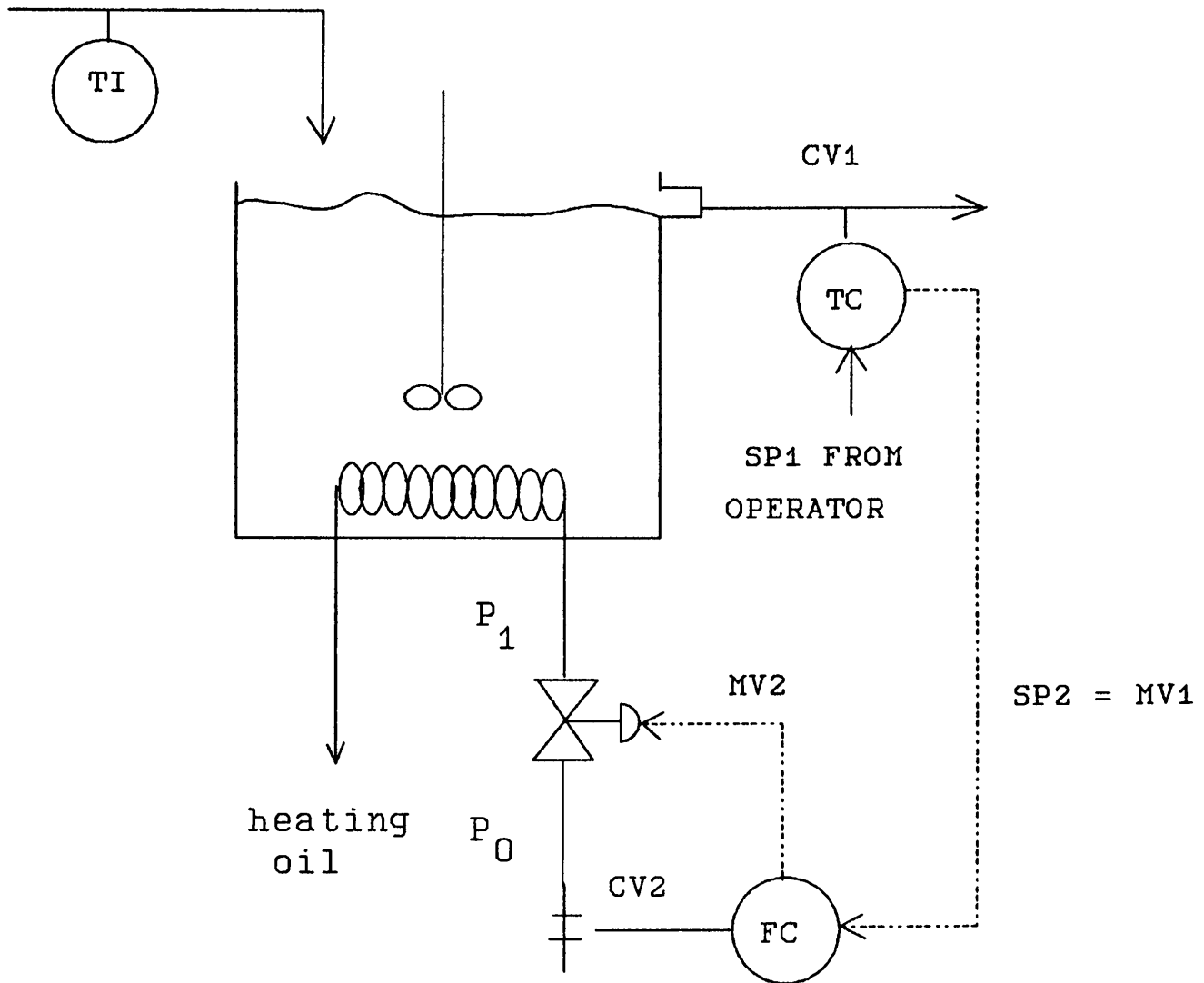
The control performance of a single variable can be improved by applying the results in [Table 156.1](#), without changing the process, by modifying the control structure. Since the feedback process dynamics impose the greatest limit on good performance, the two modified structures described here reduce the feedback dynamics between the manipulated and the controlled variables. The first new structure is **cascade** control, which uses a secondary measured variable to provide an early indication of selected disturbances. As shown in [Fig. 156.4](#), the secondary variable, the measured heating medium flow rate, is controlled using feedback principles, and the set point of the flow controller is adjusted via feedback to achieve the desired outlet temperature. Each of the controllers can use the PID controller algorithm, and the inner or secondary controller must be tuned first. The cascade design provides good performance for only some disturbances; for example, the design in [Fig. 156.4](#) gives good performance for disturbances in the pressure of the heating medium source, but it does not improve performance for inlet temperature disturbances.

**Table 156.1** Summary of Factors Affecting Single-Loop PID Controller Performance

Key Factor	Typical Parameter	Effect on Control Performance
Feedback process gain	$K_p$	The key factor is the product of the process and controller gains. For example, a small process gain can be compensated by a large controller gain. Note that the manipulated variable must have sufficient range.
Feedback process "speed"	$\theta + \tau$	Control performance is always better when this term is small.
Feedback fraction process dead time	$\frac{\theta}{\theta + \tau}$	Control performance is always better when this term is small.
Inverse response	Numerator term in transfer function, $(\tau s + 1)$ with $\tau < 0$	Control performance degrades for large inverse response.
Magnitude of disturbance effect	$K_d  \Delta D $	Control performance is always better when this term is small.
Disturbance dynamics	$\tau_D$	Control performance is best when the disturbance is slow, that is, the time constant is large.
	$\omega_D$	Feedback control is effective for low frequency disturbances and is least effective at the resonant frequency.
	$\theta_D$	Disturbance dead time does not influence performance.
Sensor		Measurement should be accurate. Dynamics should be fast with little noise.
Filter, $1/(1+\tau_f s)$	$\tau_f / (\theta + \tau)$	Filters higher frequency components of measurement. Reduces the variability of the manipulated variable but degrades control as filter time constant is increased.
Final element		Dynamics should be fast without sticking or hysteresis.
		Range should be large enough for response to demands.
Controller execution period ( $\Delta t$ )	$\frac{\Delta t}{\theta + \tau}$	Control performance is best when this parameter is small. Continuous PID tuning correlations can be used by modifying the dead time, $\theta' = \theta + \Delta t/2$ .
Controller tuning	$K_c K_p$	Determined from tuning correlations based on control objectives.
	$\frac{T_I}{(\theta + \tau)}$ $\frac{T_D}{(\theta + \tau)}$	
Modeling errors		Errors in identifying the process model parameters leads to poorer control performance and, potentially, instability. Tuning should consider the estimate of model errors.
Limitations on manipulated variables	$\min < mv(t) < \max$	Limitations on manipulated variables reduce the operating window, that is, the range of achievable conditions. An active limit would cause steady state offset from the set point.

Source: Marlin, T. 1994. *Process Control: Designing Processes and Control Systems for Dynamic Performance*. McGraw-Hill, New York. With permission.

**Figure 156.4** Cascade control system. (Source: Marlin, T. 1994. *Process Control: Designing Processes and Control Systems for Dynamic Performance*. McGraw-Hill, New York. With permission.)



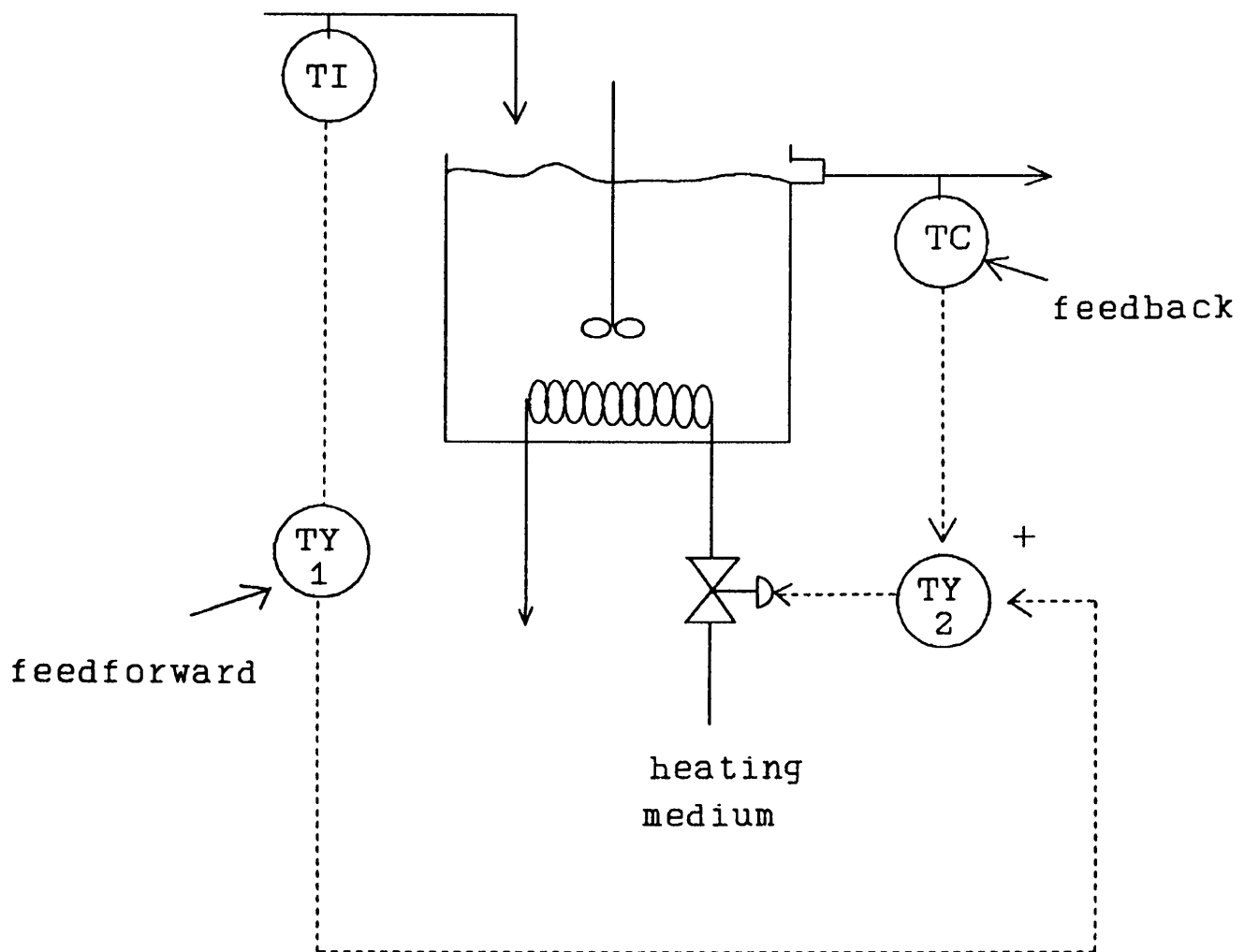
The second structural control modification provides an alternative to feedback by measuring the disturbance before it influences the process. This **feedforward** modification, shown in Fig. 156.5, eliminates the feedback dynamics from the control compensation for the measured disturbance and can theoretically provide excellent control. The feedforward control algorithm to give perfect feedforward compensation can be determined from block diagram algebra to be  $G_{ff}(s) = -G_D(s)/G_p(s)$ ; thus, the feedforward algorithm depends on the models of the process dynamics. When both the process and disturbance transfer functions can be well approximated by first order with dead time transfer functions, the feedforward controller is

$$G_{ff}(s) = \frac{MV(s)}{D(s)} = \frac{K_d}{K_p} \left( \frac{\tau s + 1}{\tau_D s + 1} \right) e^{-(\theta_D - \theta)s} \quad (156.5)$$

Equation (156.5) is often used for feedforward control and can be realized using standard gain, dead time, and lead/lag algorithms. Usually, feedback control is retained because feedforward

compensates for only the measured disturbance(s) and because the feedforward control is perfect only when the models are exact. The feedforward and feedback adjustments can be added, as in Fig. 156.5, using the property of linearity.

**Figure 156.5** Feedforward-feedback control system. (Source: Marlin, T. 1994. *Process Control: Designing Processes and Control Systems for Dynamic Performance*. McGraw-Hill, New York. With permission.)



## Multiple Input-Output Control

Most industrial processes involve many manipulated and controlled variables. Often, multiple single-loop controllers are used to automate the process; thus, the algorithms are identical to those introduced in the previous sections. However, the existence of **interactions** dramatically increases the difficulty of analysis and design. For the simple two input–two output system in Fig. 156.6, each input can influence all outputs [when  $G_{ij}(s) \neq 0$  for  $i \neq j$ ]. As a result the closed-loop

transfer function is

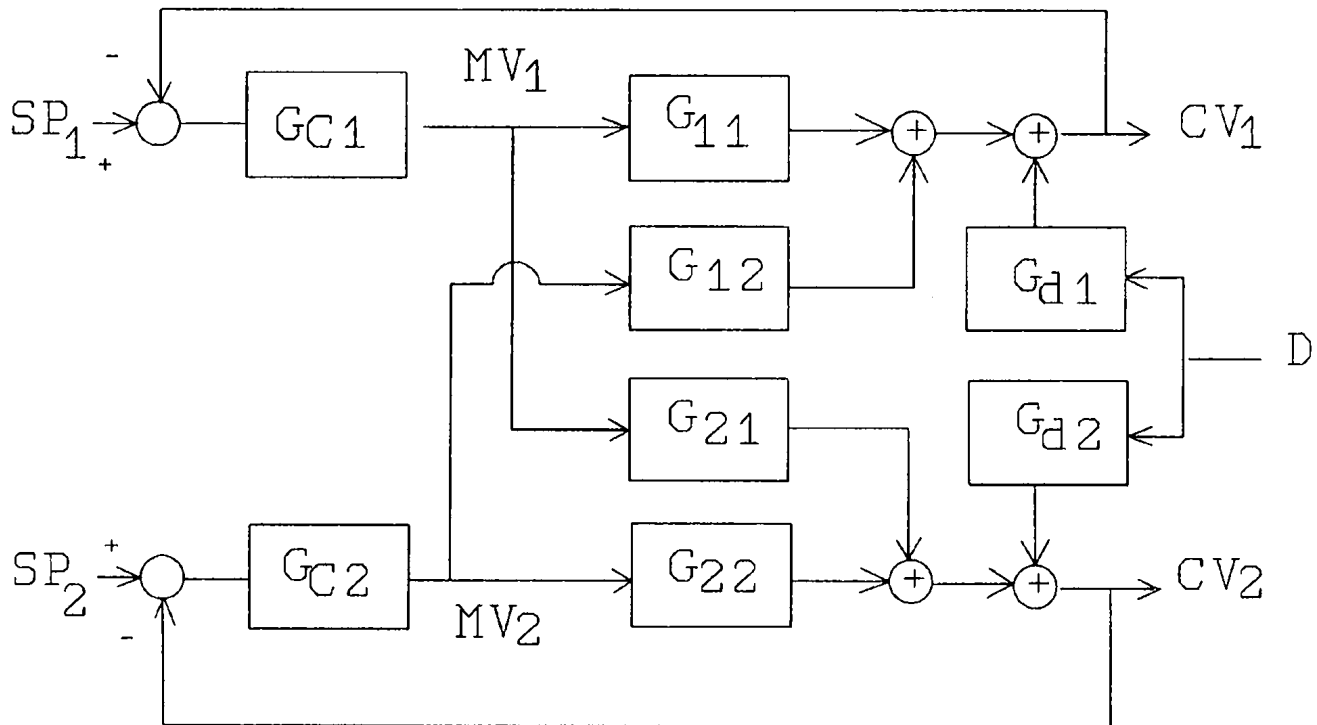
$$\frac{CV_1(s)}{SP_1(s)} = \frac{G_{c1}(s)G_{11}(s) + G_{c1}(s)G_{c2}[G_{11}(s)G_{22}(s) - G_{12}(s)G_{21}(s)]}{CE(s)} \quad (156.6)$$

with

$$CE(s) = 1 + G_{c1}(s)G_{11}(s) + G_{c2}(s)G_{22}(s) + G_{c1}(s)G_{c2}(s)[G_{11}(s)G_{22}(s) - G_{12}(s)G_{21}(s)]$$

Thus, interaction affects the stability of the closed-loop system, requiring modifications to the controller tuning, usually requiring the feedback action to be less aggressive. In addition, interaction can significantly modify the range of achievable steady state operating conditions, termed the *operating window*, and it can even affect whether a set of dependent variables can be controlled by adjusting a set of manipulated variables. Finally, interaction can significantly affect the dynamic control performance.

**Figure 156.6** Multiloop ( $2 \times 2$ ) block diagram.



The first, and often the most important, control design decision is the manner in which the controllers link the controlled and manipulated variables, termed *loop pairing*. The proper loop pairing should yield a control system that can achieve the desired operating window, control all variables independently, provide stable control with the same controller gain signs as for each single-loop system, provide stable performance when some loops are not in operation, and provide good performance for the most likely disturbances. The ability of the system to control the desired output variables depends on the process model given as follows:

$$\begin{bmatrix} CV_1(s) \\ CV_2(s) \end{bmatrix} = \begin{bmatrix} G_{11}(s) & G_{12}(s) \\ G_{21}(s) & G_{22}(s) \end{bmatrix} \begin{bmatrix} MV_1(s) \\ MV_2(s) \end{bmatrix} + \begin{bmatrix} G_{d1}(s) \\ G_{d2}(s) \end{bmatrix} D(s) \quad (156.7)$$

The controlled variables can be maintained exactly at their desired values if the feedback process can be inverted [Rosenbrock, 1974]. That is,

$$\det \begin{bmatrix} G_{11}(j\omega) & G_{12}(j\omega) \\ G_{21}(j\omega) & G_{22}(j\omega) \end{bmatrix} \neq 0 \quad (156.8)$$

Usually, the inverse cannot be taken because of the process dynamics (e.g., dead times), so *perfect* control is not possible. If the requirement is only that the controlled variables can be maintained at their desired values in the *steady state*, the requirement for controllability is

$$\det \begin{bmatrix} G_{11}(0) & G_{12}(0) \\ G_{21}(0) & G_{22}(0) \end{bmatrix} = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \neq 0 \quad (156.9)$$

The criterion in Eq. (156.9) must be satisfied for a process control system to return the controlled variables to their set points.

Before discussing further requirements, a quantitative measure of interaction, the relative gain array, is introduced [McAvoy, 1983]. The relative gain is defined as

$$\lambda_{ij} = \frac{\left\{ \frac{\partial CV_i}{\partial MV_j} \right\}_{MV_k = \text{const}, k \neq j}}{\left\{ \frac{\partial CV_i}{\partial MV_j} \right\}_{CV_k = \text{const}, k \neq i}} = \frac{\left\{ \frac{\partial CV_i}{\partial MV_j} \right\}_{\text{other loops open}}}{\left\{ \frac{\partial CV_i}{\partial MV_j} \right\}_{\text{other loops closed}}} \quad (156.10)$$

The relative gain indicates how interaction influences the steady state behavior of the process, with a value of 1.0 indicating no influence and large deviations from 1.0 indicating strong influence. The array is formed as a matrix of elements,  $\lambda_{ij}$ . The relative gain provides valuable information on the selection of the proper loop pairings; only those loop pairings with relative gains that are positive are usually considered appropriate designs. Systems with negative relative gains are disqualified since the stable multiloop system will be unstable if one or more controllers become inactive [McAvoy, 1983; Morari and Zafiriou, 1989]; this situation is termed *poor integrity*.

Other considerations for multiloop design involve control performance. First, the most important variables should be paired with manipulated variables that have sufficient range—that is, that can be adjusted to compensate for all expected disturbances. Second, all important variables should be paired with manipulated variables that provide fast feedback compensation. Third, the process equipment and operating conditions and the control structures should be selected to attenuate disturbances; this could include tankage before sensitive units and enhancements such as cascade control. Also, inventories (liquid levels and solid inventory) should be controlled in a manner that attenuates flow variation to downstream units. Fourth, inferential measurements should be used to provide real-time control where online sensors of the true controlled variable are not available.



Fifth, the loop pairings should provide the best control performance for the multiloop system; the proper pairing provides not only fast responses for each loop, but also interactions between the loops that improve the performance of the interacting loops. Favorable interaction can occur where one loop, while controlling a variable, introduces adjustments that also tend to reduce the variability of other interacting controlled variables. Both favorable and unfavorable interactions are possible; thus, the proper pairing selection is crucial. Methods are available for analyzing the likely effects of interaction [Skogestad and Morari, 1987; Marlin, 1994].

This brief summary provides some guidance on key aspects of applying automatic control in the process industries. The citations in "References" and "Further Information" provide details essential for successful process control design and application.

## Defining Terms

**Cascade:** A control structure in which one (primary) feedback controller sends its output to the set point of another (secondary) feedback controller.

**CV, MV, SP:** Symbols used for the controlled variable, manipulated variable, and the set point (reference value), respectively.

**Feedforward:** A control approach in which the adjustments to the manipulated variable are based on a measured input to maintain a system output unchanged.

**Interaction:** The situation in a multi-input, multioutput system where the inputs affect more than one output variable.

## References

- Fertik, H. 1975. Tuning controllers for noisy processes. *ISA Trans.* 14(4):292–304.
- Marlin, T. 1994. *Process Control: Designing Processes and Control Systems for Dynamic Performance*. McGraw-Hill, New York.
- McAvoy, T. 1983. *Interaction Analysis*. Instrument Society of America, Research Triangle Park, NC.
- Morari, M. and Zafiriou, E. 1989. *Robust Process Control*. Prentice Hall, Englewood Cliffs, NJ.
- Rosenbrock, H. 1974. *Computer-Aided Control System Design*. Academic Press, New York.
- Skogestad, S. and Morari, M. 1987. Effect of disturbance directions on closed-loop performance. *IEC Res.* 26:2323–2330.
- Ziegler, J. and Nichols, N. 1942. Optimum settings for automatic controllers. *Trans. ASME.* 64:759–768.

## Further Information

Many useful standards for industrial practice in process control are documented in *Standards and Practices for Instrumentation and Control*; 11th edition, 1992, published by the Instrument Society of America.

A crucial contribution of process control is improved plant safety. An introduction to approaches for designing safe process control is presented in *Guidelines for the Safe Automation of Chemical*

*Processes*, 1993, published by the American Institute of Chemical Engineers.

An introduction to advanced methods for empirical model identification is given in Cryor, 1986, *Times Series Analysis*, Duxbury Press.

Examples of the industrial practice of advanced process control are published in meetings organized by the International Federation of Automatic Control (IFAC), Symposia on ADCHEM, DYCORD, and Workshops on the Interaction Between Process Design and Control.

A topic of growing interest in process automation is statistical process control; a good introduction to issues and methods is given in Box and Kramer, 1992, Statistical process monitoring and feedback adjustment—A discussion, *Technometrics*, 34: 251–267, with further discussions on pages 268–285.

An alternative approach for multivariable control involves a centralized control algorithm, which is introduced in Garcia and Morshedi, 1986, Quadratic programming solution of dynamic matrix control (QDMC), *Chem. Eng. Commun.* 46:73–87.

Jacquot, R. G. "Digital Control  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

- 157.1 Feedback Control
- 157.2 Digital Control
- 157.3 Microcontroller Architecture
- 157.4 Linear Digital Control
- 157.5 Digital Control Stability Analysis and Design
- 157.6 Computer-Aided Design

**Raymond G. Jacquot**

*University of Wyoming*

With the decreasing cost of microcomputer systems caused by the development of very large scale integration (VLSI) technology, small, low-cost **digital computer** systems are being incorporated into low-cost automotive and home appliance systems. In the past these systems were routinely incorporated into aerospace, military, and industrial applications where their relative high cost could be tolerated. The growth of the field of digital control systems is best illustrated by the fact that two decades ago there were no computers in automobiles and now automobiles have more than two computers controlling, among other things, fuel injection, antilock brakes, active suspension, and interior climate.

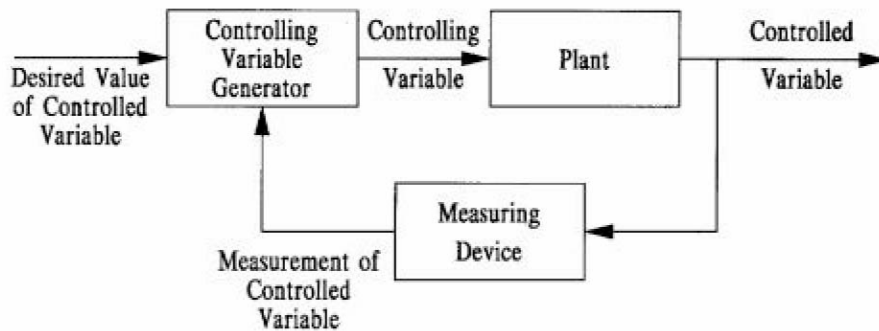
### 157.1 Feedback Control

---

In this section we shall be concerned with the idea of forcing the output of a dynamic system, called the *plant*, to take on a sequence of predetermined values in time. If a complete model of the system to be controlled is known and there are no unknown disturbances, then the input controlling variable may be manipulated to cause the output to take on the correct sequence of values. On the other hand, if the model of the system to be controlled is known only approximately or if there are unknown disturbances acting on the system (which there always are), then the output will deviate significantly from the desired values.

A better way to control such a plant is to make measurements of the output variable that are to be compared to the desired output; then the controlling variable is to be adjusted in such a way as to drive the plant output toward the desired value. This control strategy is shown diagrammatically in [Fig. 157.1](#).

**Figure 157.1** Closed-loop control strategy.



Such a control system is exemplified by a typical home heating system. There may be two controlling variables such that one drives the output one direction, whereas the other drives it in the opposite direction. One must be sure to design the controlling variable generator so that the correct action takes place.

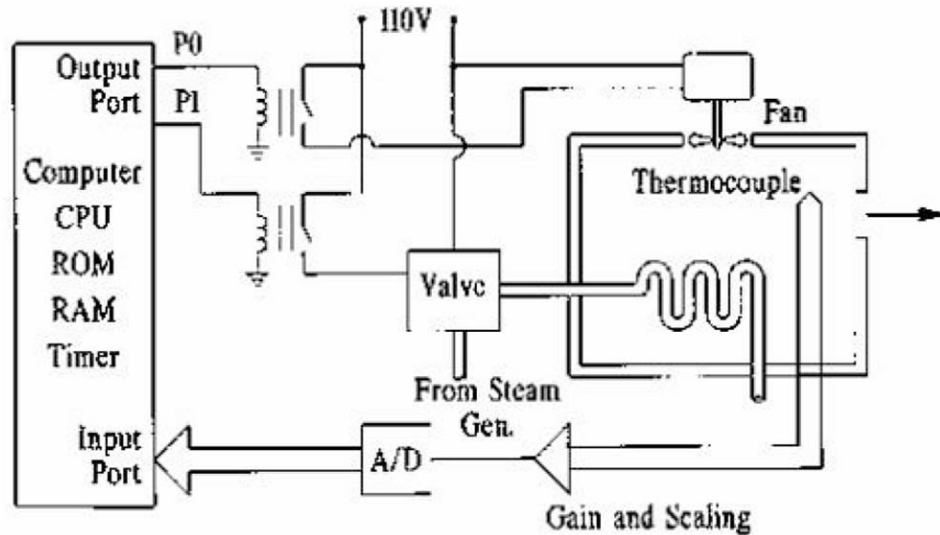
## 157.2 Digital Control

---

The previous closed-loop control scenario is entirely compatible with the incorporation of a digital processor into the control loop. The digital processor is particularly well suited to the task of control because it can process measurements in digital form to produce digital outputs that may be as simple as binary (on/off) signals. The measurement inputs can be in the form of either (1) analog-to-digital converter outputs (which are numerical samples of a signal), or (2) simply a binary signal that the controlled variable has exceeded some threshold value or that it has fallen below some other threshold value. This latter technique is particularly well suited to minimal hardware implementations because the processor can poll the bits of an input port for binary information about the controlled variable. Similarly, an output port can be employed to output simple on-off signals to control actuating elements such as valves, heaters, fan motors, and so on.

A typical application of the strategy just outlined is shown in [Fig. 157.2](#) for the control of the temperature of the interior of an environmental chamber. Disturbances to the system are changes in the exterior temperature from the normal expected temperature. The steam coils provide heat to raise the interior temperature, whereas the fan provides cool exterior air for decreasing the interior temperature.

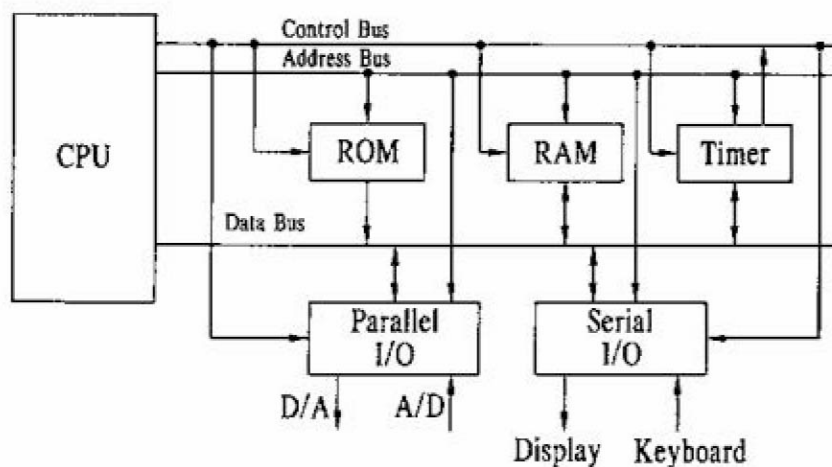
**Figure 157.2** Computer control of environmental chamber temperature.



### 157.3 Microcontroller Architecture

A microcomputer typical of those employed in control applications is illustrated in [Fig. 157.3](#). The computer consists of a central processing unit (CPU), random access memory (RAM), read-only memory (ROM), a timer, and serial and parallel I/O, all connected by three parallel buses that are functionally designated as the data bus, address bus, and control bus.

**Figure 157.3** Architecture of a typical digital control computer. (Source: Jacquot, R. G. 1995. *Modern Digital Control Systems*, 2nd ed. Marcel Dekker, New York. With permission.)



The major manufacturers of microprocessors—namely, Motorola and Intel—produce families of

microcontrollers that have many of the hardware components on board with the CPU, including both ROM and RAM, and in some instances there is also a multichannel analog-to-digital converter for acquisition of sensor data from the process or processes that are to be controlled. These microcontrollers have found wide acceptance in the automotive and appliance industries.

## 157.4 Linear Digital Control

---

Consider the case where the plant with continuous-time dynamics is driven by an output from a digital-to-analog converter (D/A) and the measurements of the plant variables are interfaced to the digital processor with an analog-to-digital converter (A/D). The case for a single controlling variable and a single controlled variable is illustrated in Fig. 157.4. In this case the A/D and D/A are operated synchronously and periodically. The A/D provides numerical values of periodic samples of the output variable  $y(t)$ . One such numerical value is commonly called  $y(kT)$ , where  $k$  denotes an integer such that  $t = kT$  and  $T$  is the sampling interval. The numbers output periodically to the D/A will be termed  $u(kT)$  and called the *control effort*. The error sequence is defined as

$$e(kT) = r(kT) - y(kT) \quad (157.1)$$

where  $r(kT)$  is the sequence of reference values that are the desired values of the output. The samples of the output  $y(kT)$  cannot perfectly track the  $r(kT)$  because of friction, inertia, or capacitance (thermal or fluid) in the plant. Present and past values of the error sequence and past values of the controlling sequence are combined to form the current value of the controlling variable. One way to combine these variables is to use a linear combination of the form

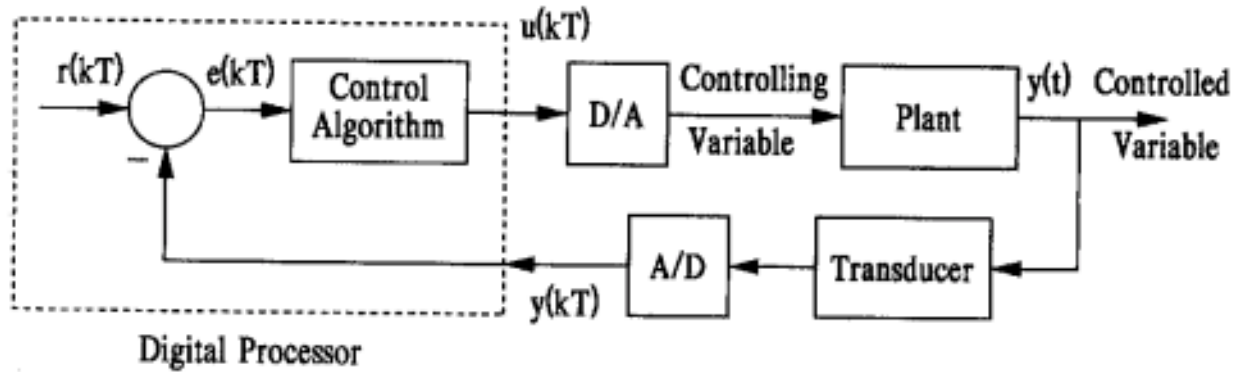
$$\begin{aligned} u(kT) = & b_n e(kT) + b_{n-1} e[(k-1)T] + \cdots + b_o e[(k-n)T] \\ & + a_{n-1} u[(k-1)T] + \cdots + a_o u[(k-n)T] \end{aligned} \quad (157.2)$$

The digital control design task is to determine the values of  $a_i$  and  $b_j$  to give the controlled system desirable performance. The simplest possible version of Eq. (157.2) is to use only the current error on the right side of the equation, which gives a control law:

$$u(kT) = b_n e(kT) \quad (157.3)$$

This control law is called the *proportional control strategy*, wherein the control action taken is proportional to the error—that is, the larger the error is, the larger the corrective action is. This is a sensible control strategy from the point of view that it always faces the plant in the direction to drive the error to zero. One problem with this strategy, however, is that if the constant  $b_n$  is too large the controller may tend to overcorrect and the result will be an unstable closed-loop system. There must therefore be some limitations on  $b_n$  for many systems to remain stable.

**Figure 157.4** Linear digital controller.



It is known in the study of analog control of systems that a controller that incorporates integral action can be very effective in the suppression of sustained errors; however, the integrator will also have a destabilizing effect on the closed-loop system. A compromise that will still control sustained errors and has more favorable stability properties is the proportional plus integral (PI) controller, which in continuous time has the form

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau \quad (157.4)$$

This control law can be approximated in the discrete-time domain by approximating the integral with trapezoidal numerical integration [Jacquot, 1995], resulting in the discrete-time control law of

$$u(kT) = u[(k-1)T] + b_1 e(kT) + b_0 e[(k-1)T] \quad (157.5)$$

where the order of the controller given by Eq. (157.2) has been selected as 1, and the coefficient  $a_0$  has been chosen as unity by virtue of the integral action. The values of  $b_1$  and  $b_0$  are the values to be selected in the course of the design process.

## 157.5 Digital Control Stability Analysis and Design

Thus far, stability and instability have only been discussed in passing; now we shall consider the analytical tools needed for design. The appropriate tool for discrete-time systems is the one-sided  $z$  transform. The  $z$  transform of a discrete-time sequence  $x(k)$  is defined by

$$X(z) = Z[x(k)] = \sum_{k=0}^{\infty} x(k)z^{-k} \quad (157.6)$$

There are many well-developed tables of  $z$  transforms; the reader is referred to one of the popular texts on the subject [Franklin *et al.*, 1990]. For linear plants with linear control laws such as those given by Eqs. (157.2), (157.3), and (157.5), the performance and stability analyses are best carried



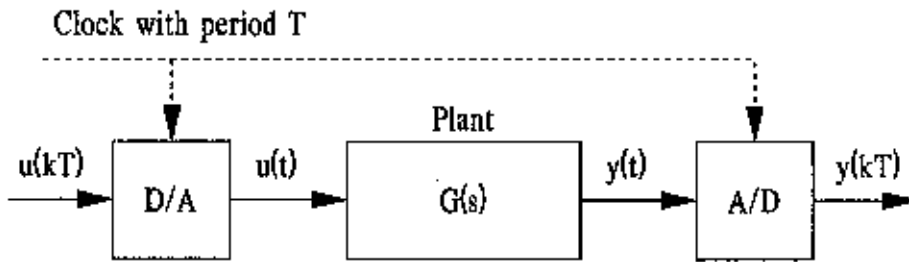
out in the complex  $z$  plane.

If the plant is linear it can be characterized by a transfer function  $G(s)$ . If (1) the plant is driven by a digital-to-analog converter, (2) the output is sampled by an analog-to-digital converter, and (3) the two converters are operated synchronously by a common clock as illustrated in Fig. 157.5, then the plant and two converters comprise, from input to output, a discrete-time system with a transfer function [Jacquot, 1995]

$$G(z) = \left( \frac{z-1}{z} \right) ZL^{-1} \left[ \frac{G(s)}{s} \right] \quad (157.7)$$

where the  $ZL^{-1}$  operations reflect the  $z$  transform of the samples taken from the inverse transform of  $G(s)/s$ .

**Figure 157.5** Linear plant driven by a digital-to-analog converter with sampled output.



The linear feedback control law of Eq. (157.2) can be characterized as a  $z$ -domain transfer function  $D(z)$  of the form

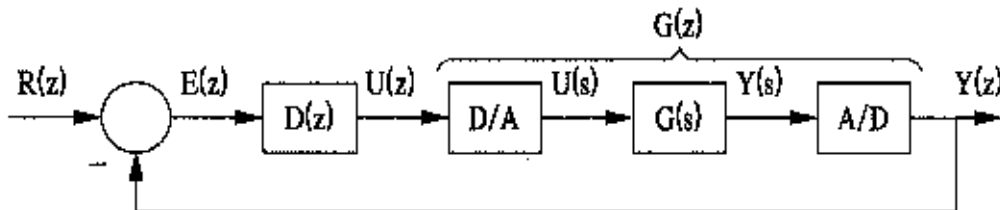
$$D(z) = \frac{U(z)}{E(z)} = \frac{b_n z^n + \cdots + b_1 z + b_0}{z^n - a_{n-1} z^{n-1} - \cdots - a_1 z - a_0} \quad (157.8)$$

and the complete closed-loop system is shown in Fig. 157.6. The closed-loop characteristic equation can be shown to be

$$1 + G(z)D(z) = 0 \quad (157.9)$$

where  $G(z)$  and  $D(z)$  are the transfer functions defined, respectively, in Eqs. (157.7) and (157.8). The criterion for the stability of the closed-loop system is that the roots of Eq. (157.9) [those values of  $z$  that satisfy Eq. (157.9)] lie interior to the unit circle of the complex  $z$  plane.

**Figure 157.6** Complete closed-loop digital control system.



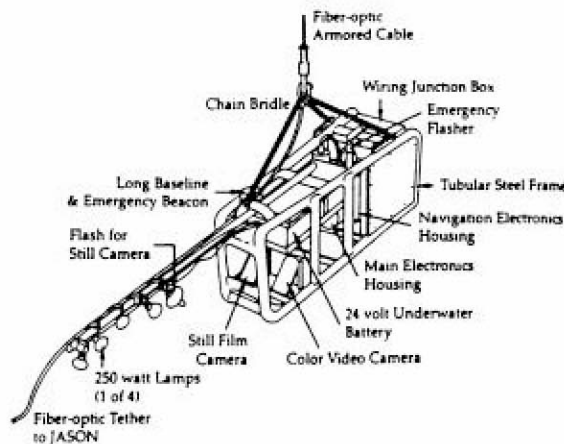
## ROV MEDEA/JASON

D. Mindell,

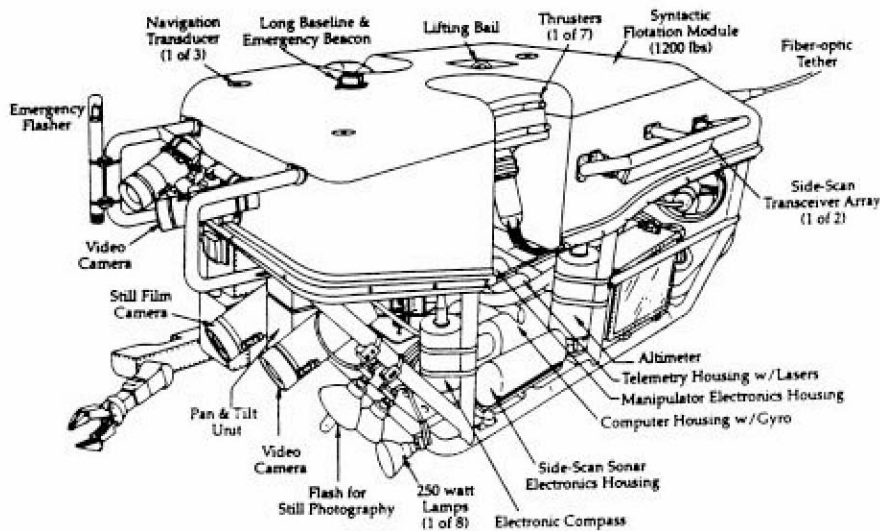
Massachusetts Institute of Technology

*Medea/Jason* is a remotely operated vehicle (ROV) system designed for scientific investigation of the deep ocean floor. It is a dual-vehicle system, with *Medea* serving as a wide area survey vehicle and *Jason* as a precision multisensory imaging and sampling platform. Both *Medea* and *Jason* are designed to operate to a maximum depth of 20000 feet (6000 meters), are transportable, and can be operated from a variety of vessels.

*Medea* is maneuvered primarily by movements of a ship utilizing dynamic positioning. *Jason* is designed for detailed survey and sampling tasks that require a high degree of maneuverability. It weighs about 2200 pounds (1000 kilograms) in air but is nearly neutrally buoyant at depth. The dynamics of *Jason* were designed to make it a very controllable platform. It is propelled by seven DC electric thrusters that provide about 300 newtons (70 pounds) in the vertical direction, 260 newtons (60 pounds) in the forward direction, and about 200 newtons (45 pounds) in the lateral direction. The vehicle has excellent passive stability in pitch and roll. (Courtesy of Woods Hole Oceanographic Institution.)



Medea



Jason

## 157.6 Computer-Aided Design

---

A convenient environment for the design of such systems is that of MATLAB™, in particular employing the Control System Toolbox. That environment is particularly useful for calculation of roots of characteristic polynomials resulting from Eq. (157.9) and the other associated frequency- and time-domain responses. The use of MATLAB has become the accepted computational environment for unified analysis, design and simulation, particularly, using the new graphical interface of SIMULINK. Another excellent simulation environment is that of VISSIM™, which is similar to SIMULINK in that it retains the analog computer device-oriented nature with which many engineers are familiar.

**Example.** Consider the control of the fluid level in the prismatic tank shown in Fig. 157.7. The flow rate  $Q$  into the tank will be controlled by a digital-to-analog converter and the fluid level  $h$  will be sampled periodically with period  $T$  with an analog-to-digital converter. The fluid will be considered to be incompressible, and the principle of conservation of volume yields

$$A \frac{dh}{dt} = -C(h)^{1/2} + Q(t) \quad (157.10)$$

where  $Q(t)$  is the incoming volumetric flow rate,  $A$  is the tank area, and  $h(t)$  is the instantaneous fluid level. Consider perturbations in the fluid level about some equilibrium height  $H$  with steady inlet flow of  $U$  and denote the perturbations in these two variables as  $y(t)$  and  $u(t)$ , respectively. The equation becomes

$$A \frac{dy}{dt} = -K(H + y)^{1/2} + U + u(t) \quad (157.11)$$

The nonlinear term on the right side may be approximated for ( $y \ll H$ ) by the first two terms of a binomial expansion. After dropping the equilibrium flow terms the resulting differential equation in the perturbation variables is

$$A \frac{dy}{dt} = -\frac{C}{2(H)^{1/2}} y + u(t) \quad (157.12)$$

This linear differential equation will have a corresponding transfer function

$$G(s) = \frac{Y(s)}{U(s)} = \frac{1/A}{s + (1/\tau)} \quad (157.13)$$

where  $\tau = 2(H)^{1/2} A/C$  is the system time constant. If the data converters are considered, then the discrete-time transfer function from flow-rate perturbation to sampled height perturbation is given by Eq. (157.7) to be

$$G(z) = \frac{\tau}{A} \left[ \frac{1 - e^{-T/\tau}}{z - e^{-T/\tau}} \right] \quad (157.14)$$

If a proportional controller with gain  $K$  is chosen, then  $D(z) = K(K > 0)$  and the closed-loop characteristic equation from Eq. (157.9) is

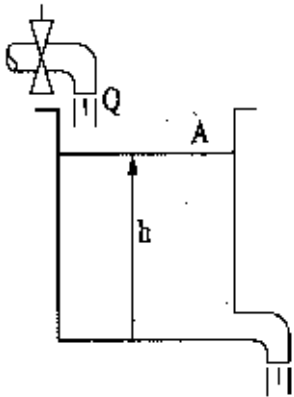
$$1 + \frac{K\tau}{A} \left[ \frac{1 - e^{-T/\tau}}{z - e^{-T/\tau}} \right] = 0 \quad (157.15)$$

The system will be stable so long as the root of Eq. (157.15) lies interior to the unit circle, and hence it is stable if

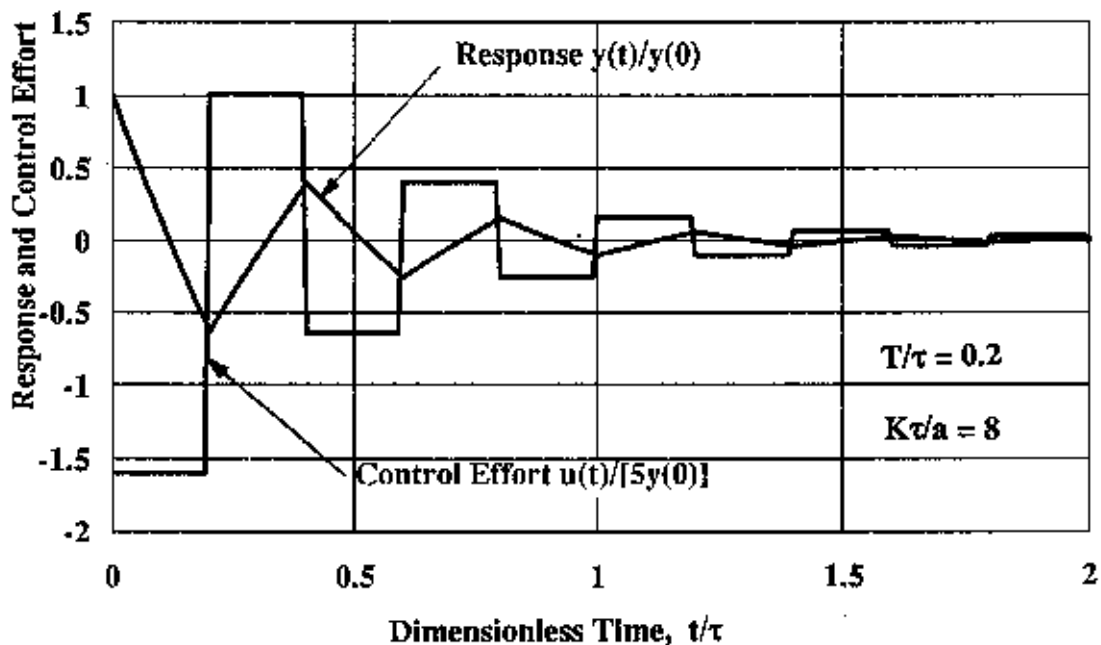
$$\frac{K\tau}{A} < \left[ \frac{1 + e^{-T/\tau}}{1 - e^{-T/\tau}} \right] \quad (157.16)$$

If the sampling period is chosen to be one fifth of the time constant ( $\tau/T = 5$ ) then the limiting value of  $K\tau/A$  is 10.03. If  $K\tau/A$  is chosen as 8 and the initial value of the perturbation height is  $y_o$ , then the response of the system is as shown in Fig. 157.8.

**Figure 157.7** Fluid tank for level control.



**Figure 157.8** Fluid height and input flow-rate perturbations for the fluid level control system.



## Defining Terms

**Digital computer:** A collection of digital devices including an arithmetic logic unit (ALU), read-only memory (ROM), random access memory (RAM), and control and interface hardware.

**Feedback control:** The regulation of a response variable of a system in a desired manner using measurements of that variable in the generation of the strategy of manipulation of the controlling variables.

## References

- Astrom, K. J. and Whittenmark, B. 1984. *Computer Controlled Systems: Theory and Design*. Prentice Hall, Englewood Cliffs, NJ.
- Franklin, G. F., Powell, J. D., and Workman, M. L. 1990. *Digital Control of Dynamic Systems*, 2nd ed. Addison-Wesley, Reading, MA.
- Houpis, C. H. and Lamont, G. B. 1992. *Digital Control Systems: Theory, Hardware, Software*, 2nd ed. McGraw-Hill, New York.
- Jacquot, R. G. 1995. *Modern Digital Control Systems*, 2nd ed. Marcel Dekker, New York.
- Kuo, B. C. 1992. *Digital Control Systems*, 2nd ed. Saunders College, Orlando, FL.
- Phillips, C. L. and Nagle, H. T. 1990. *Digital Control System Analysis and Design*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.

## Further Information

*IEEE Control Systems Magazine* is a useful information source on control systems in general and digital control in particular. Another useful source is the *IEEE Transactions on Control Systems Technology*. Highly technical articles on the state of the art in digital control may be found in the *IEEE Transactions on Automatic Control* and the *ASME Journal of Dynamic Systems, Measurement and Control*.

Bonitz, R. G. "Robots and Control"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

- 158.1 Independent Joint Control
- 158.2 Method of Computed Torque
- 158.3 Cartesian-Space Control

**Robert G. Bonitz**

*University of California, Davis*

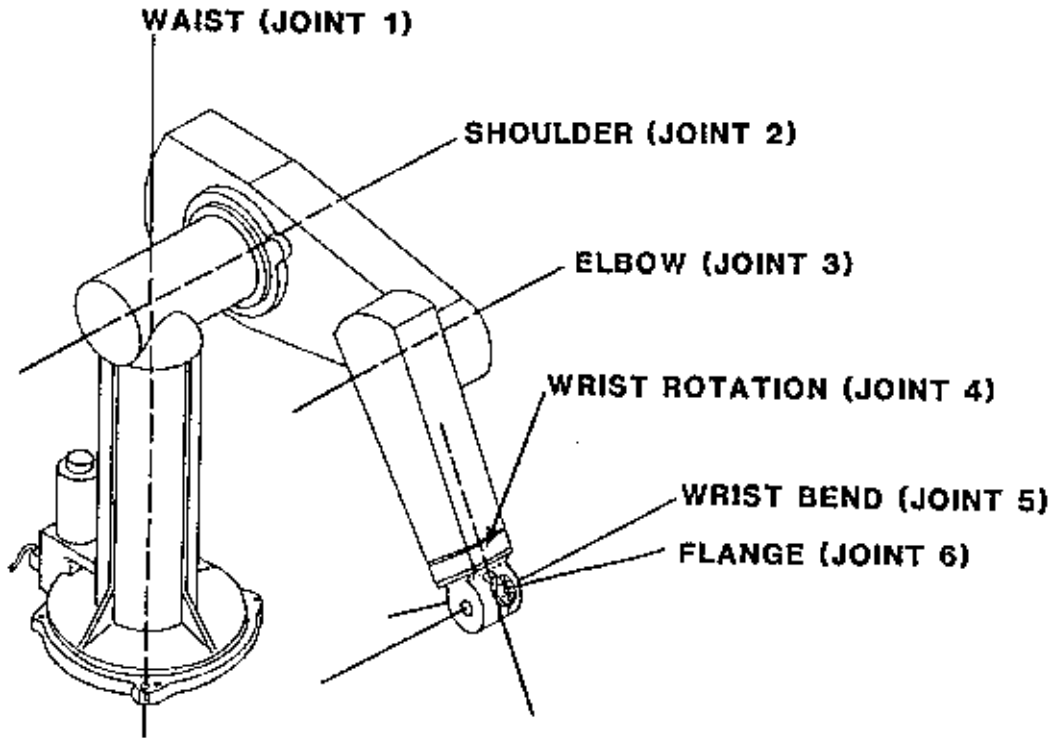
**T. C. Hsia**

*University of California, Davis*

The Robotics Industry Association defines a robot as a "reprogrammable multifunctional manipulator designed to move material, parts, tools, or specialized devices, through variable programmed motions for the the performance of a variety of tasks." A robotic manipulator generally consists of a serial chain of links connected via revolute or prismatic joints and actuated via gear trains or directly by electric motors or hydraulic or pneumatic drives. A robotic system will generally include position sensors (potentiometers, optical encoders, etc.) and may also include contact, tactile, force/torque, proximity, or vision sensors. The manipulator is typically fitted with an end effector, such as a gripper or hand, to enable it to accomplish the desired task. An example of an industrial robot is the PUMA 560 shown in [Fig. 158.1](#). The fundamental control problem in robotics is to determine the required actuator signals such that the desired motion is executed in accordance with specified performance criteria. If the robot is to perform a task while in contact with a surface, it is also necessary to control the contact force applied by the manipulator. Although the control problem is simply stated in the preceding, its solution may be quite complicated, since a robot's dynamics are described by a set of coupled nonlinear differential equations.

**Figure 158.1** PUMA 560 Robot. (*Source: Stäubli Unimation, Inc. 1985. PUMA Mark II Robot 500 Series Equipment Manual, pp. 1–3. Stäubli Unimation, Inc., Sewickly, PA. With permission.*)

**Figure 158.1**



The planning of the manipulator trajectory to achieve the desired motion is integrally linked to the control problem. The position of the robot can be described in joint space by the set of joint coordinates or in **Cartesian or task space** by the position and orientation of the end effector using coordinates along orthogonal axes. The Cartesian position and orientation can be computed from the joint positions via the **forward kinematics** function. The motion required to accomplish the desired task is generally specified in Cartesian space. The determination of the joint positions required to achieve the desired end-effector position and orientation is the **inverse kinematics** problem, which may have more than one solution, and a closed-form solution may not be attainable. The desired motion may be specified as point-to-point, in which the end effector moves from one point to another without regard to the path, or it may be specified as a continuous path, in which the end effector follows a desired path between the points. A trajectory planner generally interpolates the desired path and generates a sequence of set points for the controller. The interpolation may be done in joint or Cartesian space.

Some of the robot control schemes in use today include independent joint control [Luh, 1983]; Cartesian-space control [Luh *et al.*, 1980]; and force control strategies such as hybrid position/force control [Raibert and Craig, 1981] and impedance control [Hogan, 1985]. In independent joint control each joint is considered as a separate system and the coupling effects between the links are treated as disturbances to be rejected by the controller. The performance can be enhanced by compensating for the robot nonlinearities and interlink coupling using the method of computed torque or inverse dynamics. In Cartesian-space control the error signals are computed in Cartesian space and the inverse kinematics problem need not be solved. The position control schemes are covered in the following sections. Information on the force control schemes can be

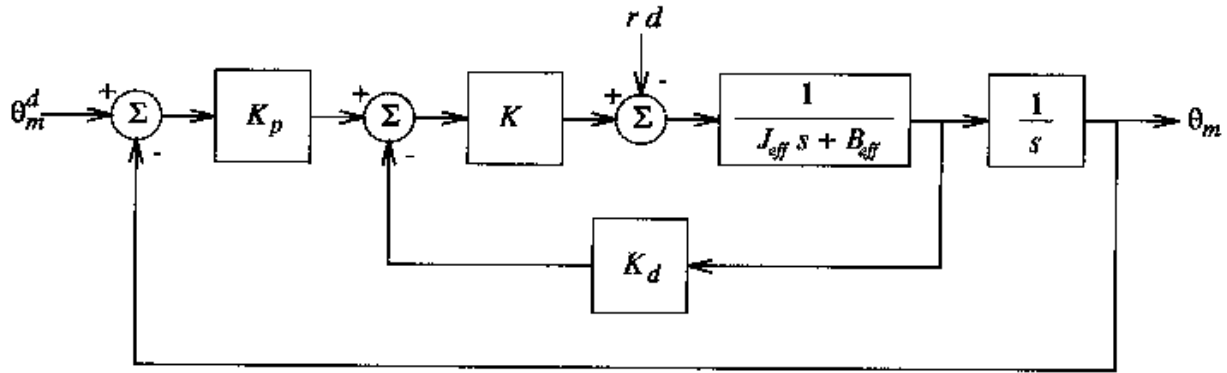


found in the references already cited.

## 158.1 Independent Joint Control

Many industrial robots employ large gear reductions (0.05 to 0.005) that significantly reduce the coupling effects between the links. For slowly varying command inputs, the drive system dominates the dynamics of each joint. Under these conditions each joint can be controlled as an independent system using linear system control techniques. Figure 158.2 depicts a closed-loop system using proportional-derivative (PD) control for a single joint of a robot driven by DC motors.

**Figure 158.2** Closed-loop system for a single robot joint. (Source: Spong, M. W. and Vidyasagar, M. 1989. *Robot Dynamics and Control*, p. 178. John Wiley & Sons, New York. With permission.)



The closed-loop transfer function from input to output is

$$T(s) = \frac{\theta_m(s)}{\theta_m^d(s)} = \frac{K K_p}{J_{\text{eff}} s^2 + (B_{\text{eff}} + K K_d) s + K K_p} \quad (158.1)$$

where  $\theta_m$  is the motor shaft angle, the superscript  $d$  represents a desired quantity,  $K = K_m/R$ ,  $K_m$  is the motor torque constant,  $R$  is the motor armature winding resistance,  $K_p$  is the proportional gain,  $J_{\text{eff}} = J_m + r^2 J_l(\theta)$ ,  $J_m$  is motor and gear inertia,  $r$  is the gear ratio defined by  $r = \theta/\theta_m$ ,  $J_l(\theta)$  is the link inertia,  $\theta$  is the  $n \times 1$  vector of joint positions,  $n$  is the number of links,  $B_{\text{eff}} = B_m + K_b K_m/R$ ,  $B_m$  is the viscous torque constant, and  $K_b$  is the back emf constant. The winding inductance has been ignored because it is frequently small enough such that the time constant of its pole is much smaller than the mechanical time constant.  $J_{\text{eff}}$  depends on the link inertia,  $J_l(\theta)$ , which is dependent on the configuration of the robot links. It is common practice to choose a constant average value for  $J_{\text{eff}}$  considering the variation in link inertia over the robot work space. The range of inertias and gear ratios for the first three links of the PUMA 560 robot are given in Table 158.1. The effect of the gearing is to significantly reduce the contribution of the

link inertia,  $J_l(\theta)$ , to the effective inertia,  $J_{\text{eff}}$ , seen at the joint motor shaft. The characteristic equation for the closed-loop system of Eq. (158.1) is

$$s^2 + \frac{B_{\text{eff}} + K K_d}{J_{\text{eff}}} s + \frac{K K_p}{J_{\text{eff}}} = 0 \quad (158.2)$$

$K_p$  and  $K_d$  can be chosen to achieve the desired damping and natural frequency. It is common practice to choose critical damping in robotic applications to achieve fast response while avoiding overshoot.

**Table 158.1** Inertias and Gear Ratios for PUMA 560 Robot

Link	$J_l(\theta)_{\min}$ (Kg · m <sup>2</sup> )	$J_l(\theta)_{\max}$ (Kg · m <sup>2</sup> )	$J_m$ (Kg · m <sup>2</sup> )	$r$
1	1.43	3.85	0.00029	0.01597
2	1.34	2.82	0.00041	0.00931
3	0.33	0.33	0.00029	0.01863

*Source:* Armstrong, B., Khatib, O., Burdick, J. 1986. The explicit dynamic model and inertial parameters of the PUMA 560 arm. *Proc. IEEE Int. Conf. Robot. Automat.*

The transfer function from the disturbance input to the output is

$$T_d(s) = \frac{\theta(s)}{D(s)} = -\frac{r}{J_{\text{eff}} s^2 + (B_{\text{eff}} + K K_d)s + K K_p} \quad (158.3)$$

where  $d$  is the disturbance input caused by the interlink coupling. For a constant disturbance the steady state error is

$$e_{\text{ss}} = -\frac{r d}{K K_p} \quad (158.4)$$

We see that the error is reduced by the gear reduction and can be made smaller by increasing the proportional gain. For a robot the disturbance in the steady state is due to the effects of gravity on the links. When in motion, the disturbance is dependent on the link velocities and, thus, this control scheme is valid only for slowly time-varying command inputs. In practice there is a limit as to how much  $K_p$  can be increased due to actuator saturation and a necessity to keep the natural frequency well below the structural resonant frequency. An alternative to increasing  $K_p$  to reduce the steady state error is to add integral control to the PD compensator, which reduces the steady state error to zero for constant disturbances.

## 158.2 Method of Computed Torque

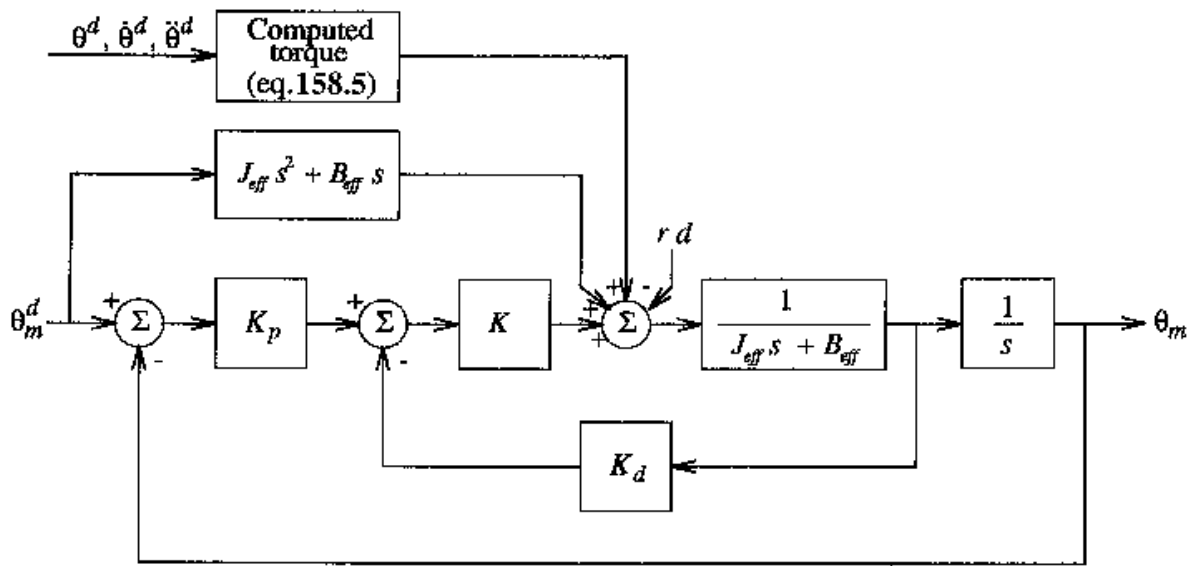
If the robot is a direct-drive type without gear reduction or if the command inputs are not slowly varying, the control scheme of the previous section exhibits poor performance characteristics and

instability may even result. One method of compensating for the effects of the interlink coupling is to use feedforward disturbance cancellation. The disturbance torque is computed from the robot dynamic equation,

$$\tau = D(\theta)\ddot{\theta} + C(\theta, \dot{\theta}) + G(\theta) + F(\dot{\theta}) \quad (158.5)$$

where  $\tau$  is the  $n \times 1$  vector of joint forces/torques,  $D(\theta)$  is the  $n \times n$  inertia matrix,  $C(\theta, \dot{\theta})$  is the  $n \times 1$  vector of Coriolis and centrifugal forces,  $G(\theta)$  is the  $n \times 1$  vector of gravity forces, and  $F(\dot{\theta})$  is the  $n \times 1$  vector of forces due to friction. In the feedforward disturbance cancellation scheme, the right-hand side of Eq. (158.5) is computed using the desired value of the joint variables and injected at the disturbance summing node as shown in Fig. 158.3. If the plant is minimum phase (has no right-half  $s$ -plane zeros), its inverse is also fed forward to achieve tracking of any reference trajectory.

**Figure 158.3** Feedforward computed-torque control.



Another version of computed-torque control, which is known as *inverse dynamics control*, involves setting the control torque to

$$\tau = D(\theta)v + C(\theta, \dot{\theta}) + G(\theta) + F(\dot{\theta}) \quad (158.6)$$

which results in  $\ddot{\theta} = v$ , a double integrator system. The value of  $v$  is now chosen as  $v = \ddot{\theta}^d + K_d(\dot{\theta}^d - \dot{\theta}) + K_p(\theta^d - \theta)$ . This results in the tracking error,  $e = \theta^d - \theta$ , which satisfies

$$\ddot{e} + K_d \dot{e} + K_p e = 0 \quad (158.7)$$

The terms  $K_p$  and  $K_d$  can be chosen for the desired error dynamics (damping and natural frequency). Computed-torque control schemes are computationally expensive due to the complicated nature of Eq. (158.5) and require accurate knowledge of the robot model. They have been successfully implemented using today's powerful and inexpensive computers.

## 158.3 Cartesian-Space Control

The basic concept of Cartesian-space control shown in Fig. 158.4 is that the error signals used in the control algorithm are computed in Cartesian space, obviating the solution of the inverse kinematics. The position and orientation of the robot end effector can be described by a  $3 \times 1$  position vector,  $p$ , and the three orthogonal axes of an imaginary frame attached to the end effector as shown in Fig. 158.5. The axes are known as the normal ( $n$ ), sliding ( $s$ ), and approach ( $a$ ) vectors. The control torque is computed from

$$\tau = D(\theta)\ddot{\theta} + C(\theta, \dot{\theta}) + G(\theta) + F(\dot{\theta}) \quad (158.8)$$

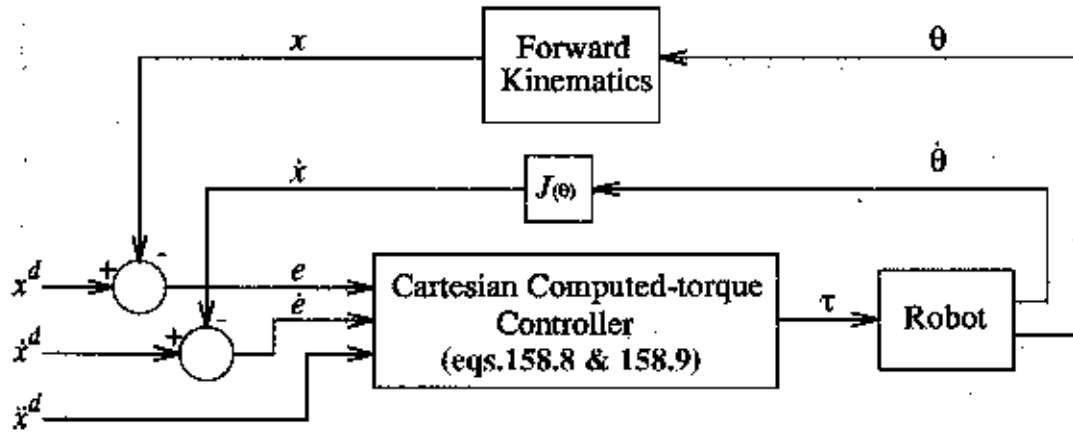
$$\ddot{\theta} = J(\theta)^{-1}[\ddot{x}^d + K_d \dot{e} + K_p e - \dot{J}(\theta, \dot{\theta})\dot{\theta}] \quad (158.9)$$

where  $J(\theta)$  is the manipulator **Jacobian** that maps the joint velocity vector to the Cartesian velocity vector,  $\ddot{x}^d$  is the  $6 \times 1$  desired acceleration vector,  $e = [e_p^T \ e_o^T]^T$ ,  $e_p$  is the  $3 \times 1$  position error vector,  $e_o$  is the  $3 \times 1$  orientation error vector,  $K_d$  is the  $6 \times 6$  positive-definite matrix of velocity gains, and  $K_p$  is the  $6 \times 6$  positive-definite matrix of position gains. The actual position and orientation of the end effector is computed from the joint positions via the forward kinematics. The position error is computed from  $e_p = p^d - p$  and, for small error, the orientation error is computed from  $e_o = \frac{1}{2}[n \times n^d + s \times s^d + a \times a^d]$ . The control law of Eqs. (158.8) and (158.9) results in the Cartesian error equation,

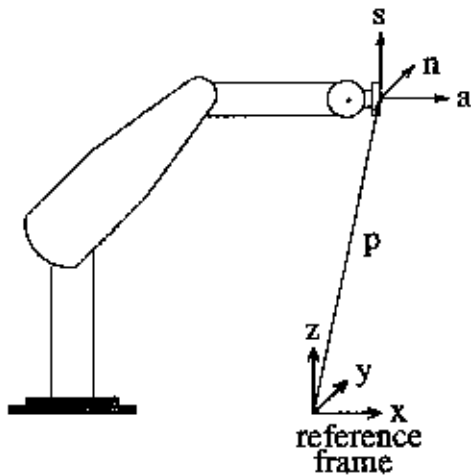
$$\ddot{e} + K_d \dot{e} + K_p e = 0 \quad (158.10)$$

The gain matrices  $K_d$  and  $K_p$  can be chosen to be diagonal to achieve the desired error dynamics along each Cartesian direction.

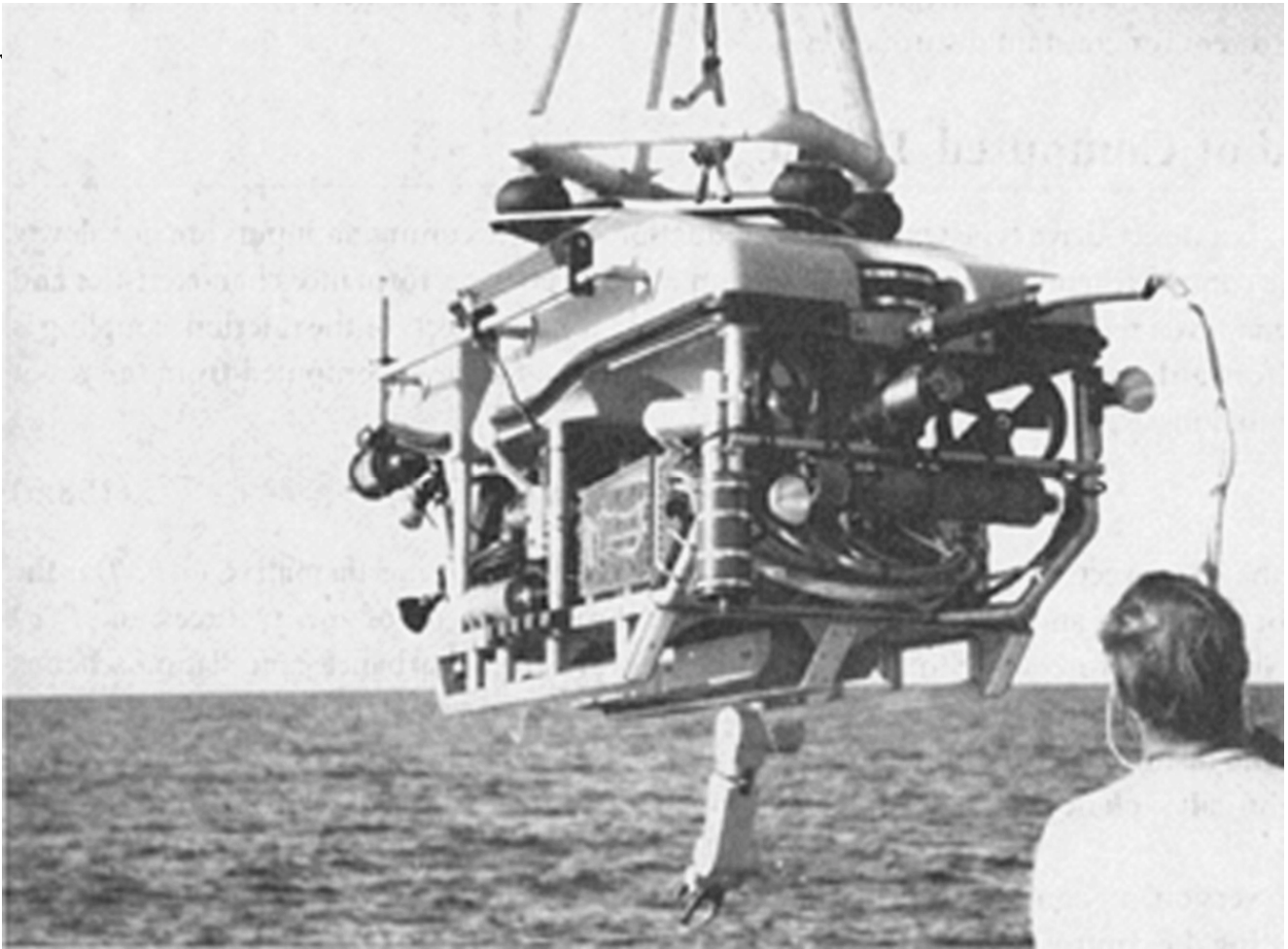
**Figure 158.4** Cartesian computed-torque controller.



**Figure 158.5** End-effector position and orientation.



The Cartesian-space controller has the disadvantage that the inverse of the Jacobian is required, which does not exist at **singular configurations**. The planned trajectory must avoid singularities, or alternative methods such as the SR pseudoinverse [Nakamura, 1991] must be used to compute the Jacobian inverse.



This photograph of JASON was taken in 1993 in Ireland on an expedition to explore the wreck of the Lusitania. JASON has also excavated an ancient shipwreck site 2000 meters down in the Mediterranean (1989) and explored the wrecks of two 1812 warships in Lake Ontario (1990), in addition to many mapping missions for science and the U.S. Navy. (Photo courtesy of D. Mindell.)

#### CONTROL SYSTEM FOR JASON

*D. Mindell, Massachusetts Institute of Technology*

JASON is a remotely operated vehicle (ROV) specifically designed for deep-ocean exploration, built and operated by the Deep Submergence Laboratory of the Woods Hole Oceanographic Institution in Woods Hole, Mass. All of its components are qualified to operate at depths up to 6000 m, which encompasses 95% of the ocean floor. The vehicle has worked that deep, although it typically operates in the 2000–3000 m range. The vehicle connects through a short tether (~50–100 m) to MEDEA. MEDEA, a heavy steel cage, connects through a long, armored cable back to a surface ship, effectively decoupling JASON's tether from ship and cable motion. This arrangement gives the ROV freedom to maneuver and explore. MEDEA also carries some lights and a video camera to enable operators to view JASON from above while working. The cable has three power conductors, for 1800 V at 400 Hz AC transmitted from a power supply on the surface

(originally designed as an auxiliary supply for jet aircraft on the ground). Also in the cable are three single-mode optical fibers, each multiplexed for two channels. Two of the fibers are dedicated to FM video uplink from the vehicle to the surface, providing four video channels, and the other fiber is used for data transmission (500 Mbits/s in both directions).

JASON's control system consists of a network of parallel processors running on the surface ship, connected through the fiber-optic tether to a similar network running on the vehicle itself. On the surface, the control system connects to data logging devices. GPS navigation, the dynamic positioning system of the ship, and the pilot interface. The subsea portion of the control system performs all vehicle functions, and enables the pilot on the surface to teleoperate the vehicle using feedback from video and sensors and to operate JASON's remote manipulator arm. JASON's control system can also operate closed-loop, based on high-precision navigation data. An acoustic navigation system, called EXACT, uses a network of battery-powered transponders on the seabed to locate JASON with an accuracy of 1 cm<sup>3</sup> within a 100 m range. JASON's control system uses these data as feedback to automatically hover or run precision surveys. In conjunction with data from a scanning sonar, for example, precise tracklines produce the data necessary to make 3-D computer models of structures on the seafloor. JASON has performed such surveys on the wreck of the Lusitania, two ships from the war of 1812, and hydrothermal vents ("black smokers") in the Pacific Ocean and the Sea of Cortez.

## Defining Terms

**Cartesian or task space:** The set of vectors describing the position and orientation of the end effector using coordinates along orthogonal axes. The position is specified by a  $3 \times 1$  vector of the coordinates of the end-effector frame origin. The orientation is specified by the  $3 \times 1$  normal, sliding, and approach vectors describing the directions of the orthogonal axes of the frame. Alternately, the orientation may be described by Euler angles, roll/pitch/yaw angles, or axis/angle representation.

**Forward kinematics:** The function that maps the position of the joints to the Cartesian position and orientation of the end effector. It maps the joint space of the manipulator to Cartesian space.

**Inverse kinematics:** The function that maps the Cartesian position and orientation of the end effector to the joint positions. It is generally a one-to-many mapping, and a closed-form solution may not always be possible.

**Jacobian:** The function that maps the joint velocity vector to the Cartesian translational and angular velocity vector of the end effector:  $\dot{x} = J(\theta)\dot{\theta}$ .

**Singular configuration:** A configuration of the manipulator in which the manipulator Jacobian loses full rank. It represents configurations from which certain directions of motion are not possible or when two or more joint axes line up and there is an infinity of solutions to the inverse kinematics problem.

## References

- Armstrong, B., Khatib, O., Burdick, J. 1986. The explicit dynamic model and inertial parameters of the PUMA 560 arm. *Proc. IEEE Int. Conf. Robot. Automat.* pp. 510–518.
- Hogan, N. 1985. Impedance control: An approach to manipulation, Parts I, II, and III. *ASME J. Dyn. Sys. Meas. Control.* 107:1–24.
- Luh, J. Y. S. 1983. Conventional Controller Design for Industrial Robots—A Tutorial. *IEEE Trans. Syst., Man, Cybern.* SMC-13(3): 298–316.
- Luh, J. Y. S., Walker, M. W., and Paul, R. P. 1980. Resolved-acceleration control of mechanical manipulators. *IEEE Trans. Automat. Control.* AC-25(3):468–474
- Nakamura, Y. 1991. *Advanced Robotics, Redundancy, and Optimization*. Addison-Wesley, Reading, MA.
- Raibert, M. H. and Craig, J. J. 1981. Hybrid position/force control of manipulators. *ASME J. Dyn. Syst., Meas., Control.* 103:126–133.
- Spong, M. W. and Vidyasagar, M. 1989. *Robot Dynamics and Control*. John Wiley & Sons, New York.
- Stäubli Unimation, Inc. 1985. *PUMA Mark II Robot 500 Series Equipment Manual*. Stäubli Unimation, Inc., Sewickly, PA.

## Further Information

- IEEE Transactions on Robotics and Automation*, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854-4150. Phone: (800) 678-IEEE.
- Robotics Today*, Society of Manufacturing Engineers, One SME Drive, P.O. Box 930, Dearborn, MI 48121. Phone: (313) 271-1500.
- Robotics Industry Association (RIA), 900 Victors Way, P.O. Box 3724, Ann Arbor, MI 48106.
- Craig, J. J. 1989. *Introduction to Robotics*, 2nd ed. Addison-Wesley, Reading, MA.
- Lewis, F. L., Abdallah, C. T., and Dawson, D. M. 1993, *Control of Robot Manipulators*. Macmillan, New York.
- Murray, R. M., Li, Z., and Sastry, S. S. 1994. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, Boca Raton, FL.
- USENET newsgroups comp.robotics.misc and comp.robotics.research.



Vincent, T. L. "State Variable Feedback"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## State Variable Feedback

---

159.1 Linear State Space Control Systems

159.2 Controllability and Observability

159.3 Eigenvalue Placement

159.4 Observer Design

**Thomas L. Vincent**

*University of Arizona*

The fundamental concept in control system design deals with creating a **feedback loop** between the **output** and the **input** of a dynamical system. This is done to improve the stability characteristics of the system. State variable feedback is associated with the idea of using every **state variable** in the feedback loop. State variable feedback is used in both nonlinear and linear systems. However, it is only with linear systems that a relatively complete theory has been worked out, with some of the results given in this chapter. The advantage of using state variable feedback for linear control system design is in its simplicity. However, in order to use state variable feedback, every state variable must be either measured or estimated. Since measuring every state variable is impractical for most control applications, a state estimator must usually be included as a part of the total state variable feedback control system.

### 159.1 Linear State Space Control Systems

---

A large class of control systems are either given by or approximated by a linear state space model of the form

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \quad (159.1)$$

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u} \quad (159.2)$$

where the dot denotes differentiation with respect to time,  $\mathbf{x}$  is an  $N_x \times 1$  dimensional state vector,  $\mathbf{u}$  is an  $N_u \times 1$  dimensional control vector,  $\mathbf{y}$  is an  $N_y \times 1$  dimensional output vector,  $\mathbf{A}$  is a square  $N_x \times N_x$  matrix,  $\mathbf{B}$  is  $N_x \times N_u$ ,  $\mathbf{C}$  is  $N_y \times N_x$ , and  $\mathbf{D}$  is  $N_y \times N_u$ . The state vector

$$\mathbf{x} = [x_1 \cdots x_{N_x}]^T \quad (159.3)$$

where  $()^T$  denotes transpose, is defined by the solution to the differential equation (159.1). The

control vector

$$\mathbf{u} = [u_1 \cdots u_{N_u}]^T \quad (159.4)$$

represents all control inputs to the system. Control inputs include both **open-loop command inputs** external to the system and **closed-loop control inputs** internal to the system. Control is ultimately transmitted to the system through the use of actuators that may or may not be modeled in Eq. (159.1). It follows from Eq. (159.1) that the state of the system cannot be determined until all inputs, along with initial conditions for the state, have been specified. The output vector

$$\mathbf{y} = [y_1 \cdots y_{N_y}]^T \quad (159.5)$$

represents all of the system quantities measured by means of sensors. The sensor output may be of the general form of Eq. (159.2), where it is a function of both the state and control inputs (e.g., accelerometer) or it may be simply one or more of the state variables (e.g., position sensor, tachometer). The control system design problem is to determine an automatic control algorithm such that the relationship between a command input  $r(t)$  and the output  $y(t)$  yields acceptable performance in terms of tracking, stability, uncertain inputs, and so forth. These qualitative performance criteria are determined by the controlled system eigenvalues.

## 159.2 Controllability and Observability

---

The concepts of controllability and observability are fundamental to state space design. If a system is controllable, then it will always be possible, using state variable feedback, to design a stable controlled system. In fact, the controlled systems eigenvalues may be arbitrarily placed. If a system is observable, then it will always be possible to design an estimator for the state.

The system in Eq. (159.1) is controllable if for every initial state  $\mathbf{x}(0)$  there exists a control  $\mathbf{u}(t), t \in [0, T]$ , where  $T$  is some finite time interval, that will drive the system to any other point in state space. The system in Eqs. (159.1) and (159.2) is observable if, given a control  $\mathbf{u}(t), t \in [0, T]$ , the initial state  $\mathbf{x}(0)$  can be determined from the observation history  $\mathbf{y}(t), t \in [0, T]$ .

Controllability and observability for linear systems in the form of Eqs. (159.1) and (159.2) may be checked using the Kalman controllability and observability criteria. The Kalman controllability criterion is that the system in Eq. (159.1) is controllable if and only if

$$\text{rank} [\mathbf{P}] = N_x \quad (159.6)$$

where  $\mathbf{P}$  is the controllability matrix

$$\mathbf{P} = [\mathbf{B}, \mathbf{AB}, \cdots, \mathbf{A}^{N_x-1} \mathbf{B}] \quad (159.7)$$

The Kalman observability criterion is that the system in Eqs. (159.1) and (159.2) is observable if

and only if

$$\text{rank } [\mathbf{Q}] = N_x \quad (159.8)$$

where  $\mathbf{Q}$  is the observability matrix

$$\mathbf{Q} = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \mathbf{CA}^2 \\ \vdots \\ \mathbf{CA}^{N_x-1} \end{bmatrix} \quad (159.9)$$

State variable feedback design requires that both the controllability and observability criteria be satisfied, which will be assumed in what follows. State space design may be broken into two parts, eigenvalue placement, assuming full state information is available, and observer design to supply any missing state information.

### 159.3 Eigenvalue Placement

---

Assume that every component of the state  $\mathbf{x}$  is measured—that is,  $\mathbf{C} = \mathbf{I}$  and  $\mathbf{D} = 0$  so that  $\mathbf{y} = \mathbf{x}$ . Given that the desired controlled system eigenvalues are known, the design problem is reduced to finding an  $N_u \times N_x$  feedback gain matrix  $\mathbf{K}$  such that Eq. (159.1) under state variable feedback of the form

$$\mathbf{u} = \mathbf{F}\mathbf{r}(t) - \mathbf{K}\mathbf{x}(t) \quad (159.10)$$

will have the prescribed eigenvalues. Here,  $\mathbf{r}(t)$  is an  $N_x \times 1$  vector of command inputs and  $\mathbf{F}$  is an  $N_x \times N_r$  input matrix. The  $\mathbf{F}$  matrix will in general be required, since the dimension of the command inputs may not equal the dimension of the control vector. It also provides for scaling of the command input. Substituting Eq. (159.10) into Eq. (159.1) yields

$$\dot{\mathbf{x}} = \hat{\mathbf{A}}\mathbf{x} + \hat{\mathbf{B}}\mathbf{r} \quad (159.11)$$

where

$$\hat{\mathbf{A}} = \mathbf{A} - \mathbf{BK} \quad (159.12)$$

$$\hat{\mathbf{B}} = \mathbf{BF} \quad (159.13)$$

The controlled system of Eq. (159.11) is of the same form as the original system in Eq. (159.1), except the matrix  $\hat{\mathbf{A}}$  now depends on the matrix of feedback gains  $\mathbf{K}$ , and the control input,  $\mathbf{u}$ , is now the command input  $\mathbf{r}$ . For a constant command input the equilibrium solutions to Eq. (159.11) depend on both  $\mathbf{F}$  and  $\mathbf{K}$ . However,  $\mathbf{K}$  alone affects the eigenvalues of the controlled system. If Eq.

(159.1) is completely controllable, then any desired set of eigenvalues for  $\bar{A}$  can be obtained through an appropriate choice for  $\mathbf{K}$  [Davison, 1968; Wonham, 1967].

For single-input systems ( $u$  is a scalar) of low dimensions, determining  $\mathbf{K}$  for direct eigenvalue placement is easy to do. For example, consider a second-order system of the form of Eq. (159.1) with

$$A = \begin{bmatrix} 0 & 1 \\ a_1 & a_2 \end{bmatrix} \quad B = \begin{bmatrix} 0 \\ b \end{bmatrix} \quad (159.14)$$

This system satisfies Eq. (159.6) so that state space design is possible. Under state variable feedback,

$$\bar{A} = \begin{bmatrix} 0 & 1 \\ a_1 - bk_1 & a_2 - bk_2 \end{bmatrix} \quad (159.15)$$

which has the characteristic equation

$$\lambda^2 + (bk_2 - a_2)\lambda + (bk_1 - a_1) = 0 \quad (159.16)$$

This system can now be made to behave like a second-order system with eigenvalues  $\bar{\lambda}_1$  and  $\bar{\lambda}_2$  satisfying the characteristic equation

$$(\lambda - \bar{\lambda}_1)(\lambda - \bar{\lambda}_2) = \lambda^2 + 2\zeta\omega_n\lambda + \omega_n^2 = 0 \quad (159.17)$$

by matching coefficients to yield

$$k_1 = (a_1 + \omega_n^2)/b; \quad k_2 = (a_2 + 2\zeta\omega_n)/b \quad (159.18)$$

This procedure is also applicable when  $\mathbf{u}$  is a vector, but it is complicated by the fact that the gain matrix  $\mathbf{K}$  is not unique. However, a general procedure to handle this situation is available [Wonham, 1985]. The procedure can also be readily adapted for numerical solution so that higher dimensional problems can be handled as well.

Unless one is certain about what eigenvalues to use, the question of where to place the controlled system eigenvalues remains a fundamental design problem. There are methods available for handling this aspect of state variable feedback design as well. Some of the more commonly used methods are optimal (LQ) design [Bryson and Ho, 1975] and robust control design [Green and Limebeer, 1994]. These methods focus on other system performance criteria for determining the feedback gain matrix  $\mathbf{K}$ , which, in turn, indirectly determine the controlled system eigenvalues. Software is available, based on these methods, for doing control design on a PC [Shahian and Hassul, 1993].

## 159.4 Observer Design

---

Even if every component of the state vector  $\mathbf{x}$  is not directly measured, it is possible to build a device called an *observer* [Luenberger, 1971] that will approximate  $\mathbf{x}$ . An observer that reconstructs the entire state vector is called an *identity observer*. One that only reconstructs states not directly measured is called a *reduced-order observer*. Identity observers are of a simpler form.

An observer uses the system model with an additional output feedback term to generate an estimate for the state. If we let  $\hat{\mathbf{x}}$  denote the estimate of  $\mathbf{x}$ , then an identity observer is obtained from the equations

$$\dot{\hat{\mathbf{x}}} = \mathbf{A}\hat{\mathbf{x}} + \mathbf{B}u - \mathbf{G}[\hat{\mathbf{y}} - y] \quad (159.19)$$

where  $\hat{\mathbf{x}}(0) = \mathbf{0}$ ,  $\mathbf{G}$  is an  $N_x \times N_y$  matrix to be determined, and

$$\hat{\mathbf{y}} = \mathbf{C}\hat{\mathbf{x}} + \mathbf{D}u \quad (159.20)$$

is the predicted measurement in accordance with Eq. (159.2). The structure of the state estimator from Eqs. (159.19) and (159.20) is the same as a Kalman filter [Kalman, 1960], which is used for state estimation in the presence of Gaussian random noise inputs.

Substituting (159.20) into (159.19) and rearranging terms yields the identity observer

$$\dot{\hat{\mathbf{x}}} = [\mathbf{A} - \mathbf{GC}]\hat{\mathbf{x}} + [\mathbf{B} - \mathbf{GD}]u + \mathbf{G}y \quad (159.21)$$

The components of the  $\mathbf{G}$  matrix are chosen so that the error equation

$$\dot{\mathbf{e}} = \hat{\mathbf{x}} - \mathbf{x} = [\mathbf{A} - \mathbf{GC}]\mathbf{e} \quad (159.22)$$

will be stable. The **return time** of the error equation should be chosen so that it is faster than the return time of the controlled system. The stability properties of the error equation are completely under the designer's control, provided that the system in Eqs. (159.1) and (159.2) is observable. For controllable and observable linear systems, observers do not change the closed-loop eigenvalues of the controlled system. Rather, they simply adjoin their own eigenvalues. That is, the eigenvalues for the controlled system using an identity observer are those associated with  $[\mathbf{A} - \mathbf{BK}]$  and  $[\mathbf{A} - \mathbf{GC}]$ , respectively.

As an example, consider the second-order system given by Eq. (159.14), in which the output is the first state variable  $x_1$ , that is,

$$\mathbf{C} = [1 \quad 0] \quad \mathbf{D} = 0 \quad (159.23)$$

The observability criterion of Eq. (159.8) is satisfied by this system so that an observer can be built. In this case

$$A - GC = \begin{bmatrix} -g_1 & 1 \\ a_1 - g_2 & a_2 \end{bmatrix} \quad (159.24)$$

which has the characteristic equation

$$\lambda^2 + (g_1 - a_2)\lambda + (g_2 - a_2g_1 - a_1) = 0 \quad (159.25)$$

By choosing the constants  $g_1$  and  $g_2$ , the observer can now be made to behave like a second-order system with eigenvalues  $\hat{\lambda}_1$  and  $\hat{\lambda}_2$  satisfying the characteristic equation

$$(\lambda - \hat{\lambda}_1)(\lambda - \hat{\lambda}_2) = \lambda^2 + 2\hat{\zeta}\hat{\omega}_n\lambda + \hat{\omega}_n^2 = 0 \quad (159.26)$$

Equating coefficients between Eqs. (159.25) and (159.26) yields

$$g_1 = a_2 + 2\hat{\zeta}\hat{\omega}_n; \quad g_2 = a_1 + a_2g_1 + \hat{\omega}_n^2 \quad (159.27)$$

The observer in this case is of the form

$$\dot{\hat{x}}_1 = \hat{x}_2 + g_1(y - \hat{x}_1) \quad (159.28)$$

$$\dot{\hat{x}}_2 = a_1\hat{x}_1 + a_2\hat{x}_2 + g_2(y - \hat{x}_1) + bu \quad (159.29)$$

These equations must be solved on-line to yield  $\hat{x}_1$  and  $\hat{x}_2$ . The resulting controller is

$$u = r - k_1\hat{x}_1 - k_2\hat{x}_2 \quad (159.30)$$

In this case the actual measurement  $y = x_1$  could be used in Eq. (159.30) instead of  $\hat{x}_1$ . However, if the measurement of  $y$  contains noise, this noise can be filtered through the observer by using the full state estimate  $\hat{x}$  as used in Eq. (159.30).

The observer differential equations may be solved on-line by either digital or analog methods. The initial conditions for solving these equations will generally not be known. However, in most situations, this should not present difficulties with a properly designed observer. Any error resulting from setting all the initial conditions equal to zero will rapidly tend to zero according to the error equation.

## Defining Terms

**Closed-loop control inputs:** Inputs to a control system, given as a function of the output, that are determined by an automatic control algorithm within the system (usually computed by an analog or digital device).

**Feedback loop:** Any connection between the input and the output of a dynamical system.

**Input:** The inputs to a dynamical system are quantities that can affect the evolution of the state of

the system.

**Open-loop command inputs:** Inputs to a control system, given as a function of time, that are initiated by a human operator or some other external device used to specify a desired output.

**Output:** Those functions of the state and control, possibly the states themselves, that can be measured.

**Return time:** For asymptotically stable systems, with eigenvalues  $\lambda_i$ , the return time is defined by  $T = 1/\min |\operatorname{Re}(\lambda_i)|$ ,  $i = 1, \dots, N_x$  and the eigenvalue(s) corresponding to  $T$  is called the *dominant eigenvalue(s)*. The dominate eigenvalue(s) corresponds to the slowest time constant in the system.

**State variables:** Those variables that identify the state of a system. The state of a dynamical system are those dynamical components of a system that completely identify it at any moment in time.

## References

- Bryson, A. E. and Ho, Y. C. 1975. *Applied Optimal Control*. Halsted, New York.
- Davison, E. J. 1968. On pole assignment on multivariable linear systems. *IEEE Trans. Automatic Control*. AC-13(6):747–748.
- Green, M. and Limebeer, D. 1994. *Linear Robust Control*. Prentice Hall, Englewood Cliffs, NJ.
- Kalman, R. E. 1960. A new approach to linear filtering and prediction problems. *Trans. ASME, J. Basic Eng.* 82(1):34–45.
- Luenberger, D. G. 1971. An introduction to observers. *IEEE Trans. Automatic Control*. AC-16(6):596–602.
- Shahian, B. and Hassul, M. 1993. *Control System Design Using MATLAB*. Prentice-Hall, Englewood Cliffs, NJ.
- Wonham, W. M. 1967. On pole assignment in multi-input controllable linear systems. *IEEE Trans. Auto. Control*. AC-12(6):660–665.
- Wonham, W. M. 1985. *Linear Multivariable Control*, 3rd ed. Springer-Verlag, New York.

## Further Information

- Brogan, W. L. 1985. *Modern Control Theory*. Prentice-Hall, Englewood Cliffs, NJ.
- Dorf, R. C. 1995. *Modern Control Systems*. Addison-Wesley, Reading, MA.
- Grantham, W. J. and Vincent, T. L. 1993. *Modern Control Systems Analysis and Design*. John Wiley & Sons, New York.
- Kalman, R. E., Ho, Y. C., and Narendra, K. S. 1963. Controllability of linear dynamical systems. *Contrib. to Differential Equations*. 1(2):189–213.
- Kuo, B. C. 1991. *Automatic Control Systems*. Prentice Hall, Englewood Cliffs, NJ.



Kreith, F. "Manufacturing"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



Roseville, California is the site of one of many Hewlett-Packard manufacturing plants. This "touch-up line" at the Roseville site is where printed circuit boards are finalized before becoming part of the newly manufactured HP 9000 line of computer systems. This manufacturing process line assures the quality of each Hewlett-Packard printed circuit board prior to system completion. (Photo courtesy of Hewlett-Packard.)

# XXV

## Manufacturing

---

**Frank Kreith**

*University of Colorado*

**160 Types of Manufacturing** *R. J. Schonberger*

Job-Shop and Batch Production • Mass Production • Continuous Production • Mixtures and Gray Areas • Capital Investment, Automation, Advanced Technology, Skills, and Layout

**161 Quality** *M. P. Stephens and J. F. Kmec*

Measurement • Statistical Quality Control • Tolerances and Capability

**162 Flexible Manufacturing** *A. Kusiak and C.-X. Feng*

Flexible Machining • Flexible Assembly • The Economic Justification of Flexibility

**163 Management and Scheduling** *E. M. Knod, Jr.*

Management: Definition and Innovations • Scheduling

**164 Design, Modeling, and Prototyping** *W. L. Chapman and A. T. Bahill*

The System Design Process • Rapid Prototyping • When to Use Modeling and Prototyping

**165 Materials Processing and Manufacturing Methods** *S. H. Risbud*

Processing Metals and Alloys • Ceramics, Glasses, and Polymers • Joining of Materials

**166 Machine Tools and Processes** *Y. C. Shin*

Economic Impact • Types of Machine Tools • Control of Machine Tools • Machine Tool Accuracy

**167 Human Factors and Ergonomics** *W. Karwowski and B. Jamaldin*

The Concept of Human-Machine Systems • Ergonomics in Industry • The Role of Ergonomics in Prevention of Occupational Musculoskeletal Injury • Fitting the Work Environment to the Workers

**168 Pressure and Vacuum** *P. Biltoft, C. Borzileri, D. Holten, and M. Traini*

Pressure • The Vacuum Environment

**169 Food Engineering** *R. P. Singh*

Liquid Transport Systems • Heat Transfer

**170 Agricultural Engineering** *D. J. Hills*

Equipment Sizing Criteria • Equipment Selection

**171 System Reliability** *R. Ramakumar*

Catastrophic Failure Models • The Bathtub Curve • Mean Time to Failure • Average Failure Rate • A Posteriori Failure Probability • Units for Failure Rates • Application of the Binomial Distribution • Application of the Poisson Distribution • The Exponential Distribution • The Weibull Distribution • Combinatorial Aspects • Modeling Maintenance • Markov Models • Binary Model for a Repairable Component • Two Dissimilar Repairable Components • Two Identical Repairable Components • Frequency and Duration Techniques • Applications of Markov Process • Some Useful Approximations

DURING THE PAST DECADE, competition on a worldwide scale has intensified, with every country attempting to increase its share in the world economy. As a result, the manufacturing

processes have to adapt frequently to changing market demands for products, and innovations in design have to be accommodated. Competition also requires improved quality, cost reduction, improved management efficiency, and flexibility. All of these requirements are part and parcel of modern engineering.

It is commonly accepted engineering wisdom that new ideas are useless if they cannot be implemented, and outstanding engineering designs are worthless if they cannot be produced. The engineer responsible for conceiving and implementing production methods and for creating useful products from ideas and designs is the key to success in competition. The manufacturing engineer must work with severe constraints in terms of the design, the cost, and the time available to create a new product. This section covers the main ingredients required in the field of manufacturing, including quality control, flexible manufacturing, management, design, materials processing, and ergonomics. In addition, the section covers topics such as agricultural engineering and reliability. These are enormously broad and important topics, and the coverage in a handbook is necessarily limited. However, the reader will find the material useful in gaining an understanding of the various elements of manufacturing and can consult the references at the end of the chapters for more detailed and in-depth information.

Schonberger, R. J. "Types of Manufacturing"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Types of Manufacturing

160.1 Job-Shop and Batch Production

160.2 Mass Production

160.3 Continuous Production

160.4 Mixtures and Gray Areas

160.5 Capital Investment, Automation, Advanced Technology, Skills, and Layout

**Richard J. Schonberger**

*University of Washington*

Although there are many ways to categorize manufacturing, three general categories stand out. These three (which probably have emerged from production planning and control lines of thought) are:

1. **Job-shop production.** A job shop produces in small lots or batches.
2. **Mass production.** Mass production involves machines or assembly lines that manufacture discrete units repetitively.
3. **Continuous production.** The **process industries** produce in a continuous flow.

Primary differences among the three types center on output volume and variety and process flexibility. [Table 160.1](#) matches these characteristics with the types of manufacturing and gives examples of each type. The following discussion begins by elaborating on [Table 160.1](#). Next are comments on hybrid and uncertain types of manufacturing. Finally, five secondary characteristics of the three manufacturing types are presented.

**Table 160.1** Types of Manufacturing—Characteristics and Examples

Volume	Very low	High	Highest
Variety	Highest	Low	Lowest
Flexibility	Highest	Low	Lowest
Job-shop production	Tool and die making Casting (foundry) Baking (bakery)		
Mass production	Auto assembly Bottling Apparel manufacturing		
Continuous production	Paper milling Refining Extrusion		

## 160.1 Job-Shop and Batch Production

---

As [Table 160.1](#) shows, job-shop manufacturing is very low in volume but is highest in output variety and process flexibility. In this mode, the processes—a set of resources including labor and equipment—are reset intermittently to make a variety of products. (Product variety requires flexibility to frequently reset the process.)

In tool and die making, the first example, the volume is generally one unit—for example, a single die set or mold. Since every job is different, output variety is at a maximum, and operators continually reset the equipment for the next job.

Casting in a foundry has the same characteristics, except that the volume is sometimes more than one. That is, a given job order may be to cast one, five, ten, or more pieces. The multipiece jobs are sometimes called lots or **batches**.

A bakery makes a variety of products, each requiring a new series of steps to set up the process—for example, mixing and baking a batch of sourdough bread, followed by a batch of cinnamon rolls.

## 160.2 Mass Production

---

Second in [Table 160.1](#) is mass production. Output volume, in discrete units, is high. Product variety is low, entailing low flexibility to reset the process.

Mass production of automobiles is an example. A typical automobile plant will assemble two or three hundred thousand cars a year. In some plants just one model is made per assembly line; variety is low (except for option packages). In other plants assembly lines produce mixed models. Still, this is considered mass production since assembly continues without interruption for model changes.

In bottling, volumes are much higher, sometimes in the millions per year. Changing from one bottled product to another requires a line stoppage, but between **changeovers** production volumes are high (e.g., thousands). Flexibility, such as changing from small to large bottles, is low; more commonly, large and small bottles are filled on different lines.

Similarly, mass production of apparel can employ production lines, with stoppages for pattern changes. More conventionally, the industry has used a very different version of mass production: Cutters, sewers, and others in separate departments each work independently, and material handlers move components from department to department to completion. Thus, existence of an assembly line or production line is not a necessary characteristic of mass production.

## 160.3 Continuous Production

---

Products that flow—liquids, gases, powders, grains, slurries—are continuously produced, the third type in [Table 160.1](#). In continuous process plants, product volumes are very high (relative to, for



example, a job-shop method of making the same product). Because of designed-in process limitations (pumps, pipes, valves, etc.) product variety and process flexibility are very low.

In a paper mill a meshed belt begins pulp on its journey through a high-speed multistage paper-making machine. The last stage puts the paper on reels holding thousands of linear meters. Since a major product changeover can take hours, plants often limit themselves to incremental product changes. Special-purpose equipment design also poses limitations. For example, a tissue machine cannot produce newsprint, and a newsprint machine cannot produce stationery. Thus, in paper making, flexibility and product variety for a given machine are very low.

Whereas a paper mill produces a solid product, a refinery keeps the substance in a liquid (or sometimes gaseous) state. Continuous refining of fats, for example, involves centrifuging to remove undesirable properties to yield industrial or food oils. As in paper making, specialized equipment design and lengthy product changeovers (including cleaning of pipes, tanks, and vessels) limit process flexibility; product volumes between changeovers are very high, sometimes filling multiple massive tanks in a tank farm.

Extrusion, the third example of continuous processing in [Table 160.1](#), yields such products as polyvinyl chloride (PVC) pipe, polyethylene film, and reels of wire. High process speeds produce high product volumes, such as multiple racks of pipe, rolls of film, or reels of wire per day. Stoppages for changing extrusion heads and many other adjustments limit process flexibility and lead to long production runs between changeovers. Equipment limitations (e.g., physical dimensions of equipment components) keep product variety low.

## 160.4 Mixtures and Gray Areas

---

Many plants contain a mixture of manufacturing types. A prominent example can be found in the process industries, where production usually is only partially continuous. Batch mixing of pulp, fats, or plastic granules precedes continuous paper making, refining of oils, and extrusion of pipe. Further processing may be in the job-shop mode: slitting and length-cutting paper to customer order, secondary mixing and drumming of basic oils to order, and length-cutting and packing of pipe to order.

Mixed production also often occurs in mass production factories. An assembly line (e.g., assembling cars or trucks) may be fed by parts, such as axles, machined in the job-shop mode from castings that are also job-shop produced. Uniform mass-made products (e.g., molded plastic hard hats) may go to storage where they await a customer order for final finishing (e.g., decals) in the job-shop mode. An apparel plant may mass-produce sportswear on one hand and produce custom uniforms for professional sports figures in the job-shop mode on the other.

More than one type of manufacturing in the same plant requires more than one type of production planning, scheduling, and production. The added complexity in management may be offset, however, by demand-side advantages of offering a fuller range of products.

Sometimes, a manufacturing process does not fit neatly into one of the three basic categories. One gray area occurs between mass production and continuous production. Some very small



products—screws, nuts, paper clips, toothpicks—are made in discrete units. But because of small size, high volumes, and uniformity of output, production may be scheduled and controlled not in discrete units but by volume, thus approximating continuous manufacturing. Production of cookies, crackers, potato chips, and candy resembles continuous forming or extrusion of sheet stock on wide belts, except that the process includes die-cutting or other separation into discrete units—like mass production. Link sausages are physically continuous, but links are countable in whole units.

Another common gray area is between mass and job-shop production. A notable example is high-volume production of highly configured products made to order. Products made for industrial uses—such as specialty motors, pumps, hydraulics, controllers, test equipment, and work tables—are usually made in volumes that would qualify as mass production, except that end-product variety is high, not low.

These types of manufacturing with unclear categories do not necessarily create extra complexity in production planning and control. The difficulty and ambiguity are mainly terminological.

## **160.5 Capital Investment, Automation, Advanced Technology, Skills, and Layout**

---

The three characteristics used to categorize manufacturing—volume, variety, and flexibility—are dominant but not exhaustive. To some extent, the manufacturing categories also differ with respect to capital investment, automation, technology, skills, and layout.

Typically, continuous production is highly capital-intensive, whereas mass production is often labor-intensive. The trend toward automated, robotic assembly, however, is more capital-intensive and less labor-intensive, which erodes the distinction. Job-shop production on conventional machines is intermediate as to capital investment and labor intensiveness. However, computer numerically controlled (CNC) machines and related advanced technology in the job shop erodes this distinction as well.

As technology distinctions blur, so do skill levels of factory operatives. In conventional high-volume assembly, skill levels are relatively low, whereas those of machine operators in job shops—such as machinists and welders—tend to be high. But in automated assembly, skill levels of employees tending the production lines elevate toward technician levels—more like machinists and welders. In continuous production skill levels range widely—from low-skilled carton handlers and magazine fillers to highly skilled process technicians and troubleshooters.

Layout of equipment and related resources is also becoming less of a distinction than it once was. The classical job shop is laid out by type of equipment: all milling machines in one area, all grinding machines in another. Mass and continuous production have been laid out by the way the product flows: serially and linearly. Many job shops, however, have been converted to cellular layouts—groupings of diverse machines that produce a family of similar products. In most work cells the flow pattern is serial from machine to machine, but the shape of the cell is not linear; it is U-shaped or, for some larger cells, serpentine. Compact U and serpentine shapes are thought to

provide advantages in teamwork, material handling, and labor flexibility.

To some degree, such thinking has carried over to mass production. That is, the trend is to lay out assembly and production lines in U and serpentine shapes instead of straight lines, which was the nearly universal practice in the past. In continuous production of fluids the tendency has always been toward compact facilities interconnected by serpentine networks of pipes. Continuous production of solid and semisolid products (wide sheets, extrusions, etc.), on the other hand, generally must move in straight lines, in view of the technical difficulties in making direction changes.

## Defining Terms

**Batch:** A quantity (a lot) of a single item.

**Changeover (setup):** Setting up or resetting a process (equipment) for a new product or batch.

**Continuous production:** Perpetual production of goods that flow and are measured by area or volume; usually very high in product volume, very low in product variety, and very low in process flexibility.

**Job-shop production:** Intermittent production with frequent resetting of the process for a different product or batch; usually low in product volume, high in product variety, and high in process flexibility.

**Mass production:** Repetitive production of discrete units on an assembly line or production line; usually high in product volume, low in product variety, and low in process flexibility.

**Process:** A set of resources and procedures that produces a definable product (or service).

**Process industry:** Manufacturing sector involved in continuous production.

## Further Information

*Industrial Engineering.* Published monthly by the Institute of Industrial Engineers.

*Manufacturing Engineering.* Published monthly by the Society of Manufacturing Engineers.

Schonberger, R. J., and Knod, E. M. 1994. *Operations Management: Continuous Improvement*, 5th ed. Richard D. Irwin, Burr Ridge, IL. See, especially, chapters 11, 12, and 13.

Stephens, M. P., Kmec, J. F. "Quality"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 161

## Quality

---

### 161.1 Measurement

Normal Distribution

### 161.2 Statistical Quality Control

Control Charts for Variables • Control Charts for Attributes

### 161.3 Tolerances and Capability

**Matthew P. Stephens**

*Purdue University*

**Joseph F. Kmec**

*Purdue University*

Although no universally accepted definition of **quality** exists, in its broadest sense quality has been described as "conformance to requirements," "freedom from deficiencies," or "the degree of excellence which a thing possesses." Taken within the context of the manufacturing enterprise, quality—or, more specifically, manufacturing quality—shall be defined as "conformance to requirements." This chapter focuses on the evaluation of product quality, with particular emphasis directed at statistical methods used in the measurement, control, and tolerances needed to achieve the desired quality. Factors that define product quality are ultimately determined by the customer and include such traits as reliability, affordability or cost, availability, user friendliness, and ease of repair and disposal. To ensure that quality goals are met, manufacturing firms have initiated a variety of measures that go beyond traditional product inspection and record keeping, which, by and large, were the mainstays of quality control departments for decades. One such initiative is total quality management (TQM) [Saylor, 1992], which focuses on the customer, both inside and outside the firm. It consists of a disciplined approach using quantitative methods to continuously improve all functions within an organization. Another initiative is registration under the ISO 9000 series [Lamprecht, 1993], which provides a basis for U.S. manufacturing firms to qualify their finished products and processes to specified requirements. More recently, the U.S. government has formally recognized outstanding firms through the coveted Malcolm Baldrige Award [ASQC, 1994] for top quality among U.S. manufacturing companies. One of the stipulations of the award is that recipient companies share information on successful quality strategies with their manufacturing counterparts.

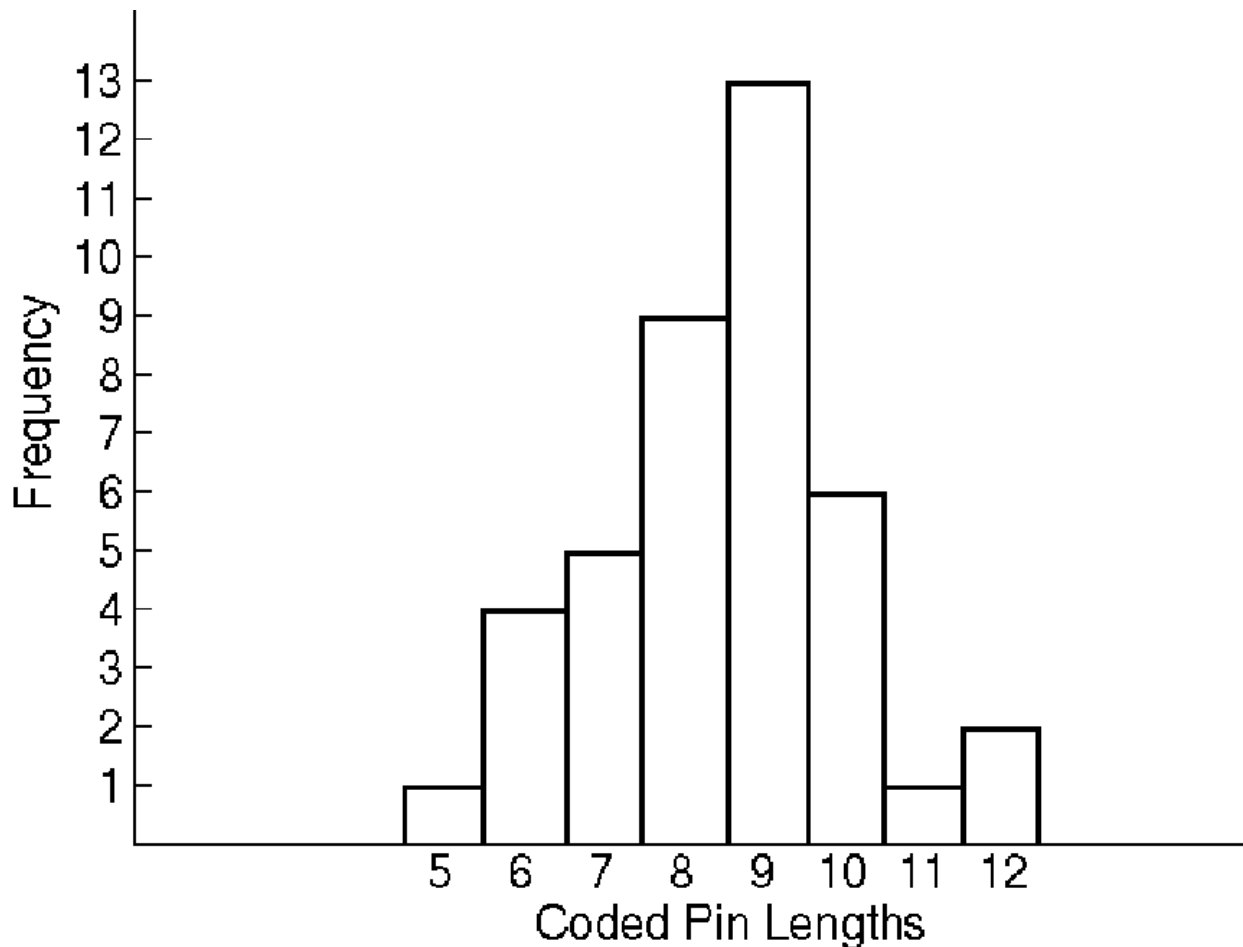
## 161.1 Measurement

---

The inherent nature of the manufacturing process is variation. Variation is present due to any one or a combination of factors including materials, equipment, operators, or the environment. Controlling variation is an essential step in realizing product quality. To successfully control variation, manufacturing firms rely on the measurement of carefully chosen parameters. Because measurement of the entire population of products or components is seldom possible or desirable, **samples** from the **population** are chosen. The extent to which sample data represent the population depends largely on such items as sample size, method of sampling, and time-dependent variations.

Measured data from samples taken during a manufacturing process can be plotted in order to determine the shape of the **frequency distribution**. The frequency distribution can give a visual clue to the process average and dispersion. The latter is referred to as **standard deviation**. [Figure 161.1](#) shows a frequency distribution plot of 40 coded pin lengths expressed in thousands of an inch above 1 inch. Thus the coded length 6 represents an actual length of 1.006 in. For the data shown, average coded pin length is 8.475 and standard deviation is 1.585.

**Figure 161.1** Coded pin lengths.



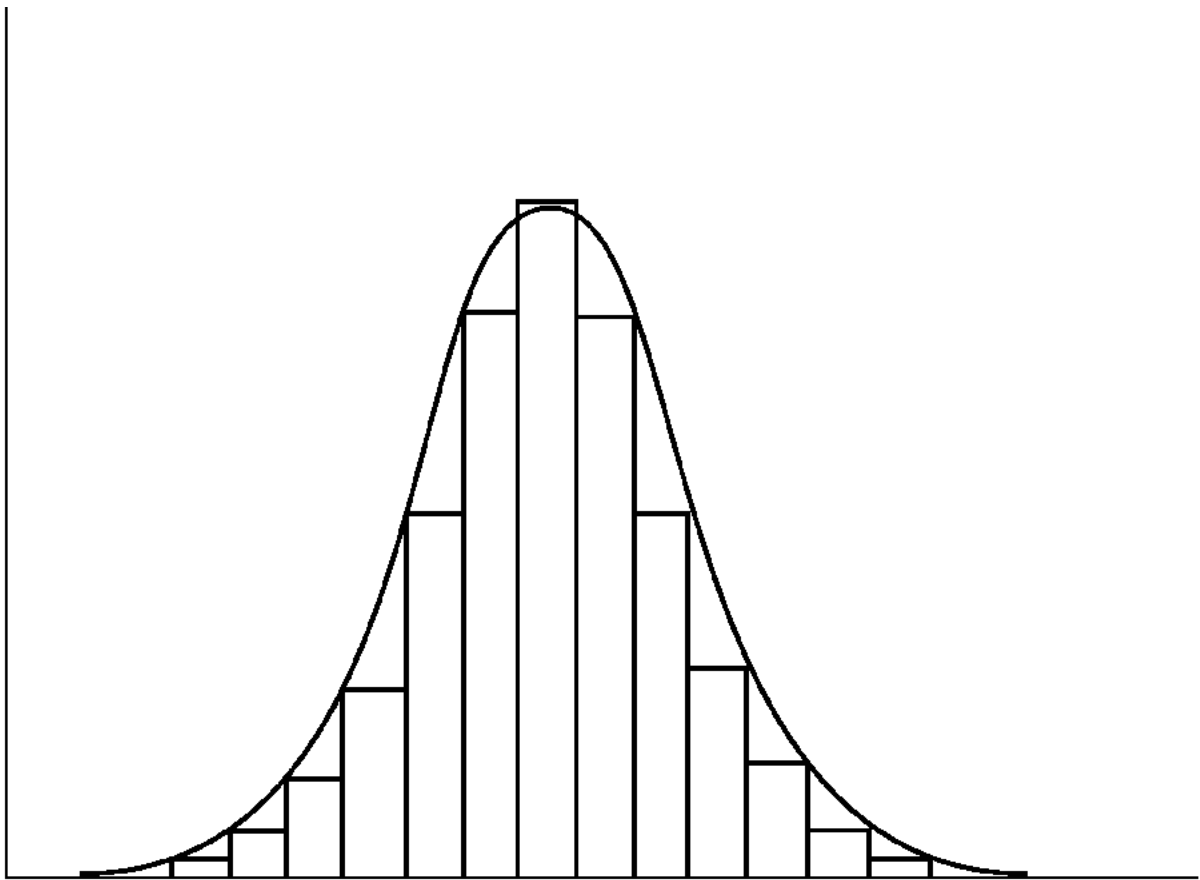
## Normal Distribution

Although there is an infinite variety of frequency distributions, the variation of measured parameters typically found in the manufacturing industry follows that of the normal curve. The normal distribution is a continuous bell-shaped plot of frequency versus some parameter of interest and is an extension of a histogram whose basis is a large population of data points. [Figure 161.2](#) shows a normal distribution plot superimposed on a histogram. Some important properties of the normal distribution curve are:

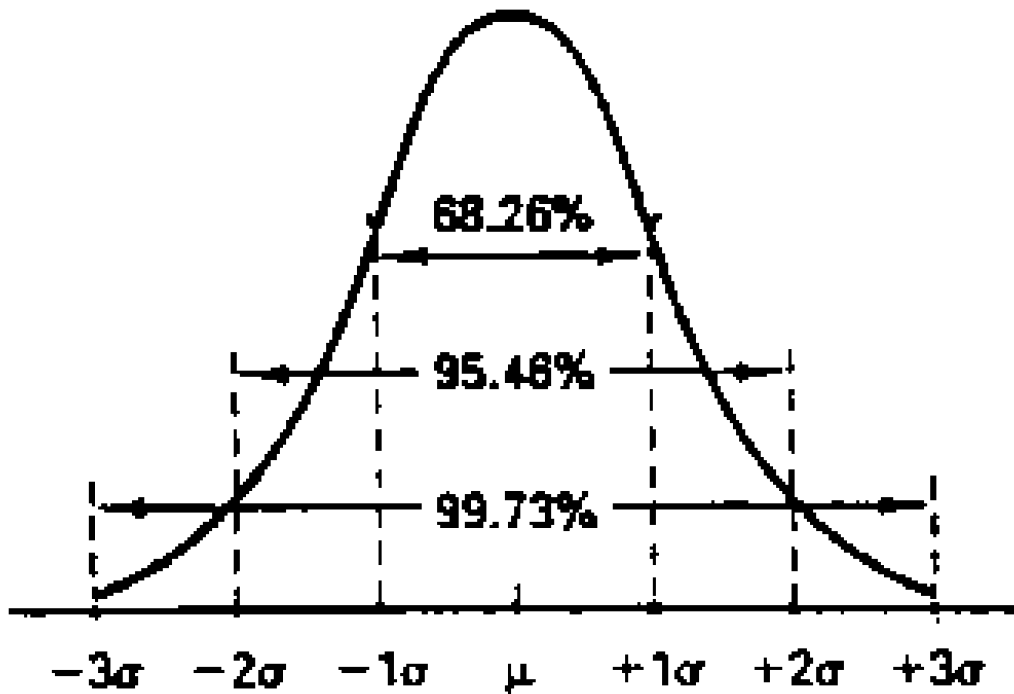
1. The distribution is symmetrical about the population mean  $\mu$ .
2. The curve can be described by a specific mathematical function of population mean  $\mu$  and population standard deviation  $\sigma$ .

An important relationship exists between standard deviation and area under the normal distribution curve. Such a relationship is shown in [Fig. 161.3](#) and may be interpreted as follows: 68.26% of the readings (or area under the curve) will be between  $\pm 1\sigma$  limits, 95.46% of the readings will be between  $\pm 2\sigma$  limits, and 99.73% of the readings will be between  $\pm 3\sigma$  limits. The significance of this relationship is that the standard deviation can be used to calculate the percentage of the population that falls between any two given values in the distribution.

**Figure 161.2** Normal distribution curve.



**Figure 161.3** Percentages under the normal curve.



## 161.2 Statistical Quality Control

---

Statistical quality control (SQC) deals with collection, analysis, and interpretation of data to monitor a particular manufacturing or service process and ensure that the process remains within its capacity. In order to understand process capability, it is necessary to realize that variation is a natural phenomenon that will occur in any process. Parts will appear identical only due to the limitation of the inspection or measurement instrument. The sources of these variations may be the material, process, operator, time of the operation, or any other significant variables. When these factors are kept constant, the minor variations inherent in the process are called *natural* (or *chance*) *variations*, as opposed to variations due to **assignable causes**.

Control charts are utilized to determine when a given process variation is within the expected or natural limits. When the magnitude of variation exceeds these predetermined limits, the process is said to be *out of control*. The causes for out-of-control conditions are investigated and the process is brought back in control. Control charts or the control limits for the natural or chance-cause variations are constructed based on the relationship between the normal distribution and the standard deviation of the distribution. The property of normal distribution and its relationship to the standard deviation of the distribution was discussed in **Chapter 160**. As stated earlier, since approximately 99.73% of a normal distribution is expected to fall between  $\pm 3\sigma$  of the distribution, control limits are established at  $\bar{X} \pm 3\sigma$  for the process. Therefore, any sample taken from the process is expected to fall between the control limits or the  $\bar{X} \pm 3\sigma$  of the process 99.73% of the

time. Any sample not within these limits is assumed to indicate an out-of-control condition for which an assignable cause is suspected.

Control charts can be divided into two major categories: control charts for variables (measurable quality characteristics, i.e., dimension, weight, hardness, etc.) and control charts for attributes (those quality characteristics not easily measurable and therefore classified as conforming or not conforming, good or bad, etc.).

## Control Charts for Variables

The most common charts used for variables are the  $\bar{X}$  and  $R$  charts. The charts are used as a pair for a given quality characteristic. In order to construct control charts for variables, the following steps may be followed:

1. Define the quality characteristic that is of interest. Control charts for variables deal with only one quality characteristic; therefore, if multiple properties of the product of the process are to be monitored, multiple charts should be constructed.
2. Determine the sample (also called the *subgroup*) size. When using control charts, individual measurements or observations are not plotted, but, rather, sample averages are utilized. One major reason is the nature of the statistics and their underlying assumptions. Normal statistics, as the term implies, assumes a normal distribution of the observations. Although many phenomena may be normally distributed, this is not true of all distributions. A major statistical theory called the *central limit theorem* states that the distribution of sample averages will tend toward normality as the sample size increases, regardless of the shape of the parent population. Therefore, plotting sample averages ensures a reasonable normal distribution so that the underlying assumption of normality of the applied statistics is met.

The sample size (two or larger) is a function of cost and other considerations, such as ease of measurement, whether the test is destructive, and the required sensitivity of the control charts. As the sample size increases, the standard deviation decreases; therefore, the control limits will become tighter and more sensitive to process variation.

3. For each sample calculate the sample average,  $\bar{X}$ , and the sample **range**. For each sample record any unusual settings (e.g., new operator, problem with raw material) that may cause an out-of-control condition.
4. After about 20 to 30 subgroups have been collected, calculate

$$\bar{\bar{X}} = \frac{\sum \bar{X}}{g}; \quad \bar{R} = \frac{\sum R}{g}$$

where  $\bar{\bar{X}}$  is the average of averages,  $\bar{R}$  is the average of range, and  $g$  is the number of samples or subgroups.

5. Trial upper and lower control limits for the  $\bar{X}$  and  $R$  chart are calculated as follows:



$$\begin{aligned} \text{UCL}_{\bar{X}} &= \bar{\bar{X}} + A_2 \bar{R}; & \text{UCL}_R &= D_4 \bar{R} \\ \text{LCL}_{\bar{X}} &= \bar{\bar{X}} - A_2 \bar{R}; & \text{LCL}_R &= D_3 \bar{R} \end{aligned}$$

$A_2$ ,  $D_3$ , and  $D_4$  are constants and are functions of sample sizes used. These constants are used to approximate process standard deviation from the range. Tables of these constants are provided in Banks [1989], De Vor *et al.* [1992], Grant and Leavenworth [1988], and Montgomery [1991].

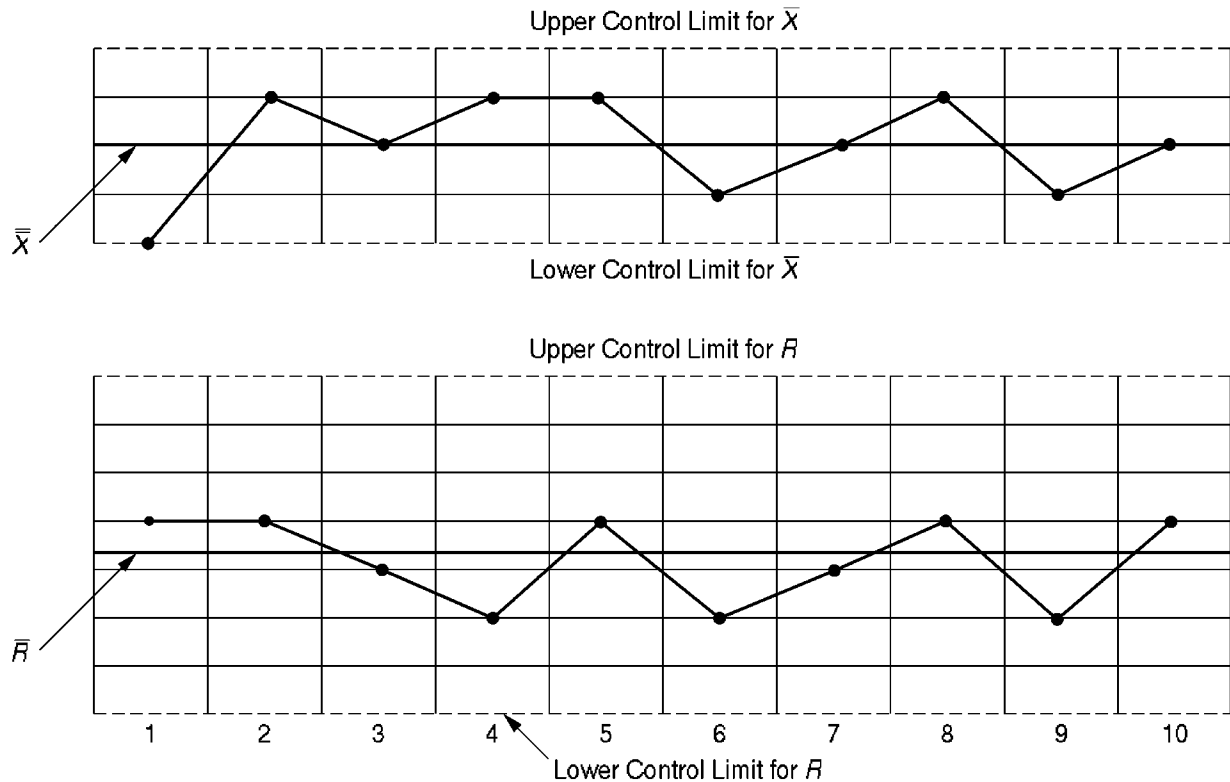
6. Plot the sample averages and ranges on the  $\bar{X}$  and the  $R$  chart, respectively. Any out-of-control point that has an assignable cause (new operator, etc.) is discarded.
7. Calculate the revised control limits as follows:

$$\bar{X}_o = \frac{\sum \bar{X} - \sum \bar{X}_d}{g - g_d}; \quad R_o = \frac{\sum R - \sum R_d}{g - g_d}; \quad \sigma_o = \frac{R_o}{D_2}$$

$$\begin{aligned} \text{UCL}_{\bar{X}_o} &= \bar{X}_o + A\sigma_o & \text{UCL}_R &= D_2\sigma_o \\ \text{LCL}_{\bar{X}_o} &= \bar{X}_o - A\sigma_o & \text{LCL}_R &= D_1\sigma_o \end{aligned}$$

The subscript  $o$  and  $d$  stand for revised and discarded terms, respectively. The revised control charts will be used for the next production period by taking samples of the same size and plotting the sample average and sample range on the appropriate chart. The control limits will remain in effect until one or more factors in the process change. Figure 161.4 shows control charts of  $\bar{X}$  and  $R$  values for ten subgroups. Each subgroup contained five observations because none of the ten data points lie outside of either upper and lower control limits; the process is designated "in control."

**Figure 161.4** Control charts for  $\bar{X}$  and  $R$ .



The control charts can be used to monitor the out-of-control conditions of the process. It is imperative to realize that patterns of variation as plotted on the charts should give clear indications to a process that is headed for an out-of-control condition or one that displays an abnormal pattern of variations. Whereas no point may actually fall out of the limits, variation patterns can often point to some unusual process behavior that requires careful study of the process.

## Control Charts for Attributes

For those quality characteristics that are not easily measured—or in such cases where count of defects or defective items are involved or go-no-go gages are used—control charts for attributes are used. These charts can be grouped into two major categories:

1. Charts for defectives or nonconforming items
2. Charts for defects or nonconformities

### Charts for Nonconforming Items

The basic charts in this group are the fraction **nonconforming** chart ( $p$  chart), percent nonconforming chart ( $100p$  chart), and count of nonconforming chart ( $np$  chart). The procedure for the construction, revision, and the interpretation of control charts for attributes is similar to that for

$\bar{X}$  and  $R$  charts. The following steps may be used to construct a  $p$  chart:

1. Once sample size has been established, fraction nonconforming,  $p$ , is determined for each sample by

$$p = \frac{np}{n}$$

where  $n$  is the sample size and  $np$  is the count of defectives or nonconforming items in the sample.

2. After roughly 20–30 subgroups have been collected, calculate  $\bar{p}$ , the value of the central line, or the average fraction defective.

$$\bar{p} = \frac{\sum np}{\sum n}$$

3. Trial control limits are calculated using:

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$$

4. Plot the fraction defective for each subgroup. The out-of-control subgroups that have assignable causes are discarded, and revised limits are calculated as follows:

$$p_o = \frac{\sum np - \sum np_d}{\sum n - n_d}$$

$$UCL = p_o + 3\sqrt{\frac{p_o(1 - p_o)}{n}}$$

$$LCL = p_o - 3\sqrt{\frac{p_o(1 - p_o)}{n}}$$

5. If the lower control limit is a negative number, it is set to zero. Sample points that fall above the upper limit indicate a process that is out of control. However, samples that fall below the lower limit, when the lower control limit is greater than zero, indicate a product that is better than expected. In other words, if a sample contains fewer nonconforming items than the process is capable of producing, the sample fraction defective will fall below the lower control limit. For this reason some practitioners may choose to set the lower limit of the attribute charts to zero. This practice, however, may mask other problems or potentials for process improvements.

Other charts for nonconforming items are simple variations of the  $p$  chart. In the case of the  $100p$  chart, all values of the  $p$  chart are expressed as percentages. In the case of the  $np$  chart, instead of plotting fraction or percent defectives, actual counts of nonconforming or defective items are plotted. See Banks [1989], De Vor *et al.* [1992], Grant and Leavenworth [1988], and Montgomery [1991] for greater detail. The formulas for the central line and the control limits for an  $np$  chart are given below. It is assumed that the revised value for universe fraction defective,  $p$ , is known. If  $p$  is not known, then the procedure for the  $p$  chart must be carried out to determine the revised value for the universe fraction defective.

$$\begin{aligned}\text{Central line} &= np_o \\ \text{Control limits} &= np_o \pm 3\sqrt{np_o(1 - p_o)}\end{aligned}$$

where  $n$  is the sample size and  $p_o$  is the universe fraction defective.

### Charts for Defects or Nonconformities

Whereas the charts for defective or nonconforming items are concerned with the overall quality of an item or sample, charts for defects look at each defect (i.e., blemish, scratch, etc.) in each item or sample. One may consider an item a nonconforming item based on its overall condition. A defect or **nonconformity** is that condition that makes an item a nonconforming or defective item.

In this category are  $c$  charts and  $u$  charts. The basic difference between the two is the sample size. The sample size,  $n$ , for a  $c$  chart is equal to one. In this case the number of nonconformities or defects are counted per a single item. For a  $u$  chart, however,  $n > 1$ . See Banks [1989], De Vor *et al.* [1992], Grant and Leavenworth [1988], and Montgomery [1991] for the formulas and construction procedures.

## 161.3 Tolerances and Capability

---

As stated earlier, *process capability* refers to the range of process variation that is due to chance or natural process deviations. This was defined as  $\bar{X} \pm 3\sigma$  (also referred to as  $6\sigma$ ) which is the expected or natural process variation. **Specifications** or **tolerances** are dictated by design engineering and are the maximum amount of acceptable variation. These specifications are often stated without regard to process spread. The relationships between the process spread or the natural process variation and the engineering specifications or requirements are the subject of process capability studies. Process capability can be expressed as:

$$C_p = \frac{US - LS}{6\sigma}$$

where

$C_p$  = process capability index

US = upper engineering specification value

LS = lower engineering specification value

A companion index,  $C_{pk}$ , is also used to describe process capability, where

$$C_{pk} = \frac{US - \bar{X}}{3\sigma}$$

or

$$C_{pk} = \frac{\bar{X} - LS}{3\sigma}$$

The lesser of the two values indicates the process capability. The  $C_{pk}$  ratio is used to indicate whether a process is capable of meeting engineering tolerances and whether the process is centered around the target value  $\bar{X}$ . If the process is centered between the upper and the lower specifications,  $C_p$  and  $C_{pk}$  are equal. However, if the process is not centered,  $C_{pk}$  will be lower than  $C_p$  and is the true process capability index. See De Vor *et al.* [1992], Grant and Leavenworth [1988], and Montgomery [1991] for additional information.

A capability index less than one indicates that the specification limits are much tighter than the process spread. Hence, although the process may be in control, the parts may well be out of specification. Thus the process does not meet engineering requirements. A capability index of one means that as long as the process is in control, parts are also in-spec. The most desirable situation is to have a process capability index greater than one. In such cases, not only are approximately 99.73% of the parts in-spec when the process is in control, but even if the process should go out of control, the product may still be within the engineering specifications. Process improvement efforts are often concerned with reducing the process spread and, therefore, increasing the process capability indices.

An extremely powerful tool for isolating and determining those factors that significantly contribute to process variation is statistical design and analysis of experiments. Referred to as "design of experiments," the methodology enables the researcher to examine the factors and determine how to control these factors in order to reduce process variation and therefore increase process capability index. For greater detail, see Box *et al.* [1978].

## Defining Terms

**Assignable causes:** Any element that can cause a significant variation in a process.

**Frequency distribution:** Usually a graphical or tabular representation of data. When scores or measurements are arranged, usually in an ascending order, and the occurrence (frequency) of each score or measurement is also indicated, a frequency distribution results.

No. of Defectives	Frequency
0	10
1	8
2	7
3	8
4	6
5	4
6	2
7	1
8	1
9	0

The frequency distribution indicates that ten samples were found containing zero defectives.

**Nonconforming:** A condition in which a part does not meet all specifications or customer requirements. This term can be used interchangeably with *defective*.

**Nonconformity:** Any deviation from standards, specifications, or expectation; also called a *defect*. Defects or nonconformities are classified into three major categories: critical, major, and minor. A critical nonconformity renders a product inoperable or dangerous to operate. A major nonconformity may affect the operation of the unit, whereas a minor defect does not affect the operation of the product.

**Population:** An entire group of people, objects, or phenomena having at least one common characteristic. For example, all registered voters constitute a population.

**Quality:** Quality within the framework of manufacturing is defined as conformance to requirements.

**Range:** A measure of variability or spread in a data set. The range of a data set,  $R$ , is the difference between the highest and the lowest values in the set.

**Sample:** A small segment or subgroup taken from a complete population. Because of the large size of most populations, it is impossible or impractical to measure, examine, or test every member of a given population.

**Specifications:** Expected part dimensions as stated on engineering drawings.

**Standard deviation:** A measure of dispersion or variation in the data. Given a set of numbers, all of equal value, the standard deviation of the data set would be equal to zero.

**Tolerances:** Allowable variations in part dimension as stated on engineering drawings.

## References

- ASQC. 1994. *Malcolm Baldrige National Quality Award<sup>®</sup> 1994 Award Criteria*. American Society for Quality Control, Milwaukee, WI.
- Banks, J. 1989. *Principles of Quality Control*. John Wiley & Sons, New York.
- Box, G. E. P., Hunter, W. G., and Hunter, J. S. 1978. *Statistics for Experimenters*. John Wiley & Sons, New York.

- De Vor, R. E., Chang, T. H., and Sutherland, J. W. 1992. *Statistical Quality Design and Control*. Macmillan, New York.
- Grant, E. L. and Leavenworth, R. S. 1988. *Statistical Quality Control*, 6th ed. McGraw-Hill, New York.
- Lamprecht, J. L. 1993. *Implementing the ISO 9000 Series*. Marcel Dekker, New York.
- Montgomery, D. C. 1991. *Statistical Quality Control*, 2nd ed. John Wiley & Sons, New York.
- Saylor, J. H. 1992. *TQM Field Manual*. McGraw-Hill, New York.

## **Further Information**

*Statistical Quality Control*, by Eugene Grant and Richard Leavenworth, offers an in-depth discussion of various control charts and sampling plans.

*Statistics for Experimenters*, by George Box, William Hunter, and Stewart Hunter, offers an excellent and in-depth treatment of design and analysis of design of experiments for quality improvements.

Most textbooks on statistics offer detailed discussions of the central limit theorem. *Introduction to Probability and Statistics for Engineers and Scientists*, written by Sheldon Ross and published by John Wiley & Sons, is recommended.

American Society for Quality Control, P.O. Box 3005, Milwaukee, WI 53201-3005, phone: (800)248-1946, is an excellent source for reference material, including books and journals, on various aspects of quality.

Kusiak, A., Feng, C. X. "Flexible Manufacturing"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



## 162.1 Flexible Machining

An Example of FMS

## 162.2 Flexible Assembly

## 162.3 The Economic Justification of Flexibility

**Andrew Kusiak**

*University of Iowa*

**Chang-Xue Feng**

*University of Iowa*

Flexible manufacturing systems (FMSs) have emerged to meet frequently changing market demands for products. The distinguishing feature of an FMS is in its software (e.g., numerical control programs) rather than the hardware (using, for example, relays and position switches). Fixed automation, programmable automation, and flexible automation are the three different forms of automation [Groover *et al.*, 1986]. Numerical control (NC) and computer numerical control (CNC) technologies are essential in FMSs. Developments in robotics, automated guided vehicles (AGVs), programmable controllers (PCs), computer vision, group technology (GT), and statistical quality control (SQC) have accelerated the applications of FMSs. The computer technology (hardware and software) is a critical factor in determining the performance of flexible manufacturing systems. An FMS is frequently defined as a set of CNC machine tools and other equipment that are connected by an automated material-handling system, all controlled by a computer system [Askin and Standridge, 1993].

The FMS concept has been applied to the following manufacturing areas [Kusiak, 1986]: (1) fabrication, (2) machining, and (3) assembly. An FMS can be divided into three subsystems [Kusiak, 1985]: (1) management system, (2) production system, and (3) material-handling system. The management system incorporates computer control at various levels; the production system includes CNC machine tools and other equipment (e.g., inspection stations, washing stations); the material handling system includes AGVs, robots, and automated storage/retrieval systems (AS/RSs).

The planning, design, modeling, and control of an FMS differs from the classical manufacturing system. The structure and taxonomy of an FMS has been studied in Kusiak [1985]. Some planning tools—for example, IDEF (the U.S. Air Force ICAM definition language) [U.S. Air Force, 1981]—have been proposed for modeling of material flow, information flow, and dynamic

(simulation) modeling of an FMS. Numerous mathematical models of FMSs are discussed in Askin and Standridge [1993]. An FMS can be structured as a five-level hierarchical control model [Jones and McLean, 1986] (see Table 162.1).

**Table 162.1** FMS Control Architecture (National Institute of Standards and Technology) [Jones and McLean, 1996]

Level	Planning Horizon	Functions
1. Facility	Months to years	Information management Manufacturing engineering Production management
2. Shop	Weeks to months	Task management Resource allocation
3. Cell	Hours to weeks	Task analysis Batch management Scheduling Dispatching Monitoring
4. Workstation	Minutes to hours	Setup Equipment sequencing In-process inspection
5. Equipment	Milliseconds to minutes	Machining Measurement Handling Transport Storage

Manufacturing concepts such as just-in-time (JIT) manufacturing, lean manufacturing, and agile manufacturing relate to the flexible manufacturing approach. Here the term *flexible* describes the system's ability to adjust to customers' preferences; JIT reduces the throughput time and the level of inventory; "lean" stresses efficiency and cost reduction; and "agile" addresses the speed of a manufacturing system in responding to the changing market demands. Other terms such as flexible manufacturing cells (FMCs), computer-aided manufacturing (CAM), and computer integrated manufacturing (CIM) are frequently used to describe flexible manufacturing.

## 162.1 Flexible Machining

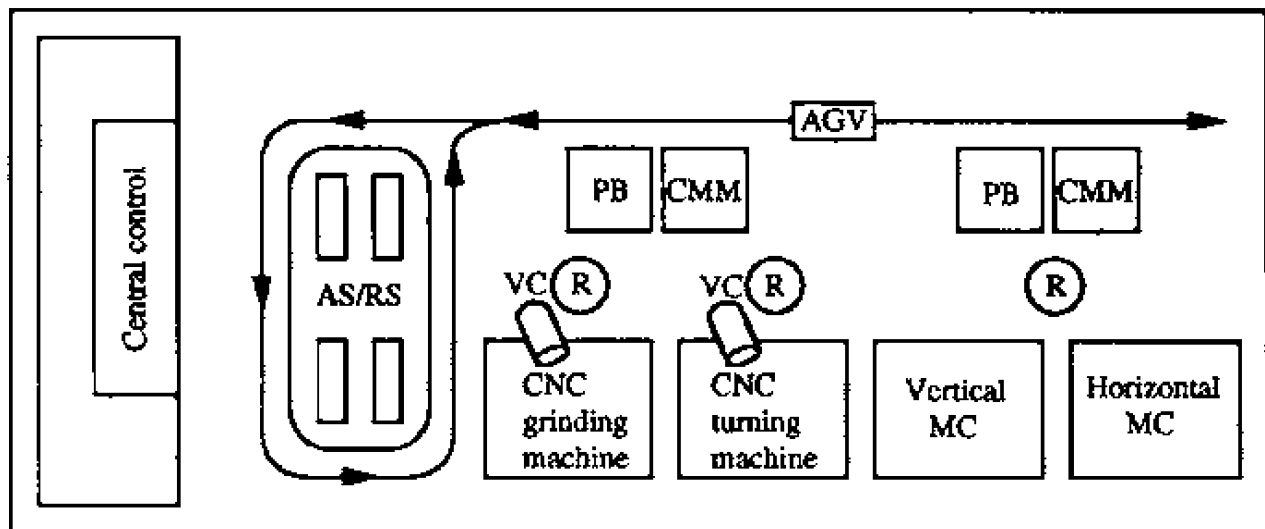
The basic features of a flexible machining system (FMS), as well as the differences between a flexible machining system and the equivalent classical machining system, are presented in Kusiak [1990] in the form of eight observations. These observations are based on the analysis of over 50 flexible machining systems; for each observation, a brief justification is provided.

### An Example of FMS

Figure 162.1 illustrates one type of flexible machining system (FMS). The FMS produces parts for

a family of DC motors used in the FANUC CNC system. The parts are of two types: rotational (e.g., a motor shaft) and prismatic (e.g., a supporting house). The production system includes a CNC turning machine tool, a CNC grinding machine tool, a vertical machining center, and a horizontal machining center, as well as two coordinate-measuring machines (CMMs). Programmable controllers (PCs) are also used in the CNC control system to perform some auxiliary functions. Quite frequently, a PC can be used as an independent CNC control device in a position control mode (as opposed to a path control mode), for example, in a CNC drilling machine to directly drive the servo motor. A CMM measures automatically the precision of the machined parts, including their nominal values and tolerances. It includes a moving head, a probe, a microcomputer, and input/output (I/O) devices. It is controlled by an NC program, and it feeds the measured data to the corresponding CNC system and records them for the purpose of statistical quality control (SQC). The essential function of the SQC system is to analyze the data from the CMM and to construct an X-bar chart, which shows the measured values and their mean value based on a predetermined sample size and grouping. This mean value and the changing range are then compared to the designed nominal value and tolerance, respectively, to determine whether the process is under control.

**Figure 162.1** Physical layout of a flexible machining system. AS/RS: automated storage and retrieval system; PB: part buffer; CMM: coordinate-measuring machine; R: robot; VC: video camera; MC: machining center.



The automated material handling system in Fig. 162.1 includes an AGV, three robots, two part buffers, and the automated storage and retrieval system (AS/RS). The function of the AGV in Fig. 162.1 is to transfer loads to remote locations following a guided path. For the long-distance handling in this FMS, a robot cannot provide the mobility of an AGV, and a conveyor does not offer the flexibility desired. The AGV sends the raw material from the AS/RS to the part buffer of each workstation and brings back the finished products from the part buffers to the AS/RS. An industrial robot is a reprogrammable, multifunctional manipulator designed to move materials,

parts, tools, or special devices through variable programmed motions for the performance of a variety of tasks (Robotics Industries Association). In the system in Fig. 162.1 the three robots load parts from part buffers to each CNC machine tool, transfer parts from the CNC machine tools to the CMMs, and unload parts from machines to the part buffers. The AS/RS is used to temporarily store the raw material and finished product. Although the AGV and the robots in the system in Fig. 162.1 function separately, it is possible for them to work jointly to provide both mobility and flexibility. For example, an AGV can be interfaced with a robot to provide more flexible loading and unloading capability, and a robot can be mounted on an AGV to achieve mobility.

The control and management system includes the central control minicomputer, the microcomputer resided in each CNC machine and robot, and two industrial video cameras. The computer control system performs two fundamental functions: monitoring and control. For example, it stores process plans and NC programs; dispatches programs to the individual pieces of equipment; controls the production, traffic, tools, and quality; monitors the material handling system; and monitors and reports the system performance. The video camera (VC) monitors the critical machining processes. A VC might or might not have the capability to analyze and process the data collected. In the former case it can process the image based on computer vision technology and inform the CNC system; in the latter case a human expert monitors the image taken by the VC and then manually informs the corresponding CNC system. Other types of monitoring might be used in an FMS, for example, measuring the cutting force, torque, or power. In some types of monitoring a measuring head may be placed in the tool magazine and used to measure the machining quality of certain operations as required. The use of the manufacturing automation protocol (MAP) in manufacturing is reaching a momentum. The equipment purchased from different vendors is linked through a shop local-area network (LAN), as the vendors tend to adopt the MAP standard. Tool management is one of the functions of the computer control system [Kusiak 1990].

## 162.2 Flexible Assembly

---

Most of the observations regarding the flexible machining systems [Kusiak, 1990] apply to the flexible assembly systems (FASs) as well. Any differences that may arise are of terminology rather than concept. The earliest FASs were designed for assembling printed circuit boards. More recent FASs have been developed for mechanical assembly. The automated material-handling system—including AGVs, conveyors, and robots—is more important in an FAS than an FMS. The design of a product is more closely related to the design and operations of an FAS than to that of a component with a flexible machining system. Boothroyd *et al.* [1982] discusses in depth rules for design of products for automated assembly, including flexible assembly systems. Another widely used design for assembly method, the assembly evaluation method (AEM), was developed at Hitachi. The basic design for assembly (DFA) rules include:

1. Minimize the number of components in a product.
2. Ensure that a product has a suitable base (part) on which the assembly can be built.

3. Ensure that the product base has features that will enable it to be readily located in a suitable position in the horizontal plane.
4. Minimize the number of assembly directions; if possible, use only top-down, vertical assembly motions.
5. If possible, design the product so that it can be assembled in layers.
6. Provide chamfers or tapers to help the component be easily positioned.
7. Avoid expensive and time-consuming fastening operations and avoid using assembly tools (e.g., screwdrivers).
8. Design the product so that the assembly does not have to be lifted or rotated, since the lifting and rotating results in complicated fixtures and grippers, more degrees of freedom of the robot, and increased cycle time.
9. Design the product to simplify packaging. The more uniform and simple the product packaging is, the easier it will be to apply a robotic packaging station that uses standard carts and pallets.
10. Use symmetric features whenever possible; otherwise, exaggerate asymmetry to facilitate identification and orientation of components.
11. Design a component to avoid tangling when feeding.
12. Design product variations to allow common handling and assembly methods.

An example of mechanical product design for flexible assembly is presented in Elmaraghy and Knoll [1989]. The authors discuss design for flexible assembly of a family of DC motors and the design of the flexible assembly system.

The IBM Proprinter is a frequently cited example of an electronic product that was designed for flexible automated assembly. The goals for the printer design were relatively simple:

- Develop a modular design.

- Minimize the total number of parts.

- All parts are to be self-aligning for ease of assembly.

Other principles of design for assembly that were used in this example are as follows:

- No fasteners

- No springs

- No pulleys

- No excess modules

- No adjustment

- No labels

- No paint

- No extra parts

- No multidirectional assembly

- No alignment/location tools or fixtures

- No external tests

- No engineering change notices (ECNs)

No custom builds or options  
No manual assembly

## 162.3 The Economic Justification of Flexibility

---

Flexibility is a critical objective in design, planning, and operating an FMS. An FMS is intended to accommodate a large number of types of products and allow for more changes of products and higher equipment utilization, whereas in lean and agile manufacturing, flexibility means not only more changes and more product types, but also quicker changes, plus a reasonable rather than ultimately high utilization. Although none of the definitions of manufacturing flexibility have been generally accepted, the following types of flexibility are frequently considered: machine flexibility, process flexibility, product flexibility, routing flexibility, volume flexibility, expansion flexibility, process sequence flexibility, operation flexibility, and production flexibility. How to price the product produced in an FMS causes difficulties and often creates conflicts between accounting and engineering departments. There is some evidence that the U.S. and Japan operate FMSs in radically different ways. Jaikumar [1986] shows that Japanese users had achieved at the time greater flexibility and yet shorter development time than users in the U.S.

### References

- Askin, R. and Standridge, C. 1993. *Modeling and Analysis of Manufacturing Systems*. John Wiley & Sons, New York.
- Boothroyd, G., Poli, C., and Murch, L. 1982. *Automatic Assembly*. Marcel Dekker, New York.
- Elmaraghy, H. and Knoll, L. 1989. Flexible assembly of a family of DC motors. *Manufacturing Review*. 2(4): 250–256.
- Groover, M., Weiss, M., Nagel, R., and Odrey, N. 1986. *Industrial Robotics: Technology, Programming, and Applications*. McGraw-Hill, New York.
- Jaikumar, R. 1986. Post-industrial manufacturing. *Harv. Bus. Rev.* 86(6): 69–76.
- Jones, A. and McLean, C. 1986. A proposed hierarchical control model for automated manufacturing systems. *J. Manufacturing Syst.* 5(1): 15–25.
- Kusiak, A. 1985. Flexible manufacturing systems: A structural approach. *Int. J. Prod. Res.* 23(6): 1057–1073.
- Kusiak, A. 1986. Parts and tools handling systems. In *Modelling and Design of Flexible Manufacturing Systems*, ed. A. Kusiak, p. 99–110. Elsevier, New York.
- Kusiak, A. 1990. *Intelligent Manufacturing Systems*. Prentice Hall, Englewood Cliffs, NJ.
- U.S. Air Force. 1981. *U.S. Air Force Integrated Computer Aided Manufacturing (ICAM) Architecture Part II, Volume IV<sup>3/4</sup>Functional Modeling Manual (IDEF0)*. AFWAL-tr-81-4023. Air Force Materials Laboratory, Wright-Patterson AFB, OH.

## Further Information

The *Handbook of Flexible Manufacturing Systems* published by Academic Press in 1991 provides a good overview of various aspects of FMSs from theoretical and application point of view.

The *International Journal of Flexible Manufacturing Systems* (Kluwer) publishes papers in design, analysis, and operations of flexible fabrication and assembly systems.

The *Journal of Intelligent Manufacturing* (Chapman and Hall) emphasizes the theory and applications of artificial intelligence in design and manufacturing.

A comprehensive review of existing FMS models and future directions is provided in A. Gunasekran, T. Martikainen, and P. Yli-Olli. 1993. Flexible manufacturing systems: An investigation for research and applications. *European Journal of Operational Research*. 66(1): 1–26.

For up-to-date information on flexible manufacturing, see *Handbook of Design, Manufacturing and Automation*, edited by R. Dorf and A. Kusiak and published by John Wiley & Sons, New York, in 1994.

*The Machine that Changed the World*<sup>3/4</sup>*The Story of Lean Production*, by J. Womack, D. Jones, and D. Roos of the MIT International Automobile Program, published by HarperCollins, New York, 1991, coins the term *lean production*, discusses the differences between the mass production and lean production. *Just-in-Time Manufacturing Systems: Operational Planning and Control Issues*, edited by A. Satir, includes a collection of papers presented in the International Conference on Just-in-Time Manufacturing held in Montréal, Canada, 2–4 October 1991.

In 1987 the Society of Manufacturing Engineers (SME) published *Automated Guided Vehicles and Automated Manufacturing* by R. Miller. It discusses AGV principles and their applications in FMSs.

*Flexible Assembly Systems* by A. Owen was published by Plenum Press, New York, 1984. In 1987 G. Boothroyd and P. Dewhurst published the book *Product Design for Assembly Handbook* (Boothroyd Dewhurst, Inc.), which presents comprehensive design for assembly rules and examples. A commercial software package is also available. For design for assembly of electronic products, see S. Shina. 1991. *Concurrent Engineering and Design for Manufacture of Electronic Products*. Van Nostrand Reinhold, New York.

Edward M. Knod, Jr.. “Management and Scheduling”  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



## Management and Scheduling

---

### 163.1 Management: Definition and Innovations

Duties and Activities • Requisite Skills and Attributes • Trends and Innovations • Principles for Managing Operations

### 163.2 Scheduling

Relationship to Planning • Effects of Manufacturing Environment • Effects of Contemporary Management Practices

### **Edward M. Knod, Jr.**

*Western Illinois University*

Prescriptions for how managers ought to establish and maintain world-class excellence in their organizations changed substantially during the 1980s and 1990s. Emerging markets, shifting patterns of global competitiveness and regional dominance in key industries, the spread of what might be called Japanese management and manufacturing technologies, and the philosophy and tools of the total quality movement are among the factors that combined to usher in a heightened focus on the overall capacity required to provide continuous improvement in meeting evolving customer needs. The effects of this contemporary management approach first appeared in manufacturing [[Schonberger, 1982](#); [Hall, 1983](#)] and continue to have profound influence in that sector. Changes in the way managers approach scheduling serve to exemplify the new thinking.

This chapter begins with an overview of contemporary management, continues with a discussion of scheduling in various manufacturing environments, and concludes with references and suggested sources of additional information.

### **163.1 Management: Definition and Innovations**

---

In a somewhat general-to-specific progression, management in contemporary competitive manufacturing organizations may be described by (1) duties and activities, (2) requisite skills and attributes, (3) trends and innovations, and (4) principles for managing operations.

### **Duties and Activities**

Briefly, the goal of management is to ensure organizational success in the creation and delivery of goods and services. Popular definitions of management often employ lists of activities that describe

what managers do. Each activity serves one or more of three general, and overlapping, duties: creating, implementing, and improving.

- *Creating.* Activities such as planning, designing, staffing, budgeting, and organizing accomplish the creativity required to build and maintain customer-serving capacity. Product and process design, facility planning and layout, work-force acquisition and training, and materials and component sourcing are among the tasks that have a substantial creative component.
- *Implementing.* When managers authorize, allocate, assign, schedule, or direct, emphasis shifts from creating to implementing—putting a plan into action. (A frequent observation is that the biggest obstacle to successful implementation is lack of commitment.) During implementation, managers also perform controlling activities—that is, they monitor performance and make necessary adjustments.
- *Improving.* Environmental changes, (e.g., new or revised customer demands, challenges from competitors, and social and regulatory pressures) necessitate improvements in output goods and services. In response to—or better yet, in anticipation of—those changes, managers re-create—that is, they start the cycle again with revised plans, better designs, new budgets, and so forth.

Outcomes, desirable or otherwise, stem from these activities. A goal-oriented manager might describe the aim of the job as "increased market share" or "greater profitability," but he or she will try to attain that goal by creating, implementing, and improving.

## Requisite Skills and Attributes

Exact requirements are hard to pin down, but any skill or attribute that helps a manager make better decisions is desirable. Bateman and Zeithaml [1993] suggest that managers need technical skills, interpersonal and communications skills, and conceptual and decision skills. Extension of these broad categories into job-specific lists is perhaps unwarranted given current emphasis on cross-functional career migration and assignments to interdisciplinary project teams or product groups. Sound business acumen and personal traits such as ethical demeanor, good time-management habits, and pleasant personality, however, are general attributes that serve managers in any job. More recently, emphasis on computer (and information system) literacy and knowledge of foreign languages and customs has increased as well.

## Trends and Innovations

An array of publications, seminars, and other vehicles for disseminating "how and why" advice has bolstered the spread of contemporary management theory and research. Manufacturing managers constitute the primary target audience for much of this work. The information bounty can be reduced, tentatively, to a set of core trends and innovative approaches. [Table 163.1](#) offers a short list of seven concept areas within which numerous interrelated trends or innovations have emerged. They are dominant themes in contemporary management literature and in that regard help define what today's managers are all about.

**Table 163.1** Trends and Innovations in Management

---

**Customers at center stage.** The customer is the next person or process—the destination of one's work. The provider-customer chain extends, process to process, on to final consumers. Whatever a firm produces, customers would like it to be better, faster, and cheaper; prudent managers therefore embrace procedures that provide *total quality, quick responses, and waste-free* (economical) *operations*. These three aims are mutually supportive and form the core rationale for many of the new principles that guide managers.

**Focus on improvement.** Managers have a duty to embrace improvement. A central theme of the total quality (TQ) movement is constant improvement in output goods and services and in the processes that provide them. Sweeping change over the short run, exemplified by business process reengineering [Hammer and Champy, 1993], anchors one end of the improvement continuum; the rationale is to discard unsalvageable processes and start over so as not to waste resources in fruitless repair efforts. The continuum's other end is described as incremental continuous improvement and is employed to fine-tune already-sound processes for evenbetter results.

**Revised "laws" of economics.** Examples of the contemporary logic include the following. Quality costs less, not more. Costs should be allocated to the activities that cause their occurrence. Prevention (of errors) is more cost-effective than discovery and rework. Training is an investment rather than an expense. Value counts more than price (e.g., in purchasing). Desired market price should define (target) manufacturing cost, notthe reverse.

**Elimination of wastes.** Waste is anything that doesn't add value; it adds cost, however, and should be eliminated. Waste detection begins with two questions: "Are we doing the right things?" and "Are we doing those things in the right way?" Toyota identifies seven general categories of wastes [Suzaki, 1987, ch. 1], each with several subcategories. Schonberger [1990, ch. 7] adds opportunities for further waste reduction by broadening the targets to include nonobvious wastes. Simplification or elimination of indirect and support activities (e.g., production planning, scheduling, and control activities; inventory control; costing and reporting, etc.) is a prime arena for contemporary waste-reduction programs [Steudel and Desruelle, 1992, ch. 8].

**Quick-response techniques.** Just-in-time (JIT) management, queue-limiters, reduced setups, better maintenance, operator-led problem solving, and other procedures increase the velocity of material flows, reduce throughput times, and eliminate the need for many interdepartmental control transactions. Less tracking and reporting (which add no value) reduces overhead. Collectively, quick-response programs directly support faster customer service [Blackburn, 1991, ch. 11].

**The process approach.** The **process** approach has several advantages [Schonberger and Knod, 1994, ch. 4]. Processes cut across functional departments; attention is drawn to overall or group results, ideally by a cross-functional team that may also include customers and suppliers. Processes are studied at the job or task level, or at the more detailed operations level. Automation can be beneficial after' process waste is removed and further reduction in variation is needed. Tools for measurement and data analysis, methods improvement, and team building are among those needed for successful process improvement.

**Human resources management.** Increased reliance on self-directed teams (e.g., in cells or product groups) and/or on-line operators for assumption of a larger share of traditional management responsibilities is a product of the management revolution of the 1980s that had noticeable impact in the 1990s. Generally, line or shop employees have favored those changes; they get more control over their workplaces. There is a flipside: As employee empowerment shifts decision-making authority, as JIT reduces the need for many reporting and control activities, and as certain supervisory and support-staff jobs are judged to be non-value-adding, many organizations downsize. Lower and mid-level managers and support staff often bear the job-loss burden.

---

# Principles for Managing Operations

The final and most detailed component of this general-to-specific look at contemporary management is a set of action-oriented, prescriptive principles for managing operations in any organization; see [Table 163.2](#). The principles apply to managers at any level and define ways for increasing competitiveness in manufacturing organizations. Brief supporting rationale and techniques or procedures that exemplify each principle appear in the right-hand column; Schonberger and Knod [[1994](#), ch. 1] provide a more detailed discussion.

**Table 163.2** Principles for Managing Operations

	Principle	Rationale and Examples
1.	Get to know customers; team up to form partnerships and share process knowledge.	Providers are responsible for getting to know their customers' processes and operations. By so doing, they offer better and faster service, perhaps as a member of customers' teams.
2.	Become dedicated to rapid and continual increases in quality, flexibility, and service; and decreases in costs, response or lead time, and variation from target.	The logic of continuous improvement, or <i>kaizen</i> [ <a href="#">Imai, 1986</a> ], rejects the "if it ain't broke" philosophy; seeks discovery and then prevention of current and potential problems; and anticipates new or next-level standards of excellence.
3.	Achieve unified purpose through shared information and cross-functional teams for planning/design, implementation, and improvement efforts.	Information sharing keeps all parties informed. Early manufacturing/supplier involvement (EMI/ESI), and concurrent or simultaneous product and process design are components of the general cross-functional team design concept.
4.	Get to know the competition and world-class leaders.	Benchmarking [ <a href="#">Camp, 1989</a> ] elevates the older notion of "reverse engineering" to a more formal yet efficient means of keeping up with technology and anticipating what competitors might do. Search for best practices.
5.	Cut the number of products (e.g., types or models), components, or operations; reduce supplier base to a few good ones and form strong relationships with them.	Product line trimming removes nonperformers; component reduction cuts lead times by promoting simplification and streamlining. Supplier certifications and registrations (e.g., ISO 9000) lend confidence, allow closer partnering with few suppliers (e.g., via EDI), and reduce overall buying costs.
6.	Organize resources into multiple chains of customers, each focused on a family of products or services; create cells, flow lines, plants-in-a-plant.	Traditional functional organization by departments increases throughput times, inhibits information flow, and can lead to "turf battles." Flow lines and cells promote focus, aid scheduling, and employ cross-functional expertise.
7.	Continually invest in human resources through cross-training for mastery of multiple skills, education, job and career path rotation, health and safety, and security.	Employee involvement programs, team-based activities, and decentralized decision responsibility depend on top-quality human resources. Cross-training and education are keys to competitiveness. Scheduling—indeed, all capacity management—is easier when the work force is flexible.
8.	Maintain and improve present equipment and human work before acquiring new equipment, then automate incrementally	<i>TPM</i> , total productive maintenance, [ <a href="#">Nakajima, 1988</a> ] helps keep resources in a ready state and facilitates scheduling by decreasing unplanned downtime, thus increasing capacity.

	when process variability cannot otherwise be reduced.	Also, process improvements must precede automation; get rid of wasteful steps or dubious processes first.
9.	Look for simple, flexible, movable, and low-cost equipment that can be acquired in multiple copies—each assignable to a focused cell, flow line, or plant-in-a-plant.	Larger, faster, general-purpose equipment can detract from responsive customer service, especially over the longer run. A single fast process is not necessarily divisible across multiple customer needs. Simple, dedicated equipment is an economical solution; setup elimination is an added benefit.
10.	Make it easier to make/provide goods and services without error or process variation.	The aim is to prevent problems or defects from occurring—the fail-safing ( <i>pokayoke</i> ) idea—rather than rely on elaborate control systems for error detection and the ensuing rework. Strive to do it right the first time, every time.
11.	Cut cycle times, flow time, distance, and inventory all along the chain of customers.	Time compression provides competitive advantage [Blackburn, 1991]. Removal of excess distance and inventory aids quick response to customers. Less inventory also permits quicker detection and correction of process problems.
12.	Cut setup, changeover, get-ready, and startup times.	Setup (or changeover) time had been the standard excuse for large-lot operations prior to directed attention at reduction of these time-consuming activities [Shingo, 1985]. Mixed-model processing demands quick changeovers.
13.	Operate at the customer's rate of use (or a smoothed representation of it); decrease cycle interval and lot size.	Pull-mode operations put the customer in charge and help identify bottlenecks. Aim to synchronize production to meet period-by-period demand rather than rely on large lots and long cycle intervals.
14.	Record and <i>own</i> quality, process, and problem data at the workplace.	When employees are empowered to make decisions and solve problems, they need appropriate tools and process data. Transfer of point-of-problem data away from operators and to back-office staff inhibits responsive, operator-centered cures.
15.	Ensure that front-line associates get first chance at problem solving—before staff experts.	Ongoing process problems and on-the-spot emergencies are most effectively solved by teams of front-line associates; staff personnel are best used in advisory roles and for especially tough problems.
16.	Cut transactions and reporting; control <i>causes</i> , not symptoms.	Transactions and reports often address problem symptoms (e.g., time or cost variances) and delay action. Quick-response teams, using data-driven logic, directly attack problem causes and eliminate the need for expensive reporting.

---

Source: Schonberger, R.J. and Knod, E.M., Jr. 1994. *Operations Management: Continuous Improvement*, 5th ed. chapter 1. Richard D. Irwin, Burr Ridge, IL. Adapted with permission.

## 163.2 Scheduling

---

Basically, **scheduling** refers to the activities through which managers allocate capacity for the near future. It includes the assignment of work to resources, or vice versa, and the determination of

timing for specific work elements and thus answers the questions "who will do what" and "when will they do it." In manufacturing the scheduling time horizon is usually weekly, daily, or even hourly. In general, scheduling (1) flows from and is related to planning, (2) is determined by manufacturing environment, and (3) can be simplified when appropriate management principles are followed.

## Relationship to Planning

The planning activity also answers the "who," "what," and "when" questions, but in more general or aggregate terms for the longer time horizon—typically months, quarters, or years into the future. So, in the temporal sense, scheduling is short-term planning. But planning involves more. With the other creative activities, planning also addresses the characteristics of what is to be done (e.g., designs), quantity and variety (e.g., the product mix), how work will be accomplished (e.g., methods and procedures), utilization of funds (e.g., budgeting), and so on—including how scheduling itself will be accomplished. When planning (or design) has been thorough and things go according to plan, little creativity should be required in scheduling; it ought to be mostly an implementation activity.

In manufacturing, aggregate demand forecasts are filtered by business plans and strategies—what the company wants to do—to arrive at aggregate capacity and production plans. The master schedule states what the company plans to provide, in what quantities, and when. To the extent that on-hand or previously-scheduled inventories are insufficient to meet the commitments described by the master schedule, additional purchasing and production are required. Consequently, detailed planning and scheduling activities—for fabrication of components and subassemblies and for final assembly operations—come into play. Scheduling is among the production activity control duties that form the "back end" of the total manufacturing planning and control (MPC) system [Vollmann *et al.*, 1992]. Thus, it might be said that scheduling flows from planning.

## Effects of Manufacturing Environment

The type of manufacturing environment determines the nature of scheduling, as shown in Table 163.3. As the table notes, project and repetitive/continuous environments present less extreme scheduling problems; the sources listed at the end of this chapter contain suitable discussion. For project scheduling, see Evans [1993, ch. 18], Kerzner [1989, ch. 12], and Schonberger and Knod [1994, ch. 14]. For scheduling in repetitive or continuous production, see Schniederjans [1993, ch. 3 and 6] and Schonberger and Knod [1994, ch. 11 and 12].

**Table 163.3** Manufacturing Scheduling Overview

Manufacturing Environment	General Nature of Scheduling
Project	Activity scheduling and controlling (as well as overall project planning) may rely on Program Evaluation and Review Technique (PERT) or Critical Path Method (CPM). Large project complexity, cost, and uncertainties involved justify these tools. Smaller



Job or Batch	projects and single tasks may be scheduled with Gantt charts. Scheduling is time consuming due to low- or intermediate-volume output coupled with irregular production of any given item. Production typically occurs on flexible resources in functional or process layouts where considerable variation in products, routings, lot sizes, and cycle times-along with the competition among jobs (customers) for resource allocation-adds to the scheduling burden. Rescheduling may be common.
Continuous or Repetitive	Regular—if not constant—production on equipment dedicated to one or a few products (e.g., lines or cells) combine to decrease the scheduling problem. In process flow systems, scheduling is minimal except for planned maintenance, equipment or product changeovers, and so forth. In repetitive production, line balancing may be used to assign work; JIT's pull system, regularized schedules, and mixed-model scheduling can closely synchronize output with demand and can accommodate demand variation.

---

The variabilities inherent with traditional batch and job environments create the most complex scheduling (and control) problems. *Loading* (allocation of jobs to work centers), *sequencing* (determining job-processing order), *dispatching* (releasing jobs into work centers), *expediting* (rushing "hot" jobs along), and *reporting* (tracking job progress along routes) are among the specific activities. *Assignment models* can be of assistance in loading, and *priority rules* may be used for sequencing and dispatching. In batch operations, *lot splitting* and *overlapping* may help expedite urgent jobs. Unfortunately, however, throughput time in many job operations consists largely of queue time, setup time, move time, and other non-value-adding events. Perhaps even more unfortunate have been managers' attempts to "solve" the complexities of job scheduling and control by relying on more exotic scheduling and control tools.

## Effects of Contemporary Management Practices

This last section closes the discussion of scheduling by appropriately returning to the topic of management. In the 1970s North American managers allowed batch and job production scheduling and control systems to become too complicated, cumbersome, and costly. A more competitive approach lies in the simplification of the production environments themselves [Schonberger and Knod, 1994, ch. 13; Steudel and Desruelle, 1992, ch. 8; Schneiderjans, 1993, ch. 6]. Though necessary to some degree, scheduling itself adds no value to outputs; as such, it ought to be a target for elimination where possible and for streamlining when it must remain. Application of contemporary management practices, such as the principles detailed in Table 163.2, has been shown to improve scheduling, largely by removing the *causes* of the problems—that is, the factors that created a need for complicated and costly scheduling systems.

Steudel and Desruelle [1992] summarize such improvements for scheduling and related production control activities, especially in group-technology environments. Regarding scheduling, they note that manufacturing cells largely eliminate the scheduling problem. Also, sequencing is resolved at the (decentralized) cell level, and mixed-model assembly and kanban more easily handle demand variations. In similar fashion, manufacturing process simplifications foster improvements throughout the production planning and control sequence.

Just-in-time (JIT) management, for example, has been shown to greatly reduce the burden of

these activities, especially scheduling and control [Vollmann *et al.*, 1992, ch. 3 and 5]. Attempts to describe a "JIT scheduling system," however, are unnecessary; it is more productive to devote the effort to eliminating the need for scheduling at all. At this writing, it remains an oversimplification to suggest that the mere pull of customer demand is sufficient to *fully* schedule the factory, but that is a worthy aim. In sum, the less attention managers are required to devote to scheduling, the better.

## Defining Terms

**Management:** Activities that have the goal of ensuring an organization's competitiveness by creating, implementing, and improving capacity required to provide goods and services that customers want.

**Process:** A particular combination of resource elements and conditions that collectively cause a given outcome or set of results.

**Scheduling:** Activities through which managers allocate capacity for the immediate and near-term future.

## References

- Bateman, T. S. and Zeithaml, C. P. 1993. *Management: Function and Strategy*, 2nd ed. Richard D. Irwin, Burr Ridge, IL.
- Blackburn, J. D. 1991. *Time-Based Competition*. Business One-Irwin, Homewood, IL.
- Camp, R. C. 1989. *Benchmarking*. ASQC Quality Press, Milwaukee, WI.
- Evans, J. R. 1993. *Applied Production and Operations Management*, 4th ed. West, St. Paul, MN.
- Hall, R. 1983. *Zero Inventories*. Dow Jones-Irwin, Homewood, IL.
- Hammer, M. and Champy, J. 1993. *Reengineering the Corporation*. HarperCollins, New York.
- Imai, M. 1986. *Kaizen: The Key to Japan's Competitive Success*. Random House, New York.
- Kerzner, H. 1989. *Project Management: A Systems Approach to Planning, Scheduling, and Controlling*, 3rd ed. Van Nostrand Reinhold, New York.
- Nakajima, S. 1988. *Introduction to TPM: Total Productive Maintenance*. Productivity Press, Cambridge, MA.
- Schniederjans, M. J. 1993. *Topics in Just-in-Time Management*. Allyn and Bacon, Needham Heights, MA.
- Schonberger, R. J. 1982. *Japanese Manufacturing Techniques: Nine Hidden Lessons in Simplicity*. Free Press, New York.
- Schonberger, R. J. 1990. *Building a Chain of Customers*. Free Press, New York.
- Schonberger, R. J. and Knod, E. M., Jr. 1994. *Operations Management: Continuous Improvement*, 5th ed. Richard D. Irwin, Burr Ridge, IL.
- Shingo, S. 1985. *A Revolution in Manufacturing: The SMED [Single-Minute Exchange of Die] System*. Productivity Press, Cambridge, MA.
- Steudel, H. J. and Desruelle, P. 1992. *Manufacturing in the Nineties*. Van Nostrand Reinhold, New York.
- Suzaki, K. 1987. *The New Manufacturing Challenge: Techniques for Continuous*



*Improvement*. Free Press, New York.  
Vollmann, T. E., Berry, W. L., and Whybark, D. C. 1992. *Manufacturing Planning and Control Systems*, 3rd ed. Richard D. Irwin, Burr Ridge, IL.

## **Further Information**

### **Periodicals**

*Industrial Engineering Solutions*. Institute of Industrial Engineers.  
*Industrial Management*. Society for Engineering and Management Systems, a society of the Institute of Industrial Engineers.  
*Journal of Operations Management*. American Production and Inventory Control Society.  
*Production and Inventory Management Journal*. American Production and Inventory Control Society.  
*Production and Operations Management*. Production and Operations Management Society.  
*Quality Management Journal*. American Society for Quality Control.

### **Books**

Camp, R. C. 1995. *Business Process Benchmarking*. ASQC Quality Press, Milwaukee, WI.  
Orlicky, J. 1975. *Materials Requirements Planning*. McGraw-Hill, New York.  
Stonebraker, P. W. and Leong, G. K. 1994. *Operations Strategy*. Allyn & Bacon, Needham Heights, MA.

Chapman, W. L., Bahill, A. T. “Design, Modeling, and Prototyping”  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Design, Modeling, and Prototyping

---

164.1 The System Design Process

164.2 Rapid Prototyping

164.3 When to Use Modeling and Prototyping

**William L. Chapman**

*Hughes Aircraft Company*

**A. Terry Bahill**

*University of Arizona*

To create a product and the processes that will be used to manufacture it, an engineer must follow a defined system design process. This process is an iterative one that requires refining the requirements, products, and processes of each successive design generation. These intermediate designs, before the final product is delivered, are called **models** or **prototypes**.

A model is an abstract representation of what the final system will be. As such, it can take on the form of a mathematical equation, such as  $f = m \times a$ . This is a deterministic model used to predict the expected force for a given mass and acceleration. This model only works for some systems and fails both at the atomic level, where quantum mechanics is used, and at the speed of light, where the theory of relativity is used. Models are developed and used within fixed boundaries.

Prototypes are physical implementations of the system design. They are not the final design, but are portions of the system built to validate a subset of the requirements. For example, the first version of a new car is created in a shop by technicians. This prototype can then be used to test for aerodynamic performance, fit, drivetrain performance, and so forth. Another example is airborne radar design. The prototype of the antenna, platform, and waveguide conforms closely to the final system; however, the prototype of the electronics needed to process the signal often comprises huge computers carried in the back of the test aircraft. Their packaging in no way reflects the final fit or form of the unit.

### 164.1 The System Design Process

---

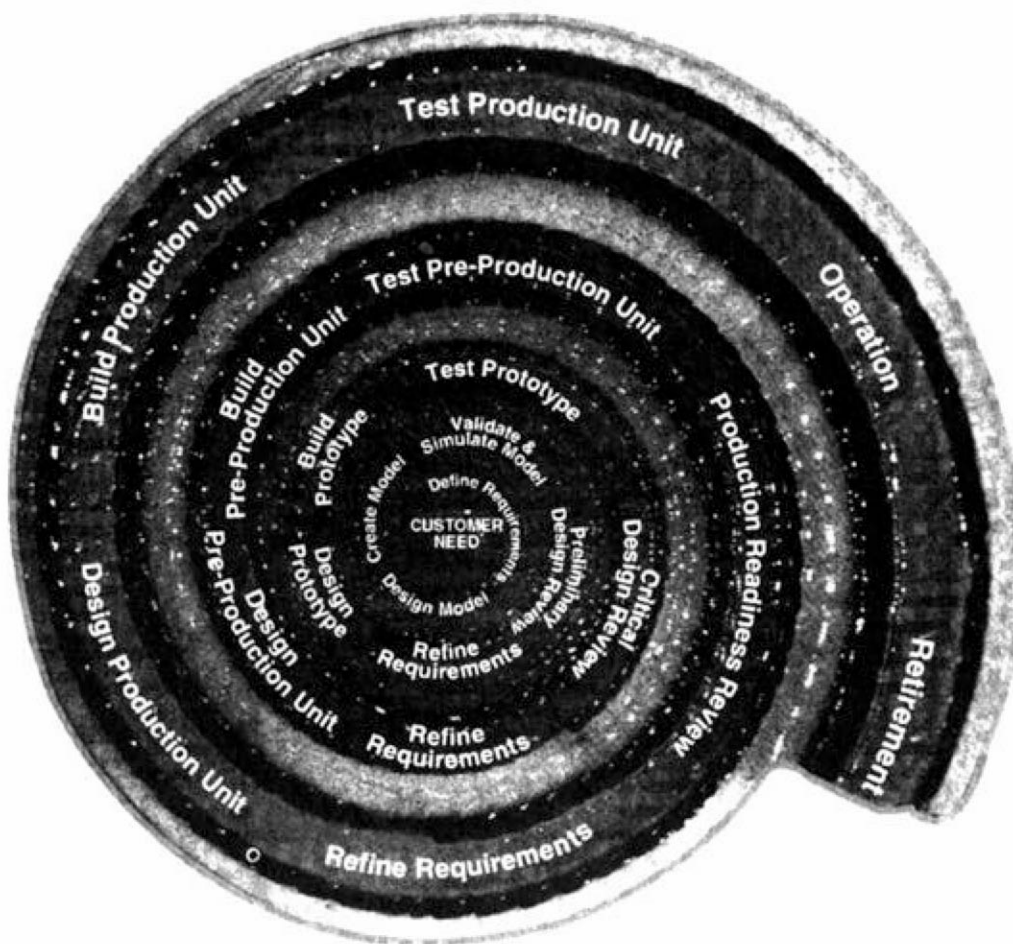
The system design process consists of the following steps.

1. Specify the requirements provided by the customer and the producer.
2. Create alternative system design concepts that might satisfy these requirements.

3. Build, validate, and simulate a model of each system design concept.
4. Select the best concept by doing a trade-off analysis.
5. Update the customer requirements based on experience with the models.
6. Build and test a prototype of the system.
7. Update the customer requirements based on experience with the prototype.
8. Build and test a preproduction version of the system and validate the manufacturing processes.
9. Update the customer requirements based on experience with the preproduction analysis.
10. Build and test a production version of the system.
11. Deliver and support the product.

This can be depicted graphically on a spiral diagram as shown in [Fig. 164.1](#).

**Figure 164.1** The system design process.



The process always begins with defining and documenting the customer's needs. A useful tool

for doing this is quality function deployment (QFD). QFD has been used by many Japanese and American corporations to document the voice of the customer. It consists of a chart called the "house of quality." On the left is listed what the customer wants. Across the top is how the product will be developed. These are often referred to as *quality characteristics*. The "whats" on the left are then related to the "hows" across the top, providing a means of determining which quality characteristics are the most important to the customer [Akao, 1990; Bahill and Chapman, 1993].

After the customer's needs are determined the design goes through successive generations as the design cycle is repeated. The requirements are set and a model or prototype is created. Each validation of a model or test of a prototype provides key information for refining the requirements.

For example, when designing and producing a new airborne missile, the initial task is to develop a model of the expected performance. Using this model, the systems engineers make initial estimates for the partition and allocation of system requirements. The next step is to build a demonstration unit of the most critical functions. This unit does not conform to the form and fit requirements but is used to show that the system requirements are valid and that an actual missile can be produced. Requirements are again updated and modified, and the final partitioning is completed. The next version is called the *proof-of-design* missile. This is a fully functioning prototype. Its purpose is to demonstrate that the design is within specifications and meets all form, fit, and function requirements of the final product. This prototype is custom-made and costs much more than the final production unit will cost. This unit is often built partly in the production factory and partly in the laboratory. Manufacturing capability is an issue and needs to be addressed before the design is complete. More changes are made and the manufacturing processes for full production readied. The next version is the proof of manufacturing or the preproduction unit. This device will be a fully functioning missile. The goal is to prove the capability of the factory for full-rate production and to ensure that the manufacturing processes are optimum. If the factory cannot meet the quality or rate production requirements, more design changes are made before the drawings are released for full-rate production. Not only the designers but the entire design and production team must take responsibility for the design of the product and processes so that the customer's requirements are optimized [Chapman *et al.*, 1992]. Also see **Chapters 80, 160, and 161** of this book for additional information on design and production.

Most designs require a model upon which analysis can be done. The analysis should include a measure of all the characteristics the customer wants in the finished product. The concept selection will be based on the measurements done on the models. See **Chapter 188** for more information on selection of alternatives. The models are created by first partitioning each conceptual design into functions. This decomposition often occurs at the same time that major physical components are selected. For example, when designing a new car, we could have mechanical or electronic ignition systems. These are two separate concepts. The top-level function—firing the spark plugs—is the same, but when the physical components are considered the functions break down differently. The firing of the spark plugs is directed by a microprocessor in one design and a camshaft in the other. Both perform the same function, but with different devices. Determining which is superior will be based on the requirements given by the customer, such as cost, performance, and reliability. These characteristics are measured based on the test criteria.

When the model is of exceptional quality a prototype can be skipped. The advances in

computer-aided design (CAD) systems have made wire-wrapped prototype circuit boards obsolete. CAD models are so good at predicting the performance of the final device that no prototype is built. Simulation is repeated use of a model to predict the performance of a design. Any design can be modeled as a complex finite state machine. This is exactly what the CAD model of the circuit does. To truly validate the model each state must be exercised. Selecting a minimum number of test scenarios that will maximize the number of states entered is the key to successful simulation. If the simulation is inexpensive, then multiple runs of this model should be done. See **Chapters 13,29,95,96,97, and 159** for more information on simulations. The more iterations of the design process there are, the closer the final product will be to the customer's optimum requirements.

For other systems, modeling works poorly and prototypes are better. Three-dimensional solids modeling CAD systems are a new development. Their ability to display the model is good, but their ability to manipulate and predict the results of fit, force, thermal stresses, and so forth is still weak. The CAD system has difficulty simulating the fit of multiple parts (such as a fender and car frame) because the complex surfaces are almost impossible to model mathematically. Therefore, fit is still a question that prototypes, rather than models, are best able to answer. A casting is usually used to verify that mechanical system requirements are met.

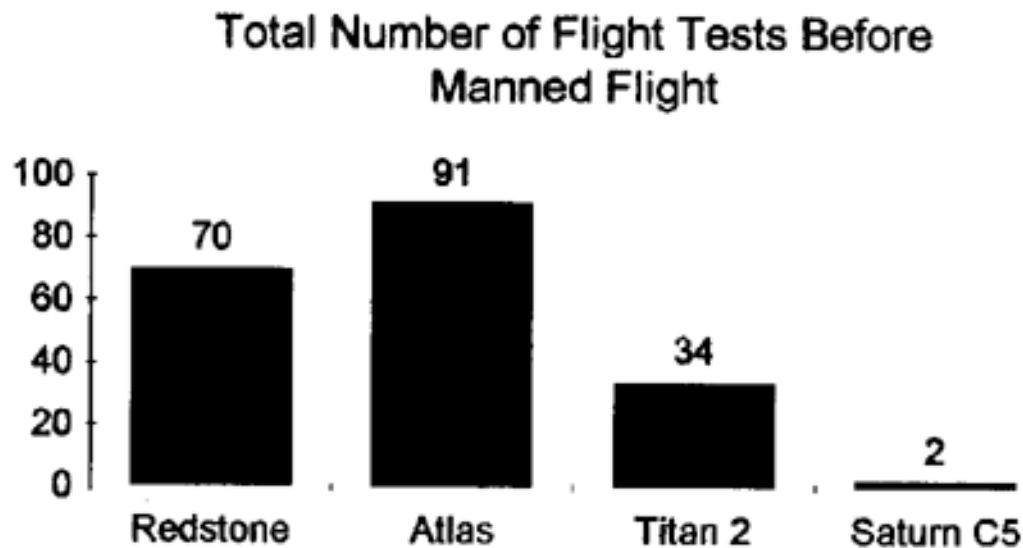
Computer-aided manufacturing (CAM) systems use the CAD database to create the tools needed for the manufacture of the product. Numerical control (NC) machine instructions can be simulated using these systems. Before a prototype is built, the system can be used to simulate the layout of the parts, the movement of the cutting tool, and the cut of the bar stock on a milling machine. This saves costly material and machine expenses.

Virtual reality models are the ultimate in modeling. Here the human is put into the loop to guide the model's progress. Aircraft simulators are the most common type of this product. Another example was demonstrated when the astronauts had to use the space shuttle to fix the mirrors on the Hubble telescope. The designers created a model to ensure that the new parts would properly fit with the existing design. They then manipulated the model interactively to try various repair techniques. The designers were able to verify fit with this model and catch several design errors early in the process. After this, the entire system was built into a prototype and the repair rehearsed in a water tank [[Hancock, 1993](#)].

Computer systems have also proved to be poor simulators of chemical processes. Most factories rely on design-of-experiments (DOE) techniques, rather than a mathematical model, to optimize chemical processes. DOE provides a means of selecting the best combination of possible parameters to alter when building the prototypes. Various chemical processes are used to create the prototypes that are then tested. The mathematical techniques of DOE are used to select the best parameters based on measurements of the prototypes. Models are used to hypothesize possible parameter settings, but the prototypes are necessary to optimize the process [[Taguchi, 1976](#)].

The progressive push is to replace prototypes with models, because an accurate fully developed model is inexpensively simulated on a computer compared to the cost and time of developing a prototype. A classic example is the development of manned rockets. [Figure 164.2](#) shows the number of test flights before manned use of the rockets.

**Figure 164.2** Flight tests for the manned space program rockets.



The necessity for prototypes diminished rapidly as confidence in computer models developed. Initially, many of the rockets exploded in their first attempts at launch. As more was learned about rocketry, sophisticated models were developed that predicted performance of the rocket based on real-time measurements of valves, temperatures, fuel levels, and so forth. Using modern computers the entire model could be evaluated in seconds and a launch decision made. This eliminated the need for many flight tests and reduced the cost of the entire Apollo moon-landing program.

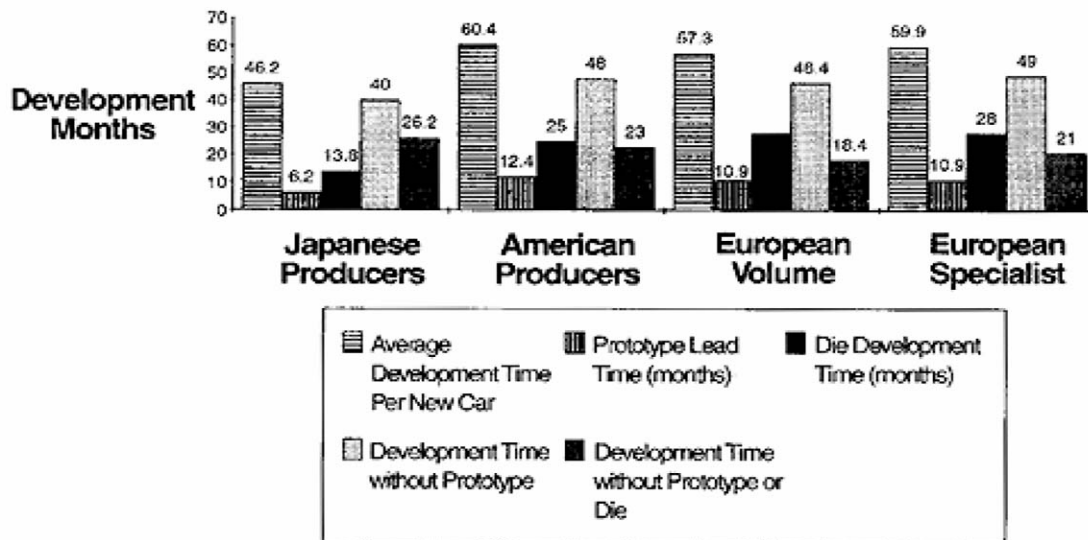
## 164.2 Rapid Prototyping

Rapid prototyping is the key to reducing design time for parts and processes. Design is an iterative process. By creating prototypes quickly, the design can be completed faster. Japanese automobile manufacturers develop prototypes of their products in 6 months, whereas American companies take 13 months [Womack *et al.*, 1990]. This advantage allows the Japanese companies to get to the market faster, or if they choose they can iterate their design one more time to improve their design's conformance to the customer's requirements. Japanese automakers' ability to create the prototype quickly is due in part to better coordination with their suppliers, but also to the exacting use of design models to ensure that the design is producible.

As seen in Fig. 164.3, the lead in prototype development accounts for 44% of the advantage the Japanese producers have in product development time. The rest of the advantage is from rapid creation of the huge dies needed to stamp out the metal forms of the automobiles. The design of these important manufacturing tools is given as much attention as the final product. By creating flexible designs and ensuring that the teams that will produce the dies are involved in the design process, the die development time is cut from 25 months in the U.S. to 13.8 months in Japan. The rapid creation of the tooling is a key to fast market response.



**Figure 164.3** Comparison of Japanese, American, and European car producers. (Based on Womack, J. P., Jones, D. T., and Roos, D. 1990. *The Machine that Changed the World*. Rawson Associates, New York.



**Stereolithography** is a new method of creating simple prototype castings. The stereolithography system creates a prototype by extracting the geometric coordinates from a CAD system and creating a plastic prototype. The solids model in the CAD system is extracted by "slicing" each layer in the  $z$  axis into a plane. Each layer is imaged by a laser into a bath of liquid photopolymer resin that polymerizes with the energy from the laser. Each plane is added, one on top of the other, to build up the prototype. The final part is cured and polished to give a plastic representation of the solids model. It illustrates the exact shape of the part in the CAD system. This technique will not, however, verify the function of the final product because the plastic material used will not meet strength or thermal requirements [Jacobs, 1992].

Software developers also use rapid prototyping. This technique is used to get a fast though barely functional version of the product into the customer's hands early in the design cycle. The prototype is created using the easiest method to simulate functionality to the viewer. The customer comments on what is seen and the developers modify their design requirements. For example, when developing expert systems, models are almost never used. One of the driving rules is to show a prototype to the customer as soon as possible and afterwards throw it away! The purpose of building the prototype was to find out what the knowledge that needs to be represented is like, so that the appropriate tool can be selected to build the product. If, as is usually the case, a nonoptimal tool was used for the prototype, then the prototype is thrown away and a new one is developed using better tools and based on better understanding of the customer's requirements. Beware, though—often a key function displayed in the prototype is forgotten when the prototype is abandoned. Be certain to get all the information from the prototype [Maude and Willis, 1991]. The fault with this technique is that the requirements are not written down in detail. They are incorporated into the code as the code is written, and they can be overlooked or omitted when transferred to a new system.

## 164.3 When to Use Modeling and Prototyping

When should modeling versus prototyping be used? The key difference is the value of the



information obtained. Ultimately, the final product must be created. The prototype or model is used strictly to improve the final product. Costs associated with the prototype or model will be amortized over the number of units built. The major problem with models is the lack of confidence in the results. Sophisticated models are too complex for any single engineer to analyze. In fact, most models are now sold as proprietary software packages. The actual algorithms, precision, and number of iterations are rarely provided. The only way to validate the algorithm (not the model) is by repeated use and comparison to actual prototypes. It is easier to have confidence in prototypes. Actual parts can be measured and tested repeatedly, and the components and processes can be examined. Prototypes are used more often than models once the complexity of the device exceeds the ability of the computer to accurately reflect the part or process. During the initial design phases, models must be used because a prototype is meaningless until a concept has been more firmly defined. At the other extreme, modeling is of limited benefit to the factory until the configuration of the part is well known. A general rule is to build a model, then a prototype, then a production unit. Create even a simple mathematical model if possible so that the physics can be better understood. If the prototype is to be skipped, confidence in the model must be extremely high. If there is little confidence in the model, then a minimum of two or three prototype iterations will have to be done.

## Defining Terms

**Model:** An abstract representation of what the final system will be. It is often a mathematical or simplified representation of a product.

**Prototype:** A physical representation of a product built to verify a subset of the system's requirements.

**Stereolithography:** A prototype-manufacturing technique used to rapidly produce three-dimensional polymer models of parts using a CAD database.

## References

- Akao, Y. (Ed.) 1990. *Quality Function Deployment: Integrating Customer Requirements into Product Design*. Productivity Press, Cambridge, MA.
- Bahill, A. T. and Chapman, W. L. 1993. A tutorial on quality function deployment. *Eng. Manage. J.* 5 (3):24–35.
- Chapman, W. L., Bahill, A. T., and Wymore, A. W. 1992. *Engineering Modeling and Design*. CRC Press, Boca Raton, FL.
- Hancock, D. 1993. Prototyping the Hubble fix. *IEEE Spectrum*. 30 (10):34–39.
- Jacobs, P. F. 1992. *Rapid Prototyping & Manufacturing: Fundamentals of StereoLithography*. McGraw-Hill, New York.
- Maude, T. and Willis, G. 1991. *Rapid Prototyping: The Management of Software Risk*. Pitman, London.
- Taguchi, G. 1976. *Experimental Designs*, 3rd ed., vols. 1 and 2. Maruzen, Tokyo.
- Womack, J. P., Jones, D. T., and Roos, D. 1990. *The Machine that Changed the World*. Rawson Associates, New York.

## **Further Information**

Pugh, S. 1990. *Total Design: Integrated Methods for Successful Product Engineering*. Addison-Wesley, London.

Suh, N. P. 1990. *The Principles of Design*. Oxford University Press, New York.

Wymore, A. W. 1993. *Model-Based Systems Engineering*. CRC Press, Boca Raton, FL.

Risbud, S. H. "Materials Processing and Manufacturing Methods"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Materials Processing and Manufacturing Methods

---

## 165.1 Processing Metals and Alloys

Casting Methods • Metal-Working Methods • Machining and Finishing

## 165.2 Ceramics, Glasses, and Polymers

Powder Processing and Sintering • Glasses and Glass-Ceramics • Sol-Gel Methods • Polymer Processing

## 165.3 Joining of Materials

### **Subhash H. Risbud**

*University of California, Davis*

One of the most critical aspects of engineering applications involves the selection, shaping, and processing of raw materials obtained either directly from the earth's crust or by special synthesis and purification methods. The history of successful cost competitive technologies and the development of modern societies is closely linked to one of two factors: (1) a new processing method for manufacturing a material with broad engineering utility, or (2) an altogether new discovery of a material with a composition that makes possible dramatically improved engineering properties. A good example of a processing method that can revolutionize engineering technologies is the now well-known discovery of zone refining and crystal growth of semiconductors, such as silicon and germanium, which are at the heart of the present day microelectronic industries. Similarly, the discovery of new materials is best exemplified by the finding in the mid-1980s that certain ceramics (mixtures of yttrium-barium-copper oxides)—long known to be ferroelectric—can actually lose all resistance to electric current and become superconductors at a relatively warm temperature of about 90 K. This discovery of a new material is likely to have a major impact in many fields of technology.

Engineers and scientists continue the rich historical trend of seeking new materials or an unusual method for shaping/processing a material. A particularly striking evidence of this is found in many of the current worldwide activities in the fast-growing field of nanotechnology. Thus, spectacular engineering properties are attainable by making extremely small structures that are many times smaller than the thickness of human hair. The principal goal is to synthesize nanometer-grained bulk materials or composite films in which at least one functional material has dimensions somewhere between 1 and 50 nm. Alternately, methods are being considered to carve out of bulk materials nanosize structures that promise to have a revolutionary impact on engineering and the physical and biological sciences. Examples of such materials and methods include the nanotechnology of making quantum wells and dots for laser optics, micromachined gears, ductile

ceramics, and molecular motors. This chapter intends to provide a "roadrunner's" view of methods for shaping and processing materials both by the well-established techniques (e.g., casting, welding, powder sintering) and the more recent evolving methods.

## 165.1 Processing Metals and Alloys

---

The major methods for making engineering parts from metallic alloys consist of casting from the molten (liquid) state, followed by cold or hot working of the casting by processes such as rolling, forging, drawing, extrusion, and spinning. Powders of metals and alloys are also made into shapes by pressing and sintering or hot pressing (HP) and hot isostatic pressing ("hipping"). Grinding, polishing, and welding are some of the final steps in the production of an engineering component. There is a growing interest in encouraging near-net-shape manufacturing methods for engineering components to minimize the wastage in machining and grinding operations. Powder metallurgy methods provide an important advantage in this regard.

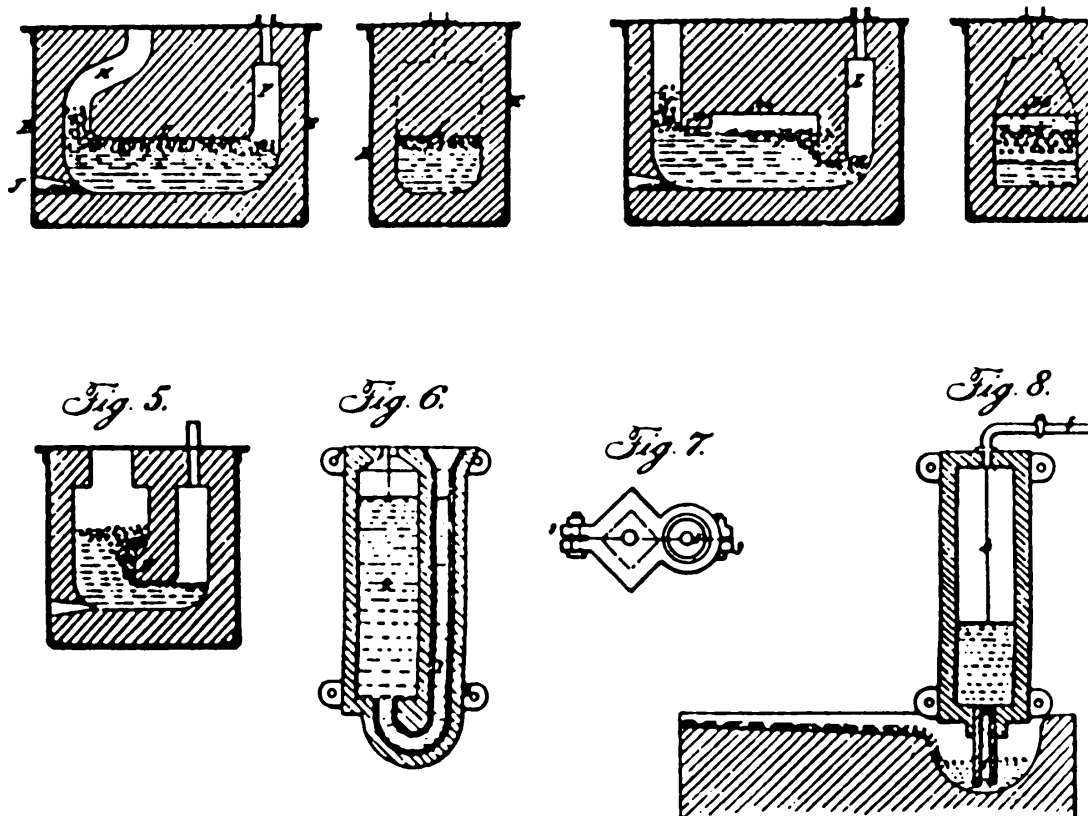
### Casting Methods

**Casting** of liquid metals and alloys is one of the oldest and best-known processes for making ingots, or shapes in final form determined by the mold or die cavity in which the liquid is poured. The solidification of the liquid metal takes place after the melt is allowed to flow to all the areas of the mold. Shrinkage during solidification makes for a smaller casting size than the mold size, and this is accounted for in the design of the mold. Molds can vary from simple cavities to multipart molds clamped together and containing ceramic cores to create hollow spaces in some parts of the casting.

Sand mixed with clay is the best-known traditional mold material and is packed around a wooden pattern that is subsequently removed when the sand casting mold is ready for pouring of liquid metal. A small channel called a *gate* provides the liquid metal access to the mold cavity, while a riser column becomes the reservoir of liquid metal that feeds the shrinkage of the liquid metal during solidification. The sand-casting process is an inexpensive method for making a small number of parts; a range of ferrous and nonferrous alloys can be made by this method. The rough finish of sand cast parts usually requires further machining and grinding.

Die casting is performed by pouring liquid alloys into preshaped metal molds machined to good accuracy and final tolerance. Thus, the finished product is of good surface quality in die casting, and complex shapes can be made. Gravity die casting involves the flowing of the liquid metal under natural gravity conditions, whereas pressure die casting involves forcing the melt into the die under pressure. Centrifugal casting relies on a rotating mold to distribute the melt in the mold cavity by the action of centrifugal forces, a method specially suitable for making large-diameter metal pipes. Aluminum, zinc, copper and magnesium alloys are often made by die-casting methods. Because machined molds are required, die casting is more expensive than sand castings, although molds are reused, and thus the process is good for large-scale production. Investment casting or lost wax casting is specially used for alloys with high melting temperatures and when high dimensional accuracy and precision are needed. The high-melting metals attack the dies used in normal die casting and thus patterns are produced from wax using a metal mold. The wax

patterns are coated with a ceramic slurry and heated to melt the wax, whereupon a dense ceramic mold is created. Molten alloys are then forced into the ceramic mold cavity by use of pressure or by a centrifugal process. Complicated shapes (e.g., aerospace turbine blades) or shapes where fine detail is sought (e.g., figurines) are made by investment or lost wax casting methods.



#### IMPROVEMENT IN THE MANUFACTURE OF IRON AND STEEL

*Henry Bessemer*

*Patented July 25, 1865*

*#49,051*

To refine crude or so-called pig iron into steel, Englishman Bessemer found that jets of air or steam injected into molten iron would burn up the carbon impurities and decarbonize the iron until it became usable steel. The heat from the decarbonization also helped keep the metal molten so it could be poured into ingots for later processing into bars, plates and rods. His first patent for

making steel was granted in 1856 with many more including this one granted for improvements in the "blast furnace" process.

Patent #49,051 improved the process by causing air or steam to decarbonize the iron without the use of nozzles that could clog as in previous versions of the process. It also specified using partially evacuated airtight molds so the metal ingots would be less cellular, an undesirable defect. (©1993, DewRay Products, Inc. Used with permission.)

## Metal-Working Methods

Hot working of metal alloys is normally the next step in the shaping of the cast alloy into a component and is also used to induce some refinement of the grain structure of the material to obtain the necessary ductility or strength. In hot working the temperature is high enough that recrystallization occurs at a rate faster than work hardening due to deformation. Forging, rolling, and extrusion are some examples of hot-working processes.

Hot forging is a process of simply pressing the hot metal between two surfaces in open, closed, or flat-faced dies. This process is closest to the ancient art of the blacksmith, but modern manufacturing plants employ high-power hydraulic presses to deliver impact blows to achieve forging of large steel crankshafts, aerospace propellers, and so forth. Hot extrusion consists of taking a cylindrical billet from the cast ingot and placing the hot billet in a die of a slightly larger size. A piston or ram extrudes the hot metal through appropriately cut orifices or openings in the die. The process is akin to squeezing toothpaste out of a tube under pressure. Direct and inverted extrusion are the two most common variations of the extrusion process. Hot rolling is a popular method of reducing the cast ingots into billets, slabs, plates, or sheets by pressing the hot metal between rotating rollers. Four high-rolling mills consist of smaller-diameter work rolls backed by larger diameter backup rolls. Many other types of cluster mills also exist, including planetary mills and pendulum mills.

Cold working of metals and alloys has the advantage of producing a clean, smooth surface finish and a harder material. Cold rolling of aluminum foil or steel strips are some examples of products in everyday use that have surface finish that can vary from a standard bright finish to matte and plated finishes, depending on the application. Drawing is another cold-working process in which cylindrical rods can be drawn through a die to make wires or sheets of metal can be shaped into cup-shaped objects by deep drawing the sheet into a die by pressure from the punch. By clamping the sheet around the edges, the pressing operation can produce domestic cookware, automobile bodies, and a variety of cup-like shapes. Circular cross sections can be made by spinning a metal blank in a lathe while applying pressure to the blank with a tool to shape it. Explosive forming is a special process used for shaping large-area sheets of metal into contoured panels such as those needed in communication reflectors. The wave generated by detonation of an explosive charge is used to convert a sheet blank into the shape of the die. [Table 165.1](#) is a summary of the processes used with a listing of the advantages and limitations of each process.

**Table 165.1** Summary of Metal-Working Processes

Process	Economic Quantity	Typical Materials	Optimum Size
Sand casting	Small–large	No limit	1–100 kg
Die casting—gravity	Large	Al, Cu, Mg, Zn alloys	1–50 kg
Die casting—pressure	Large	Al, Cu, Mg, Zn alloys	50 g–5 kg
Centrifugal casting	Large	No limit	30 mm–1 m diam.
Investment casting	Small–large	No limit	50 g–50 kg
Closed-die forging	Large	No limit	3000 cm <sup>3</sup>
Hot extrusion	Large	No limit	500 mm diam.
Hot rolling	Large	No limit	—
Cold rolling	Large	No limit	—
Drawing	Small–large	Al, Cu, Zn, mild steel	3 mm–6 mm diam.
Spinning	One-off-large	Al, Cu, Zn, mild steel	6 mm–4.5 m diam.
Impact extrusion	Large	Al, Pb, Zn, mild steel	6 mm–100 mm diam.
Sintering	Large	Fe, W, bronze	80 g–4 kg
Machining	One-off-large	No limit	—

Adapted from Bolton, W. 1981. *Materials Technology*, p. 56. Butterworth, London. With permission.

## Machining and Finishing

The cast, hot- or cold-worked metallic alloys are often subjected to some type of **machining** operation, which encompasses cutting, grinding, polishing, shearing, drilling, or other metal-removal methods. Chemical etching or polishing methods can also be used as finishing steps in the production of components from metals. Planing, turning, milling, and drilling are the main methods of removing chips from a metal part during the machining step. Surface finish and dimensional accuracy of the component are the main goals of machining. Cutting and grinding tools required in machining are usually made from hard metal (carbide) or ceramic materials; *machinability index* measures the ease with which a particular material can be machined with conventional methods. The higher the machinability index is, the easier it is to machine the metal or alloy. In addition to these methods, metal removal can also be achieved by ultrasonic abrasion, laser or electron beam cutting and drilling, electrodischarge or spark machining, arc milling, and chemical milling. Numerically controlled machining of high-precision components and sophisticated robotically controlled fabrication methods continue to be investigated and applied in the manufacture of engineering materials and components. Powder metallurgy methods are most suitable for directly converting powders of raw materials into finished shapes with little or no machining. This method, discussed later, has the greatest potential for near-net-shape manufacturing and is commonly used to consolidate metal, ceramic, and mixed composite powders.

## 165.2 Ceramics, Glasses, and Polymers

Nonmetallic materials such as ceramics, glasses, and polymeric organic materials require generally different methods for shaping and manufacturing than metals, although some common methods do



exist for all of these materials. Inorganic silicate glasses and polymer melts are processed by methods that take into account their viscoelastic properties. Ceramics are normally refractory and high-melting materials and thus require powder sintering, hot pressing or related methods for production.

## Powder Processing and Sintering

Powder sintering of metal alloy powders and ceramics is a versatile method of producing near-final shapes of a component. Ceramic powders are typically blended with solid or liquid additives, binders, and so forth, and the batched system is treated by a series of mixing, chemical dissolution, washing, de-airing, or filtration to produce a slurry, paste, or suspension of solid particles in a liquid, often referred to as the *slip*. Depending on the consistency (viscosity of the syrup) of the suspension, forming a shape is achieved by tape casting onto plastic sheets or by pouring the slip into a plaster mold, followed by drying to remove the residual liquids in the batch. The shaped porous ceramic at this stage is called a *green ceramic body*. Green machining, surface grinding, or application of coatings or glazes is done at this stage. The final step in the process is consolidation to a nearly pore-free solid by firing or sintering at a high temperature. The sintered microstructure is usually a multiphase combination of grains, secondary phases, and some degree of remaining porosity.

Another common way of sintering ceramics, and particularly metal powders, is by direct pressing of the powders with some solid additives (called *sintering aids*). The component shape is determined by a die cavity and the powders are cold pressed at relatively high pressures, followed by heat treatment to create the sintered microstructure. In a combination of these steps, pressure and heat can be applied simultaneously in hot pressing or hot isostatic pressing ("hipping"). In these methods the powder fills in the die cavity, and, as the pressure is applied by hydraulic rams, the die is heated simultaneously to consolidate the powder to a solid. The evolution of a microstructure from powders to dense ceramics with grains and grain boundaries is now a well-developed science-based technology. Thus, deliberate process control is needed, including awareness of atmospheres used in firing, particle size distribution, chemistry of additives, control over grain boundary mobility, and elimination of exaggerated grain growth. The recent development of fast sintering methods includes the use of sparks, microwaves, and plasmas to activate the surfaces of very fine powders (micrometers to as small as a few nanometer particles) and cause consolidation to occur in just a few minutes. Microwave sintering and plasma-activated sintering are examples of promising new methods on the horizon for very fast sintering to achieve small grain sizes and high-purity materials in both metallic and ceramic systems.

## Glasses and Glass-Ceramics

The manufacture of glasses from mixtures of silicon-dioxide (silica) and a number of other oxides (e.g., sodium, calcium, aluminum oxides) is based on melting the mixtures in a tank-like furnace to create a homogeneous bubble-free viscous melt. The composition of the glass is chosen such that, on cooling the melt during shaping, formation of the thermodynamically favored crystalline

material does not occur. To make glass plates or sheets (e.g., car windshields), the molten liquid is allowed to float onto the surface of a liquid bath of low-melting metal that is immiscible with the glass (the lighter glass floats on the higher-density metal melt). This float glass process is but one of a variety of ways of converting molten silicate liquids to glassy shapes. Viscous glass streaming out of a furnace can be trimmed into "gobs" that drop into a preshaped die while blowing air under pressure to force the mushy viscous gob into the final shape (e.g., a beer bottle). Molten glass compositions can also flow through bushings containing small orifices to create fiberglass materials or optical fibers. Optical fibers of high purity are also processed in drawing towers using preforms of carefully engineered clad-and-core compositions. Thousands of glass compositions with specific chemical ingredients are known for generating color, inducing a refractive index gradient, permitting selective transparency, or resisting weathering. The common window glass composition is based on a mixture of the oxides of sodium, calcium, and silicon and is popularly called soda-lime-silica glass.

Glass-ceramics are materials made by the heat treatment of a shaped piece of glass so as to allow the formation of copious amounts of small crystallites without changing the macroscopic dimensions of the object. The optical transparency of the glass changes to a translucent appearance typical of most glass-ceramics, best known in the form of domestic cookware sold under the trade name Corelle or Corningware, by Corning Glass Company. The glass-ceramic process is a useful method of obtaining fully dense ceramics from the glassy state and requires good control of the nucleating agents that are used to catalyze the crystallite formation in the glass-ceramic microstructure. The uniformity of the crystal grains in the structure and the small size are factors that give glass-ceramics the higher strength and better mechanical properties compared to glasses. The number of glass-ceramics available for use in engineering practice has grown enormously in the last 25 years; applications of glass-ceramics in electrical, magnetic, structural, and optical fields is becoming more common. Machinable glass-ceramics based on a mica-like crystalline phase have made it easier to shape glass-ceramics by conventional machine shop tools.

## HOW ALUMINUM IS MADE

### **Mining**

Bauxite, the ore rich in aluminum oxide from which most aluminum is produced, is found and mined widely around the world. Today, most bauxite mining locations are in the Caribbean area, Brazil, Australia, and Africa. Bauxite is the product of the chemical weathering of rocks containing aluminum silicates. The weathering process takes millions of years.

### **Refining**

Bauxite contains many impurities that are removed by heat and chemicals in the process known as refining. In this process, bauxite is ground, mixed with lime and caustic soda, pumped into high-pressure containers, and heated. The aluminum oxide is dissolved by the caustic soda, leaving undesired materials as a solid. The mixture is filtered to remove the solids and a crystallized material called hydrated alumina is made to drop out of the solution in a process known as seeding. The material is washed and heated to drive off water and produce a white

powder (resembling sugar) called alumina ( $\text{Al}_2\text{O}_3$ ), another name for aluminum oxide.

### **Smelting**

Aluminum is made from alumina powder by an electrolytic reduction process known as smelting. To get aluminum metal, alumina must be "reduced" to separate the aluminum and the oxygen. This is done by dissolving powdered alumina in a cryolite bath inside large, carbon-lined electrolytic reduction cells called pots. When a powerful direct electric current is passed through the bath, aluminum metal separates from the chemical solution. Heat generated by the electricity keeps the bath liquid, so more alumina can be added to make the process continuous. As liquid aluminum builds up on the bottom of the pots, it is siphoned off. The pots used in smelting are up to 15 feet wide and 40 feet long. They are electrically connected in series so that they all work at the same time. As many as 250 pots have been linked in a single potline. Most aluminum smelting plants have several potlines.

### **Fabricating**

Molten aluminum is removed from smelting pots in large containers called crucibles, then poured into furnaces for mixing with other metals in controlled amounts to make it stronger or to give it other special properties required for ultimate use. This process, known as alloying, takes place at temperatures of about 1220°F, the melting point of aluminum. In the furnace, aluminum is purified in a process called fluxing—done by forcing mixtures of gases through the hot metal. Impurities float to the surface and are removed. The aluminum then is poured into molds or cast directly into solid forms, called ingots, and worked.

**Direct chill casting**, a process invented by Alcoa, is the first step in many metal-working operations. It is done by rapid cooling with water to "freeze" molten aluminum as it passes through a mold. The result is aluminum of constant high quality, usually in rectangular or round ingots.

**Electromagnetic casting** is a moldless process that operates on the principle that current passing through a copper conductor induces an opposing current in molten aluminum. The current produces magnetic fields in both the aluminum and the copper conductor, which repel each other. Molten aluminum is thus suspended away from the copper conductor by a magnetic field. Without a mold, the hot ingot surface is in continuous contact with cooling water during solidification and a very smooth ingot surface is produced.

**Sheet and plate** ingots are often 30 feet long and weigh up to 20 tons. More operations are needed to roll sheet than the thicker plate.

**Foil** is made from aluminum sheets passed between rolls until very thin. Some is rolled to under .0002 of an inch thick, about the thickness of human hair.

**Forging**, the ancient art of pressing metal into shapes, is still widely used. Heated ingots are in some cases hammered to make precision and high-strength aluminum parts. The largest forgings are made on presses with forces up to 50,000 tons.

**Drawn bar and rod** products are pulled through similarly shaped dies, each progressively smaller in diameter. When rod is drawn to less than 3/8 of an inch, it is called wire. Wire is stranded into cable to carry electricity.

**Extrusions** are made by heating ingots until they can be easily shaped with a powerful press. An extrusion ingot (called a billet) is pushed through a shaped die just as toothpaste is squeezed from a tube. Over 100,000 shapes—some simple, some complex—have been extruded in

aluminum.

***Drawn tube*** products are pulled through a die that controls outside measurements while an insert called a mandrel controls inside measurements. This drawing process is repeated until the tube is the desired size. (Courtesy of the Aluminum Company of America.)

## Sol-Gel Methods

**Sol-gel processes** make it possible to use organic-polymeric synthesis methods to produce ceramics and glasses at relatively low temperatures. The major advantage of sol-gel techniques is that all the chemicals needed in a glass or ceramic can be blended into a solution to create a "sol" at close to room temperature. This avoids the high-temperature processing normally used in making a ceramic or a glass. The sol is then made to gel by appropriate hydrolysis reactions, and the gel is then fired into a monolithic body or made into coatings, powders, films, and so forth. The calcination and sintering steps remove water and vestiges of organic bonds to form an inorganic ceramic product. Not all ceramics can be made by sol-gel techniques—not because of any fundamental limitations, but due mainly to the lack of well-developed connections between a suitable starting precursor and the synthesis protocol to make the final ceramic or glass. For making bulk monoliths or coatings and films, the use of drying control agents in sol-gel processing is crucial to attain crack-free objects.

## Polymer Processing

The forming of organic polymeric materials essentially consists of making powders or granules of the requisite blend of polymers with additives such as carbon, chalk, and paper pulp. In the next step the granules are heated to soften the polymer mixture, and the soft material is viscoelastically shaped into the final component. The processing methods mainly involve extruding, injection molding, casting, and calendering. The particular process used depends on the quantity, size, and form in which the polymer object is desired (e.g., sheets of plastic or bulk rods). Thermosetting polymers go through an irreversible chemical change upon heating and they are mostly made by molding or casting. Examples include thermosetting resins such as urea, melamine or phenol formaldehydes, and polyesters. Thermoplastic polymers are processible by repeated softening as long as the temperature is not too high to cause decomposition and charring. Ease of flow makes these attractive for processing by injection molding and extrusion methods. Examples include polycarbonate, polyethylene, polystyrene, polyamide (nylon), and polyvinyl chloride (PVC). Inclusion of gas gives rise to foamed plastics such as polystyrene foam (for thermal insulation), whereas polymers are reinforced by fibers (typically glass fibers) to form composites with better stiffness and strength. The glass fiber polymer composites are shaped by injection molding and are useful in environments where moisture-resistant, high-fracture toughness materials are needed. Continuous sheets of plastic are made from PVC or polyethylene polymers by calendering. This process consists of feeding the heated polymer granules or powder through a set of heated rollers. The first roller converts the granules to a sheet that is subsequently reduced to the appropriate

thickness and surface finish. Laminates and coatings of polymers are formed by pressing, casting, and roll-coating technologies. Large polymer pieces such as boat hulls are put together manually by pasting several layers of gum-fiber–reinforced polyester. For a large number of parts (automobile bodies, etc.), vacuum forming of PVC or polypropylene-type polymers is the desired method. A comparison of some polymer-processing methods is shown in [Table 165.2](#)

**Table 165.2** Comparison of Various Polymer-Processing Methods

Process	Production Rate	Material Type	Optimum Size
Extrusion	Fast	Th, plastic	Few mm-1.8 m
		Th, set	
Blow molding	Fast	Th, plastic	$10^{-6}$ -2 m <sup>3</sup>
Injection molding	Fast	Th, set	15 g-6 kg
Compression molding	Fast	Th, set	Few mm-0.4 m
Transfer molding	Fast	Th, set	Few mm-0.4 m
Layup techniques	Slow	Th, set and filler	0.01-400 m <sup>2</sup>
Rotational molding	Medium	Th, plastic	$10^{-3}$ -30 m <sup>3</sup>
Thermoforming	Fast	Th, plastic	$10^{-3}$ -20 m <sup>2</sup>

Adapted from Bolton, W. 1981. *Materials Technology*, p. 67. Butterworth, London. With permission.

## 165.3 Joining of Materials

Temporary joints between the shaped finished components made from several materials are easily achieved by mechanical means such as nuts, bolts, pins, and so forth. Permanent joining of two materials requires physical or chemical changes at the joining surfaces. For joining metals and alloys, welding, brazing, and soldering are the most common methods. Fusion welding consists of melting a part of the metal at the joint interface and allowing the joint to form by flow of the liquid metal, usually with another filler metal. When the same process is assisted by compressive forces applied by clamping the two pieces, the process is called *pressure welding*. Brazing and soldering use a lower-melting filler metal alloy to join the components. The parent pieces do not undergo melting and the liquid filler acts as the glue that makes the joint. The heat required to cause melting in the welding process classifies the process. Thus, in arc welding, heat is provided by creating an electric arc between electrodes often in the presence of a shielding atmosphere. Gas welding is performed using a gas mixture fuel to create a high-temperature flame that is targeted at the joining interface. The most popular gas-welding device is the oxyacetylene welding torch commonly seen in metal-working shops. Thermite welding is a well-known process for joining large sections in remote locations where gas or electric welding are difficult to perform. The thermite process relies on the exothermic heat generated by the reaction of iron oxide with aluminum. The molten product of the exothermic reaction is brought in contact with the joining interface and causes local melting of the parent pieces, followed by cooling, to form the weld. Electron and laser beam welding are some of the modern methods for joining metals and ceramics. Other methods include electrosag welding, friction welding, and explosive welding, which is made possible by a transient wave

generated by detonating a charge of explosives.

The joining of ceramics and glasses involves some methods common to metals and some that are specific to preparing glass-metal seals. Ceramic joining and bulk ceramic processing are closely related; the main joining methods consist of adhesive bonding, brazing with metals or silicates, diffusion bonding, fusion welding, and cementitious bonding. Brazing of metals to ceramics has attracted much recent attention because of its applications in heat engines, where high-temperature ceramic components (e.g., silicon nitride based) are in use. Thus, metallic alloys based on Ti, Zr, and Fe-Ni-Co are used to braze silicon nitride parts designed for high-temperature service. Thermal expansion mismatch between joining materials is an important criterion in the choice of those materials. Studies of glass-metal, ceramic-metal, ceramic-ceramic, and ceramic-polymer interfaces have led to a sophisticated understanding of joints, coatings, and enamels used in ceramic, electronic, and thermal engineering applications.

## Defining Terms

**Casting:** The shaping of a metal alloy into a solid piece by pouring liquid metal into the cavity inside a sand mold.

**Machining:** Final steps in the manufacture of a component; may include cutting, polishing, and grinding operations.

**Metal working:** Process steps to shape the cast material into components by cold or hot work. Examples of processes include rolling, forging, extrusion, and drawing.

**Powder processing:** The conversion of powdered material into a solid piece by cold pressing and heating (sintering) or simultaneous heating and pressure application (hot pressing).

**Sol-gel process:** The use of organic or organometallic solutions to make the component into a bulk piece or coating at lower temperatures than in conventional melting and casting. Used for oxide ceramics and glasses at present.

## References

- Alexander, J. M., Brewer, R. C., and Rowe, G. W. 1987. *Manufacturing Technology, Volume 2: Engineering Processes*. Halsted Press, New York.
- Benedict, G. F. 1987. *Nontraditional Manufacturing Processes*. Marcel Dekker, New York.
- Bolton, W. 1981. *Materials Technology*. Butterworth, London.
- Chu, B.-T. B., and Chen, S.-S. (Eds.). 1992. *Intelligent Modeling, Diagnosis, and Control of Manufacturing Processes*. World Scientific, Singapore.
- de Wit, J. H. W., Demaid, A., and Onillon, M. (Eds.) 1992. *Case Studies in Manufacturing With Advanced Materials, Volume 1*. North Holland, Elsevier Science, Amsterdam.
- Kalpajian, S. 1991. *Manufacturing Processes for Engineering Materials*. Addison-Wesley, Reading, MA.
- Loehman, R. E. 1989. Interfacial reactions in ceramic-metal systems. *American Ceramic Society Bulletin*. 68:891–896.
- Neely, J. E., and Kibbe, R. R. 1987. *Modern Materials and Manufacturing Processes*. John Wiley & Sons, New York.

Selverian, J. H., O'Neil, D., and Kang, S. 1992. Ceramic-to-metal joints. *American Ceramic Society Bulletin*. 71:1511–1520.

### **Further Information**

Sources of current information on materials and manufacturing methods include the *Journal of Materials Engineering and Performance*, *American Ceramic Society Bulletin*, *Powder Metallurgy*, *Materials Science and Engineering A* and *B*, and a series of conference proceedings published as *Ceramic Science and Engineering Proceedings* by the American Ceramic Society.

Shin, Y. C. "Machine Tools and Processes"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



## Machine Tools and Processes

---

### 166.1 Economic Impact

### 166.2 Types of Machine Tools

### 166.3 Control of Machine Tools

Interpolation • Feedback Control

### 166.4 Machine Tool Accuracy

#### Yung C. Shin

*Purdue University*

Machine tools are the machinery used to process various materials in order to get desired shapes and properties. Machine tools typically consist of a base structure, which supports various components and provides overall rigidity; kinematic mechanisms and actuators, which generate necessary motions of the tools and tables; and various feedback sensors. Machine tools are often used in conjunction with jigs and fixtures, which would hold the part in position, and processing tools. An example of a machine tool used in material removal processing is shown in [Fig. 166.1](#).

The first generation of modern machine tools was introduced during the industrial revolution in the 18th century with the invention of steam engines. Machine tools opened an era of automation by providing the means to replace human work with mechanical work. In the early days automation with machine tools was performed using various kinematic mechanisms, but it evolved into programmable automation with the use of computer numerical control. Today, many machine tools are operated with electrical or hydraulic power and controlled by a computer.

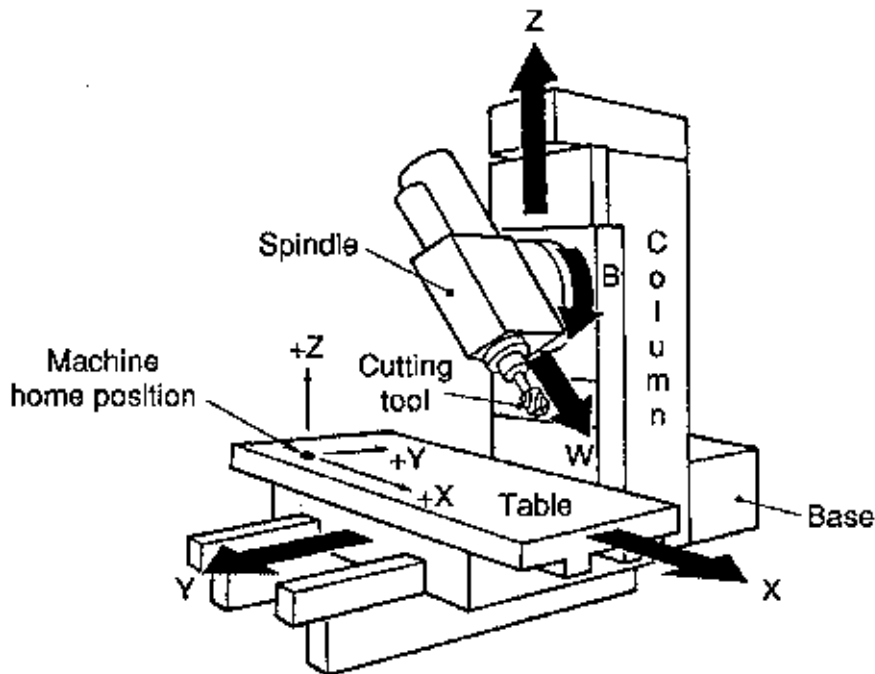
The major application of automation using machine tools in the early years was in transfer lines. Many special-purpose machines were grouped together with a part-moving system, such as a conveyor, providing the transfer of parts from one machine to another. A specific operation was performed on each station, and the part was moved to the next station until the finished product was obtained. Recently, however, **flexible automation** has been actively pursued to cope with continuous and rapid change of product designs and cycles. In such systems, several numerically controlled machines are clustered together to perform a variety of jobs. Programs are used, instead of inflexible mechanisms, to control the machines, so that the system can easily adapt to different job requirements.

### 166.1 Economic Impact

---

In most industrialized countries manufacturing contributes to about 20 to 30% of the nation's GNP. Since most manufacturing processes are performed with machine tools, the production and consumption of machine tools play a significant role in the nation's economy. Machine tool companies in the U.S. have a direct employment of about 350 000 people. If the surrounding ancillary industries are included—such as the tooling industry, material-handling industry, and sensors and controller manufacturers—the total employment would double that figure. More importantly, the machine tool industry affects a much larger user community.

**Figure 166.1** An example of machine tools.



## 166.2 Types of Machine Tools

Various manufacturing processes led to development of a variety of machine tools. Therefore, machine tools can be categorized by manufacturing processes. Manufacturing processes can be grouped into four main categories: casting, forming, material removing, and joining.

*Casting machines* are used to generate various unfinished or finished parts of defined shapes from shapeless materials. Casting is typically performed under high temperatures with molten material. The material in liquid state is poured into a mold or a die, which has the shape of the final part, and is subsequently solidified. Examples of casting processes entail sand casting, centrifugal casting, die casting, and continuous casting. No processing tools are used on this type of machine, and high accuracy is not usually required. Casting is usually followed by forming or machining processes in order to attain a final shape or to improve the dimensional accuracy.

*Forming machines* are used to change the shape of materials in the solid state. During the process, no significant change of volume or weight in the material occurs, but the mechanical properties are altered due to the large amount of bulk deformation and high stresses. Forming processes are performed at either high or low temperature. The required force and power are usually large. Examples of forming machine tools include presses; forging machines; rolling mills; and stamping, drawing, extrusion, and bending machines.

*Material removal processes* are used to obtain an accurate shape of a part that has been obtained by casting or forming processes. During the material removal process, the volume as well as the shape is changed. Machine tools used for material removal processes often require the highest accuracy. Typical processes under this category include turning, drilling, milling, boring, grinding, and so forth. Material removal processes have been used mostly for metals but have recently found new applications in nonmetallic materials such as ceramics and composites.

Machine tools for material removal processes are designed to provide multiaxis motions. They usually consist of a table used for mounting a workpiece, a spindle to hold tools, and a motor and gears to provide various speeds and feeds. In order to obtain a precision motion, axes are often controlled by microprocessor-based computer numerical controllers.

In addition to the processes already mentioned, many nontraditional machining processes are commonly used these days. Nontraditional machining processes include any that use another energy source with or without mechanical power. Typical nontraditional machining processes are electrodischarge machining (EDM), electrochemical grinding, laser machining, and electron-beam machining. These processes utilize either electric or heat energy to remove the material.

*Joining* is an operation to assemble components. Examples are welding, soldering, adhesion, and riveting. These processes are usually performed at the final stage to fabricate a product from various finished components. The types of machine tools in this category are friction-welding, arc-welding, electron-beam–welding, and laser-welding machines. In addition, robots are

## 166.3 Control of Machine Tools

---

In the early days machine tools were controlled automatically by means of mechanical systems such as cams, gears, and levers, which provided repetitive or sequential motions of machine tools. These mechanical devices shifted the operator's role from turning wheels or moving levers to supervising overall operation.

Later, hydraulic control replaced the mechanical control systems. Hydraulic systems provided more precise actions with higher power. The hydraulic control later combined with electronic switches were able to generate much more complex operations than mechanical control systems, but were eventually replaced by electronic controls.

*Numerically controlled* (NC) machines were developed to achieve automatic operation of machine tools with a program. The NC machine was first conceived by John Parson in 1948, and the first prototype was introduced by Massachusetts Institute of Technology (MIT) in 1952. In early NC machines, electric relays were grouped and hard-wired so that users could generate a complex sequence of operations. In NC machines, programs to generate commands for the sequences of operations were stored into punched tapes, which were subsequently read into a computer which interpreted the program and generated pulses to drive mechanisms.

In the beginning mainframes were used as computers to process the data. However, microprocessors and magnetic media replaced numerical control in the 1970s and opened the era of computer numerical controllers (CNC). Instead of being connected to the mainframe computer and sequentially controlled, CNC machines were designed with individual microprocessor-based controllers. CNC machines are usually equipped with a tool magazine, which holds multiple tools. These machines can perform many different operations on a single machine with automatic tool change and hence are called *machining centers*.

Computer programs used to operate the CNC machines are called *part programs*, which contain a sequence of commands. The commands used in a part program include preparatory G codes and M codes defining miscellaneous functions, as well as specific codes such as S, F, T, or X codes. The G codes are universally used regardless of the type of CNC and are defined in both ISO (International Standards Organization) and EIA (Electronic Industry Association) standards (RS-273-A). Some examples of these codes are shown in [Table 166.1](#). With NC or CNC machine tools, part programs can be generated at a remote site. The English-like languages, such as Automatic Programming Tool (APT), are commonly used for the ease of generating part programs. Nowadays, various computer-aided manufacturing (CAM) packages provide means to graphically generate part programs from design drawings created by computer-aided design (CAD) software. The part programs generated with CAM software must be postprocessed into the format of the specific CNC to be used and downloaded to the machine, where it can be edited if necessary.

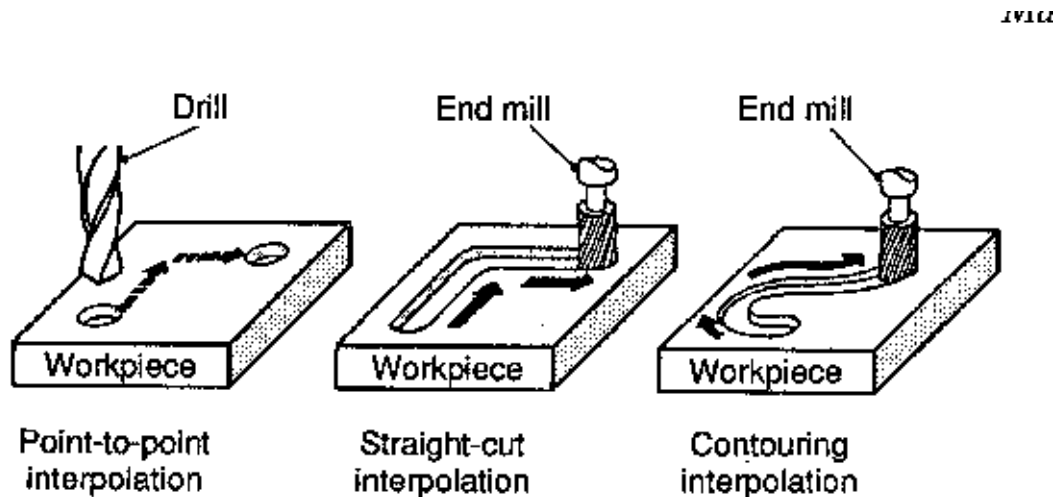
**Table 166.1** Examples of G and M Codes for Part Programs

G00	Point-to-point positioning	M03	Spindle rotation, CW
G01	Linear interpolation	M04	Spindle rotation, CCW
G02	Circular interpolation, CW	M05	Spindle stop
G03	Circular interpolation, CCW	M06	Tool change
G06	Spline interpolation	M08	Coolant No. 1 on
G17	xy-plane selection	M09	Coolant stop
G41	Tool radius compensation, left		

## Interpolation

Most CNC controllers have the capability of performing various types of interpolation. The simplest type is point-to-point interpolation. This scheme, in which beginning and final positions are of greatest importance, is usually used for rapid movement of tools or workpieces. The second type is the straight-line interpolation, which is used to generate a motion following a straight line with a specified speed. Accurate positioning and maintenance of a constant speed are required. The most complicated type of interpolation is contouring. Trajectory commands for contouring are typically generated using circular or cubic spline functions. [Figure 166.2](#) shows the graphical illustration of these interpolations.

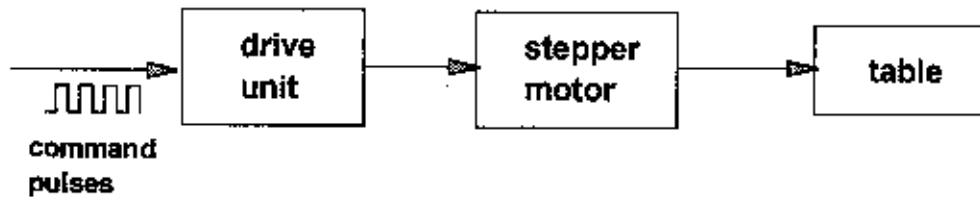
**Figure 166.2** Illustration of interpolations.



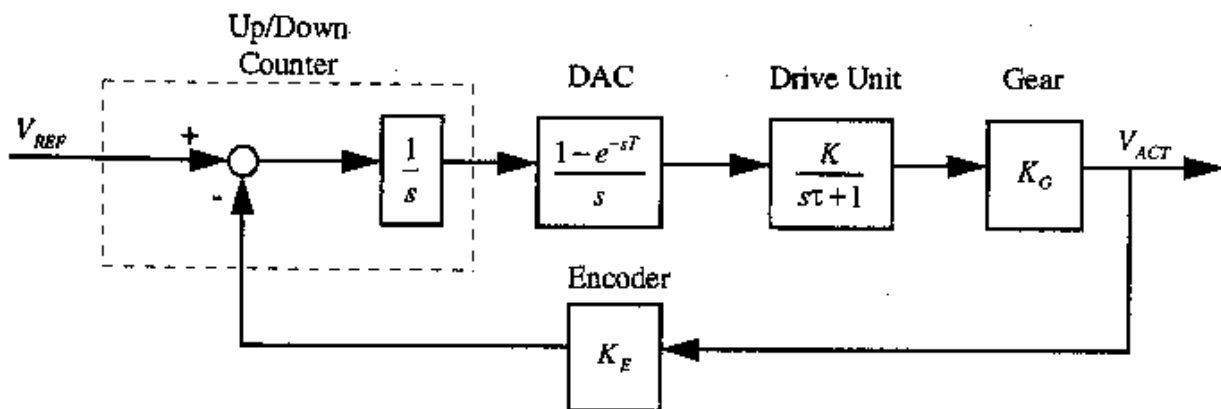
## Feedback Control

The generated commands are sent to the actuators of each axis to provide the necessary motion of the pertinent component in either open-loop or closed-loop configuration. In early CNC machines, stepper motors were often used as actuators, but their usage is at present limited to inexpensive, small machine tools. Either DC (direct current) or AC (alternating current) servomotors are popularly used for modern machine tools with feedback control. [Figure 166.3](#) shows the open-loop control system with a stepper motor, whereas [Fig. 166.4](#) depicts the closed-loop configuration.

**Figure 166.3** Open-loop control system with stepper motors.



**Figure 166.4** Closed-loop control system with stepper motors.

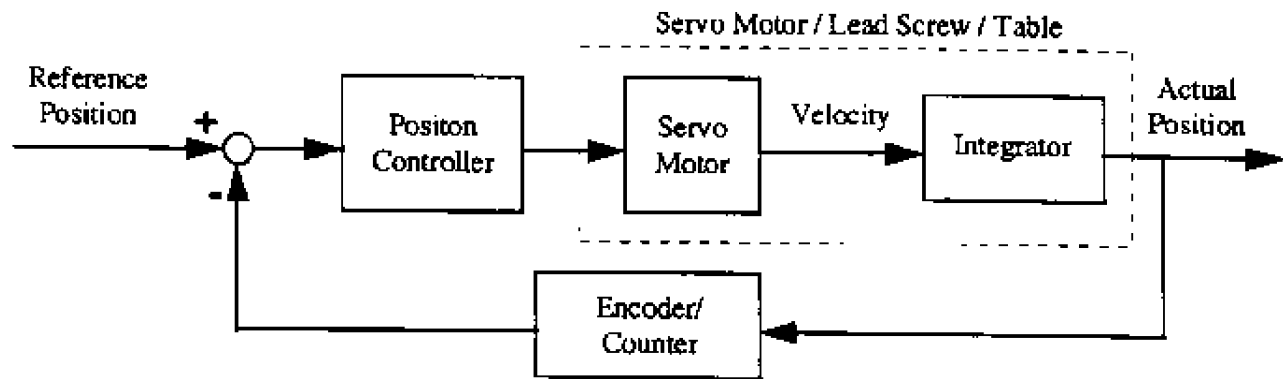


Early NC systems used a control scheme based on pulse reference and counting. In this scheme, positional reference commands are input to the system as a series of pulses, and the position feedback is obtained through an encoder. The command pulses are generated by the NC controller, with each pulse corresponding to a fixed **basic length unit** (BLU), which determines the system's accuracy. The frequency of the pulses determines the velocity of the axis, and positional accuracy is achieved by comparing the feedback pulses with the input pulse through the up-down counter.

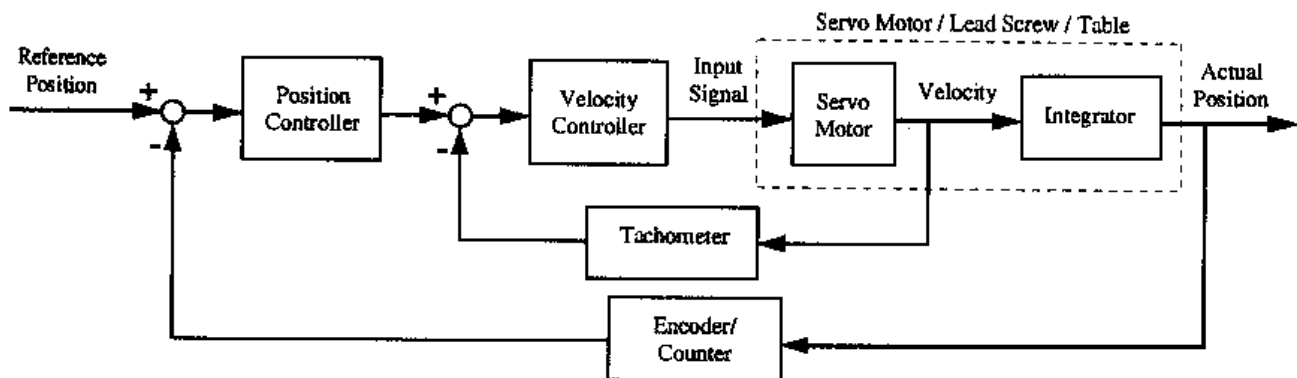
In more recent CNC systems, command generation and control are performed digitally using a digital controller with servomotors. Proportional-integral-derivative (PID) controllers are most popularly used as feedback controllers with either single- or double-feedback loops. An example of a position control system with a proportional controller is shown in Fig. 166.5. The more advanced system shown in Fig. 166.6 has two feedback loops, which perform velocity control in the inner loop and position tracking in the outer loop. The most common method uses a proportional type

for the position control and a PI controller for the velocity loop. These controllers are designed for the nominal operating condition of the machine.

**Figure 166.5** Digital position control system.



**Figure 166.6** Digital control system with position and velocity feedback.



Recently, adaptive control (AC) has opened a new era for machine tool control. Adaptive control provides the machine tool with the ability to adapt itself to the dynamic changes of the process condition to maintain optimum performance. Without adaptive control, operating conditions might have to be chosen too conservatively to avoid failure, thereby resulting in the loss of productivity. In addition, adaptive control can improve the accuracy of an existing machine, particularly in contouring. Adaptive control is particularly important for the realization of untended operation in the factory of the future.

## 166.4 Machine Tool Accuracy

The standard of accuracy has been increasing steadily over the last few decades. With the advent

of high-precision machine tool technology and measurement techniques, manufacturers can presently attain accuracy that was almost impossible to achieve a few decades ago. For example, machining accuracy of 5 to 10  $\mu\text{m}$  from regular machine tools and 0.1  $\mu\text{m}$  from precision machine tools is routinely achieved. It is projected that the accuracy will improve about tenfold every two decades.

The accuracy of a part produced by a machine tool is influenced by many factors, such as

- Machine tool accuracy
- Deflection or distortion of the part due to load or temperature
- Process operating condition
- Tool wear
- Environmental disturbances

Among these causes, the machine tool accuracy plays a very important role.

The machine tool error is defined as the deviation from the desired or planned path of the movement between the tool and the workpiece. Machine tool errors are typically grouped into geometric and kinematic errors. *Geometric error* includes positional inaccuracies of the individual link and the errors in the shape of the machine tool components. *Kinematic errors* are those occurring in coordinated movement of multiple machine elements. Therefore, kinematic errors depend on the machine's kinematic configuration and functional movements.

Machine tool errors stem from many sources, including imperfect fabrication of elements; assembly errors; friction; and errors associated with control, temperature variation, weight, and so forth. Depending on the source of error, various testing techniques must be adopted to evaluate the accuracy. Typical test methods include cutting tests, geometrical tests of moving elements, and master part tract tests. The details of test procedures are described in [Hocken, 1980; NMTBA, 1972; ANSI/ASME, 1985]. In addition, machine tool accuracy varies over time depending on the environment, operating history, and maintenance and hence can deteriorate significantly [Shin *et al.*, 1991]. Therefore, it is necessary to test the accuracy of the machine tools periodically and recalibrate the axes.

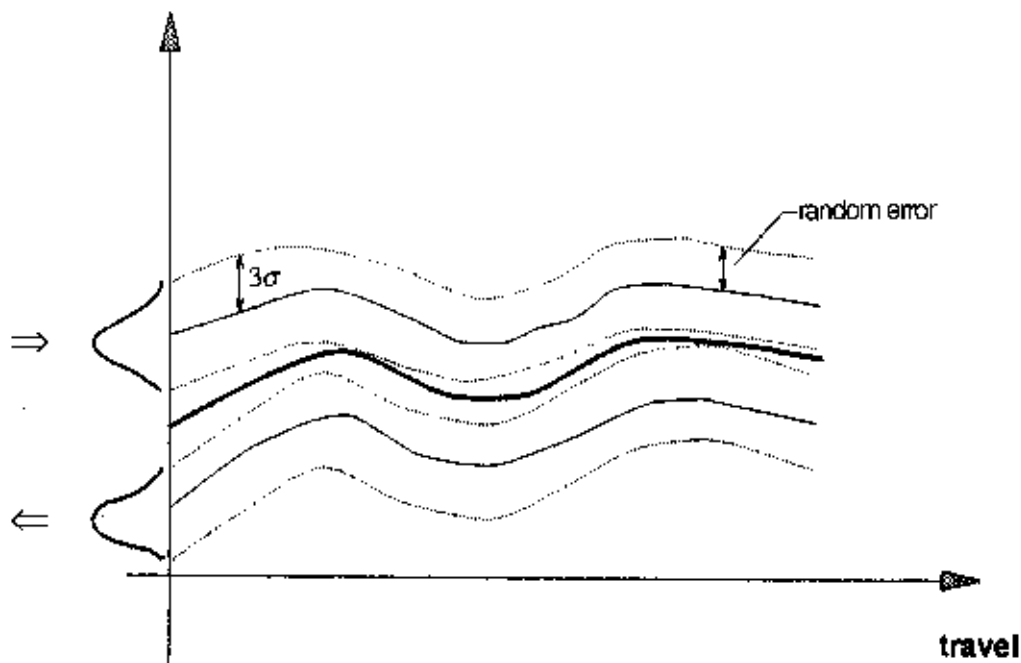
Machine tool error is divided into two categories: repeatable (deterministic) and nonrepeatable (random) errors. Typical measurements of positional errors are shown in Fig. 166.7. Due to the backlash and friction, errors often exhibit hysteresis if the measurements are performed in both directions. Generally, accuracy is defined as the worst possible error at a particular position and is represented in terms of a deterministic value defined as the mean deviation from the target position and the probabilistic distribution of random variation. Mathematically, positional accuracy is defined as

$$\varepsilon(x, y, z) = \max(|\mu + 3\sigma|, |\mu - 3\sigma|)$$

where  $\mu$  is the mean positional accuracy and  $\sigma$  is the variance of the random part. The nonrepeatable portion is used to specify the repeatability. Typically three sigma (variation) values are used to determine the repeatability.



**Figure 166.7** Presentation of errors of a machine tool axis moving in two directions.



## Defining Terms

**Basic length unit:** Smallest resolution attainable in positioning.

**Flexible automation:** Automation used to produce a variety of parts without converting the hardware set up.

## References

- ANSI/ASME. 1985. *Axes of Rotation, Methods for Specifying and Testing*. ANSI/ASME B89.3.4.M.
- Hocken, R. 1980. *Technology of Machine Tools*, vol. 5. Machine Tool Task Force, Lawrence Livermore Laboratory, Livermore, CA.
- NMTBA. 1972. *Definition and Evaluation of Accuracy and Repeatability for Numerically Controlled Machine Tools*, 2nd ed. National Machine Tool Builders Association, McLean, VA.
- Shin, Y. C., Chin, H., and Brink, M. J. 1991. Characterization of CNC machining centers. *J. of Manufacturing Sys.* 10(5):407–421.

## Further Information

A comprehensive treatment of machine tool technologies is presented in *Handbook of Machine Tools*, by Manfred Weck, (1984) and *Technology of Machine Tools*, published by Lawrence Livermore National Laboratory (1980).



Karwowski, W., Jamaldin, B. "Human Factors and Ergonomics"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Human Factors and Ergonomics

167.1 The Concept of Human-Machine Systems

167.2 Ergonomics in Industry

167.3 The Role of Ergonomics in Prevention of Occupational Musculoskeletal Injury

167.4 Fitting the Work Environment to the Workers

**Waldemar Karwowski**

*University of Louisville*

**Bill Jamaldin**

*University of Louisville*

The science of **ergonomics** originated in 1857 when Wojciech Jastrzebowski of Poland defined the term by combining two Greek words: *ergon* (work) + *nomos* (laws). This new science signified then the human work, play, thinking, and devotion as reflected in the manner of optimizing the use of four distinct human characteristics: (1) motor (physical), (2) sensory (aesthetic), (3) mental (intellectual), and (4) spiritual or moral [Karwowski, 1991]. The term *ergonomics* was independently reinvented by K. F. H. Murrell in 1949. Contemporary **human factors** (the parallel term for this new scientific discipline adopted in the U.S.), "discovers and applies information about human behavior, abilities, limitations, and other characteristics to the design of tools, machines, systems, tasks, jobs, and environments for productive, safe, comfortable, and effective human use" [Sanders and McCormick, 1993]. For example, human factors/ergonomics deals with a broad scope of problems relevant to the design and evaluation of work systems, consumer products, and working environments, whereas human-machine interactions affect human performance and product usability. The wide scope of issues addressed by ergonomics is presented in Table 167.1.

**Table 167.1** Classification Scheme for Human Factors/Ergonomics

1.	General
Human Characteristics	
2.	Psychological aspects
3.	Physiological and anatomical aspects
4.	Group factors
5.	Individual differences

6. Psychophysiological state variables
7. Task-related factors

---

#### Information Presentation and Communication

---

8. Visual communication
9. Auditory and other communication modalities
10. Choice of communication media
11. Person-machine dialogue mode
12. System feedback
13. Error prevention and recovery
14. Design of documents and procedures
15. User control features
16. Language design
17. Database organization and data retrieval
18. Programming, debugging, editing, and programming aids
19. Software performance and evaluation
20. Software design, maintenance, and reliability

---

#### Display and Control Design

---

21. Input devices and controls
22. Visual displays
23. Auditory displays
24. Other modality displays
25. Display and control characteristics

---

#### Workplace and Equipment Design

---

26. General workplace design and buildings
27. Workstation design
28. Equipment design

---

#### Environment

---

29. Illumination
30. Noise
31. Vibration
32. Whole body movement
33. Climate
34. Atmosphere
35. Altitude, depth, and space
36. Other environmental issues

---

#### System Characteristics

---

37. General system features
38. Total system design and evaluation
39. Hours of work
40. Job attitudes and job satisfaction
41. Job design
42. Payment systems
43. Selection and screening
44. Training
45. Supervision

46.	Use of support
47.	Technological and ergonomic change
Health and Safety	
48.	General health and safety
49.	Etiology
50.	Injuries and illnesses
51.	Prevention
Social and Economic Impact of the System	
52.	Trade unions
53.	Employment, job security, and job sharing
54.	Productivity
55.	Women and work
56.	Organizational design
57.	Education
58.	Law
59.	Privacy
60.	Family and home life
61.	Quality of working life
62.	Political comment and ethical
63.	Approaches and methods

Source: *Ergonomics Abstracts*, published by Taylor & Francis, Ltd., London, United Kingdom.

*Human factors design and engineering* aims to optimize the design and functioning of human-machine systems with respect to complex characteristics of people and the relationships between system users, machines, and outside environments. According to the Board of Certification in Professional Ergonomics (BCPE) a practitioner of ergonomics is a person who (1) has a mastery of a body of ergonomics knowledge, (2) has a command of the methodologies used by ergonomists in applying that knowledge to the design of a product, system, job, or environment, and (3) has applied his or her knowledge in the analysis, design testing, and evaluation of products, systems and environments. The areas of current practice in the field can be best described by examining the focus of technical groups of the Human Factors and Ergonomics Society, as illustrated in [Table 167.2](#).

**Table 167.2** Subject Interests of Technical Groups of the Human Factors and Ergonomics Society

	Technical Group	Description/Areas of Concerns
I.	Aerospace systems	Applications of human factors to the development, design, operation, and maintenance of human-machine systems in aviation and space environments (both civilian and military).
II.	Aging	Human factors applications appropriate to meeting the emerging needs of older people and special populations in a wide variety of life settings.
III.	Communications	All aspects of human-to-human communication, with an emphasis on communication mediated by telecommunications technology, including multimedia and collaborative communications, information services, and interactive broadband applications.

		Design and evaluation of both enabling technologies and infrastructure technologies in education, medicine, business productivity, and personal quality of life.
IV.	Computer systems	Human factors aspects of (1) interactive computer systems, especially user interface design issues; (2) the data-processing environment, including personnel selection, training, and procedures; and (3) software development.
V.	Consumer products	Development of consumer products that are useful, usable, safe, and desirable. Application of the principles and methods of human factors, consumer research, and industrial design to ensure market success.
VI.	Educators' professional	Education and training of human factors and ergonomics specialists in academia, industry, and government. Focus on both degree-oriented and continuing education needs of those seeking to increase their knowledge and or skills in this area, accreditation of graduate human factors programs, and professional certification.
VII.	Environmental design	Human factors aspects of the constructed physical environment, including architectural and interior design aspects of home, office, and industrial settings. Promotion of the use of human factors principles in environmental design.
VIII.	Forensics professional	Application of human factors knowledge and technique to "standards of care" and accountability established within legislative, regulatory, and judicial systems. The emphasis on providing a scientific basis to issues being interpreted by legal theory.
IX.	Industrial ergonomics	Application of ergonomics data and principles for improving safety, productivity, and quality of work in industry. Concentration on service and manufacturing processes, operations, and environments.
X.	Medical systems and functionally impaired populations	All aspects of the application of human factors principles and techniques toward the improvement of medical systems, medical devices, and the quality of life for functionally impaired user populations.
XI.	Organizational design	Improving productivity and the quality of life by an integration of psychosocial, cultural, and technological factors and with user interface factors (performance, acceptance, needs, limitations) in design of jobs, workstations, and related management systems.
XII.	Personality and individual differences in humanperformance	The range of personality and individual difference variables that are believed to mediate performance.
XIII.	Safety	Research and applications concerning human factors in safety and injury control in all settings and attendant populations, including transportation, industry, military, office, public building, recreation, and home improvements.
XIV.	System development	Concerned with research and exchange of information for integrating human factors into the development of systems. Integration of human factors activities into system development

		processes in order to provide systems that meet user requirements.
XV.	Test and evaluation	A forum for test and evaluation practitioners and developers from all areas of human factors and ergonomics. Concerned with methodologies and techniques that have been developed in their respective areas.
XVI.	Training	Fosters information and interchange among people interested in the fields of training and training research.
XVII.	Visual performance	The relationship between vision and human performance, including (1) the nature, content, and quantification of visual information and the context in which it is displayed; (2) the physics and psychophysics of information display; (3) perceptual and cognitive representation and interpretation of displayed information; (4) assessment of workload using visual tasks; and (5) actions and behaviors that are consequences of visually displayed information.

---

## 167.1 The Concept of Human-Machine Systems

---

A human-machine system can be broadly defined as "an organization of man and woman and the machines they operate and maintain in order to perform assigned jobs that implement the purpose for which the system was developed" [Meister, 1987]. The human functioning in such a system can be described in terms of perception, information processing, decision making, memory, attention, feedback, and human response processes. Furthermore, the human work taxonomy can be used to describe five distinct levels of human functioning, ranging from primarily physical tasks to cognitive tasks [Karwowski, 1992]. These basic but universal human activities are (1) tasks that produce force (primarily muscular work), (2) tasks of continuously coordinating sensory-monitor functions (e.g., assembling or tracking tasks), (3) tasks of converting information into motor actions (e.g., inspection tasks), (4) tasks of converting information into output information (e.g., required control tasks), and (5) tasks of producing information (primarily creative work).

Any task in a human-machine system requires processing of information that is gathered based on perceived and interpreted relationships between system elements. The processed information may need to be stored by either a human or a machine for later use. One of the important concepts for ergonomic design of human-machine systems is the paradigm of the stimulus-response compatibility [Wickens, 1987]. This paradigm relates to the physical relationship (compatibility) between a set of stimuli and a set of responses, as this relationship affects the speed of human response. The spatial relations between arrangements of signals and response devices in human-machine systems with respect to direction of movement and adjustments are often ambiguous, with a high degree of uncertainty regarding the effects of intended control actions. It should be noted that the information displayed to the human operator can be arranged along a continuum that defines the degree to which that information is spatial-analog (i.e., information about relative locations, transformations or continuous motion), linguistic-symbolic, or verbal (i.e., a set of instructions, alphanumeric codes, directions, or logical operations). The scope of ergonomic factors that need to be considered in design, testing, and evaluation of any

human-machine system is shown in [Table 167.3](#) in the form of an exemplary ergonomic checklist.

**Table 167.3** Examples of Factors to Be Used in the Ergonomics Checklists

**I. Anthropometric, biomechanical, and physiological factors**

1. Are the differences in human body size accounted for by the design?
2. Have the right anthropometric tables been used for specific populations?
3. Are the body joints close to neutral positions?
4. Is the manual work performed close to the body?
5. Are there any forward-bending or twisted trunk postures involved?
6. Are sudden movements and force exertion present?
7. Is there a variation in worker postures and movements?
8. Is the duration of any continuous muscular effort limited?
9. Are the breaks of sufficient length and spread over the duration of the task?
10. Is the energy consumption for each manual task limited?

**II. Factors related to posture (sitting and standing)**

1. Is sitting/standing alternated with standing/sitting and walking?
2. Is the work height dependent on the task?
3. Is the height of the work table adjustable?
4. Are the height of the seat and backrest of the chair adjustable?
5. Is the number of chair adjustment possibilities limited?
6. Have good seating instructions been provided?
7. Is a footrest used where the work height is fixed?
8. Has the work above shoulder or with hands behind the body been avoided?
9. Are excessive reaches avoided?
10. Is there enough room for the legs and feet?
11. Is there a sloping work surface for reading tasks?
12. Have the combined sit-stand workplaces been introduced?
13. Are handles of tools bent to allow for working with the straight wrists?

**III. Factors related to manual materials handling** (lifting, carrying, pushing and pulling loads)

1. Have tasks involving manual displacement of loads been limited?
2. Have optimum lifting conditions been achieved?
3. Is anybody required to lift more than 23 kg?
4. Have lifting tasks been assessed using the NIOSH (1991) method?
5. Are handgrips fitted to the loads to be lifted?
6. Is more than one person involved in lifting or carrying tasks?
7. Are there mechanical aids for lifting or carrying available and used?
8. Is the weight of the load carried limited according to the recognized guidelines?
9. Is the load held as close to the body as possible?
10. Are pulling and pushing forces limited?
11. Are trolleys fitted with appropriate handles and handgrips?

**IV. Factors related to the design of tasks and jobs**

1. Does the job consist of more than one task?
2. Has a decision been made about allocating tasks between people and machines?
3. Do workers performing the tasks contribute to problem solving?
5. Are the difficult and easy tasks performed interchangeably?
6. Can workers decide independently on how the tasks are carried out?
7. Are there sufficient possibilities for communication between workers?
8. Is there sufficient information provided to control the assigned tasks?

9. Can the group take part in management decisions?
10. Are the shift workers given enough opportunities to recover?

## **V. Factors related to information and control tasks**

### Information

1. Has an appropriate method of displaying information been selected?
2. Is the information presentation as simple as possible?
3. Has the potential confusion between characters been avoided?
4. Has the correct character/letter size been chosen?
5. Have texts with capital letters only been avoided?
6. Have familiar typefaces been chosen?
7. Is the text/background contrast good?
8. Are the diagrams easy to understand?
9. Have the pictograms been properly used?
10. Are sound signals reserved for warning purposes?

### Controls

1. Is the sense of touch used for feedback from controls?
2. Are differences between controls distinguishable by touch?
3. Is the location of controls consistent and is sufficient spacing provided?
4. Have the requirements for the control-display compatibility been considered?
5. Is the type of cursor control suitable for the intended task?
6. Is the direction of control movements consistent with human expectations?
7. Are the control objectives clear from the position of the controls?
8. Are controls within easy reach of female workers?
9. Are labels or symbols identifying controls properly used?
10. Is the use of color in controls design limited?

### Human-computer interaction

1. Is the human-computer dialogue suitable for the intended task?
2. Is the dialogue self-descriptive and easy to control by the user?
3. Does the dialogue conform to the expectations on the part of the user?
4. Is the dialogue error-tolerant and suitable for user learning?
5. Has command language been restricted to experienced users?
6. Have detailed menus been used for users with little knowledge and experience?
7. Is the type of help menu fitted to the level of user's ability?
8. Has the QWERTY layout been selected for the keyboard?
9. Has a logical layout been chosen for the numerical keypad?
10. Is the number of function keys limited?
11. Have the limitations of speech in human-computer dialogue been considered?
12. Are touch screens used to facilitate operation by inexperienced users?

## **VI. Environmental factors**

### Noise and vibration

1. Is the noise level at work below 80 dBA?
2. Is there an adequate separation between workers and source of noise?
3. Is the ceiling used for noise absorption?
4. Are the acoustic screens used?
5. Are hearing conservation measures fitted to the user?
6. Is personal monitoring to noise/vibration used?
7. Are the sources of uncomfortable and damaging body vibration recognized?
8. Is the vibration problem being solved at the source?
9. Are machines regularly maintained?
10. Is the transmission of vibration prevented?



#### Illumination

1. Is the light intensity for normal activities in the range of 200 to 800 lux?
2. Are large brightness differences in the visual field avoided?
3. Are the brightness differences between task area, close surroundings, and wider surroundings limited?
4. Is the information easily legible?
5. Is ambient lighting combined with localized lighting?
6. Are light sources properly screened?
7. Can the light reflections, shadows, or flicker from the fluorescent tubes be prevented?

#### Climate

1. Are workers able to control the climate themselves?
2. Is the air temperature suited to the physical demands of the task?
3. Is the air prevented from becoming either too dry or too humid?
4. Are draughts prevented?
5. Are the materials/surfaces that have to be touched neither too cold nor too hot?
6. Are the physical demands of the task adjusted to the external climate?
7. Are undesirable hot and cold radiation prevented?
8. Is the time spent in hot or cold environments limited?
9. Is special clothing used when spending long periods in hot or cold environments?

#### Chemical substances

1. Is the concentration of recognized hazardous chemical substances in the air subject to continuous monitoring and limitation?
2. Is the exposure to carcinogenic substances avoided or limited?
3. Does the labeling on packages of chemicals provide information on the nature of any hazard due to their contents?
4. Can the source of chemical hazards be removed, isolated, or their releases from the source reduced?
5. Are there adequate exhaust and ventilation systems in use?
6. Are protective equipment and clothing—including gas and dust masks for emergencies and gloves—available at any time if necessary?

---

Based on Dul, J. and Weerdmeester, B. 1993. *Ergonomics for Beginners: A Quick Reference Guide*. Taylor & Francis, London.

---

## 167.2 Ergonomics in Industry

---

The knowledge and expertise offered by ergonomics as applied to industrial environments can be used to (1) provide engineering guidelines regarding redesign of tools, machines, and work layouts, (2) evaluate the demands placed on the workers by the current jobs, (3) simulate alternative work methods and determine potential for reducing physical job demands if new methods are implemented, and (4) provide a basis for employee selection and placement procedures.

The basic foundations for ergonomics design are based on two components of industrial ergonomics: engineering anthropometry and biomechanics. **Occupational biomechanics** can be defined as the application of mechanics to the study of the human body in motion or at rest [Chaffin and Anderson, 1993]. Occupational biomechanics provides the criteria for application of anthropometric data to the problems of workplace design. **Engineering anthropometry** is an

empirical science branching from physical anthropology that (1) deals with physical measurements for the human body (such as body size, form (shape), and body composition), including, for example, the location and distribution of center of mass, weights, body links, or range of joint motions, and (2) applies these measures to develop specific engineering design requirements.

The recommendations for workplace design with respect to anthropometric criteria can be established by the principle of *design for the extreme*, also known as the *method of limits* [Pheasant, 1986]. The basic idea behind this concept is to establish specific boundary conditions (percentile value of the relevant human characteristic), which, if satisfied, will also accommodate the rest of the expected user population. The main anthropometric criteria for workplace design are clearance, reach, and posture. Typically, clearance problems involve the design of space needed for the legs or safe passageways around and between equipment. If the clearance problems are disregarded, they may lead to poor working postures and hazardous work layouts. Consideration of clearance requires designing for the largest user, typically by adapting the 95th percentile values of the relevant characteristics for male workers. Typical reach problems in industry include consideration of the location of controls and accessibility of control panels in the workplace. The procedure for solving the reach problems is usually based upon the fifth percentile value of the relevant characteristic for female workers (smaller members of the population). When anthropometric requirements of the workplace are not met, biomechanical stresses that manifest themselves in postural discomfort, lower-back pain, and overexertion injury are likely to occur.

## 167.3 The Role of Ergonomics in Prevention of Occupational Musculoskeletal Injury

---

Lack of attention to ergonomic design principles at work and its consequences have been linked to occupational musculoskeletal injuries and disorders. Musculoskeletal work-related injuries, such as cumulative trauma to the upper extremity and lower-back disorders (LBDs), affect several million workers each year, with total costs exceeding \$100 billion annually. For example, the upper-extremity **cumulative trauma disorders (CTDs)** account today for about 11% of all occupational injuries reported in the U.S. and have resulted in a prevalence of work-related disability in a wide range of occupations.

The dramatic increase of musculoskeletal disorders over the last 10 to 15 years can be linked to several factors, including the increased production rates leading to thousands of repetitive movements every day, widespread use of computer keyboards, higher percentage of women and older workers in the workforce, and better record keeping of employers as a result of a crackdown on industry reporting procedures by the Occupational Safety and Health Administration (OSHA). Other factors include greater employee awareness of these disorders and their relation to the working conditions, as well as a marked shift in the social policy regarding the recognition of compensation for work-related disorders. Given the epidemic proportions of the reported work-related musculoskeletal injuries, the federal government increases its efforts to introduce minimum standards and regulations aimed to reduce the frequency and severity of these disorders. For more information about these efforts, see Waters *et al.* [1993]; for NIOSH guidelines for manual lifting, see the ANSI [1994] draft document on control of CTDs.

The current state of knowledge about CTDs indicate that the chronic muscle, tendon, and nerve disorders may have multiple work-related and non-work-related causes. Therefore, CTDs are not classified as occupational diseases but rather as *work-related disorders*, where a number of factors may contribute significantly to the disorder, including work environment and human performance at work. The frequently cited risk factors of CTDs are (1) repetitive exertions; (2) posture—shoulder (elbow above mid-torso reaching down and behind), forearm (inward or outward rotation with a bent wrist), wrist (palmar flexion or full extensions), and hand (pinching); (3) mechanical stress concentrations over the base of palm, on the palmar surface of the fingers, and on the sides of the fingers; (4) vibration; (5) cold; and (6) use of gloves [Putz-Anderson, 1988]. A risk factor is defined here as an attribute or exposure that increases the probability of the disease or disorder. Cumulative trauma disorders at work are typically associated with repetitive manual tasks that impose repeated stresses to the upper body, that is, the muscles, tendons, ligaments, nerves, tissues, and neurovascular structures. For example, the three main types of disorders to the arm are (1) tendon disorders (e.g., tendonitis), (2) nerve disorders (e.g., carpal tunnel syndrome), and (3) neurovascular disorders (e.g., thoracic outlet syndrome or vibration-Raynaud's syndrome).

From the occupational safety and health perspective, the current state of ergonomics knowledge allows for management of CTDs in order to minimize human suffering, potential for disability, and the related worker's compensation costs. Ergonomics can help to (1) identify working conditions under which the CTDs might occur, (2) develop engineering design measures aimed at elimination or reduction of the known job risk factors, and (3) identify the affected worker population and target it for early medical and work intervention efforts. The ergonomic intervention should allow management to (1) perform a thorough job analysis to determine the nature of specific problems, (2) evaluate and select the most appropriate intervention(s), (3) develop and apply conservative treatment (implement the intervention), on a limited scale if possible, (4) monitor progress, and (5) adjust or refine the intervention as needed.

Most of the current guidelines for control of the CTDs at work aim to (1) reduce the extent of movements at the joints, (2) reduce excessive force levels, and (3) reduce exposure to highly repetitive and stereotyped movements. Workplace design to prevent the onset of CTDs should be directed toward fulfilling the following recommendations: (1) permit several different working postures; (2) place controls, tools, and materials between waist and shoulder heights for ease of reach and operation; (3) use jigs and fixtures for holding purposes; (4) resequence jobs to reduce the repetition; (5) automate highly repetitive operations; (6) allow self-pacing of work whenever feasible; and (7) allow frequent (voluntary and mandatory) rest breaks. For example, some of the common methods to control the wrist posture are (1) altering the geometry of tool or controls (e.g., bending the tool or handle), (2) changing the location/positioning of the part, or (3) changing the position of the worker in relation to the work object. In order to control the extent of force required to perform a task, one can (1) reduce the force required through tool and fixture redesign, (2) distribute the application of the force, or (3) increase the mechanical advantage of the (muscle) lever system.

## 167.4 Fitting the Work Environment to the Workers

---

Ergonomic job redesign focuses on fitting the jobs and tasks to capabilities of workers—for example, "designing out" unnatural postures at work, reducing excessive strength requirements, improving work layout, introducing appropriate designs of hand tools, and addressing the problem of work/rest requirements. As widely recognized in Europe, ergonomics must be seen as a vital component of the value-adding activities of the company. Even in strictly financial terms, the costs of an ergonomics management program will be far outweighed by the costs of not having one. A company must be prepared to accept a participative culture and to utilize participative techniques. The ergonomics-related problems and consequent intervention should go beyond engineering solutions and must include design for manufacturability, total quality management, and work organization alongside workplace redesign or worker training.

An important component of the management efforts to control musculoskeletal disorders in industry is the development of a well-structured and comprehensive ergonomics program. The basic components of such a program should include the following: (1) health and risk factor surveillance, (2) job analysis and improvement, (3) medical management, (4) training, and (5) program evaluation. Such a program must include participation of all levels of management; medical, safety, and health personnel; labor unions; engineering; facility planners; and workers.

The expected benefits of ergonomically designed jobs, equipment, products, and workplaces include the improved quality and productivity with reductions in errors, enhanced safety and health performance, heightened employee morale, and accommodation of people with disabilities to meet the requirements of the Americans with Disabilities Act and affirmative action. However, the recommendations offered by ergonomics can be successfully implemented in practice only with full understanding of the production processes, plant layouts, quality requirements, and total commitment from all management levels and workers in the company. Furthermore, these efforts can only be effective through participatory cooperation between management and labor through development of in-plant ergonomics committees and programs. Ergonomics must be treated at the same level of attention and significance as other business functions of the plant—for example, the quality management control—and should be accepted as a cost of doing business, rather than an add-on activity calling for action only when the problems arise.

### Defining Terms

**Cumulative trauma disorders (CTDs):** CTDs at work are typically associated with repetitive manual tasks with forceful exertions, performed with fixed body postures that deviate from neutral, such as those at assembly lines and those using hand tools, computer keyboards, mice, and other devices. These tasks impose repeated stresses to the soft tissues of the arm, shoulder, and back, including the muscles, tendons, ligaments, nerve tissues, and neurovascular structures, which may lead to tendon and/or joint inflammation, discomfort, pain, and potential work disability.

**Engineering anthropometry:** An empirical science branching from physical anthropology that deals with physical measurements for the human body—such as body size, form (shape), and body composition—and application of such measures to the design problems.

**Ergonomics/human factors:** The scientific discipline concerned with the application of the relevant information about human behavior, abilities, limitations, and other characteristics to

the design, testing, and evaluation of tools, machines, systems, tasks, jobs, and environments for productive, safe, comfortable, and effective human use.

**Occupational biomechanics:** The application of mechanics to the study of the human body in motion or at rest, focusing on the physical interaction of workers with their tools, machines, and materials in order to optimize human performance and minimize the risk of musculoskeletal disorders.

## References

- Chaffin, D. B. and Anderson, G. B. J. 1993. *Occupational Biomechanics*, 2nd ed. John Wiley & Sons, New York.
- Dul, J. and Weerdmeester, B. 1993. *Ergonomics for Beginners: A Quick Reference Guide*. Taylor & Francis, London.
- Karwowski, W. 1991. Complexity, fuzziness and ergonomic incompatibility issues in the control of dynamic work environments. *Ergonomics*. 34 (6): 671–686.
- Karwowski, W. 1992. Occupational biomechanics. In *Handbook of Industrial Engineering*, ed. G. Salvendy, p. 1005–1046. John Wiley & Sons, New York.
- Meister, D. 1987. Systems design, development and testing. In *Handbook of Human Factors*, ed. G. Salvendy, p. 17–42. John Wiley & Sons, New York.
- Pheasant, S. 1986. *Bodyspace: Anthropometry, Ergonomics and Design*. Taylor & Francis, London.
- Putz-Anderson, V. (Ed.) 1988. *Cumulative Trauma Disorders: A Manual for Musculoskeletal Diseases of the Upper Limbs*. Taylor & Francis, London.
- Sanders, M. S. and McCormick, E. J. 1993. *Human Factors in Engineering and Design*, 6th ed. McGraw-Hill, New York.
- Waters, T. R., Putz-Anderson, V., Garg, A., and Fine, L. J. 1993. Revised NIOSH equation for the design and evaluation of manual lifting tasks. *Ergonomics*. 36 (7):749–776.
- Wickens, C. D. 1987. Information processing, decision-making, and cognition. In *Handbook of Human Factors*, ed. G. Salvendy, p. 72–107. John Wiley & Sons, New York.

## Further Information

- ANSI. 1994. *Control of Cumulative Trauma Disorders*. ANSI Z-365 Draft. National Safety Council, Itasca, IL.
- Clark, T. S. and Corlett, E. N. (Eds.) 1984. *The Ergonomics of Workspaces and Machines: A Design Manual*. Taylor & Francis, London.
- Cushman, W. H. and Rosenberg, D. J. 1991. *Human Factors in Product Design*. Elsevier, Amsterdam.
- Eastman Kodak Company. 1983. *Ergonomic Design for People at Work*, Volume 1. Lifetime Learning, Belmont, CA.
- Eastman Kodak Company. 1986. *Ergonomic Design for People at Work*, Volume 2. Van Nostrand Reinhold, New York.
- Grandjean, E. 1988. *Fitting the Task to the Man*, 4th ed. Taylor & Francis, London.

Helander, M. (Ed.) 1988. *Handbook of Human-Computer Interaction*. North-Holland, Amsterdam.

Kroemer, K. H. E., Kroemer, H. B., and Kroemer-Elbert, K. E. 1994. *Ergonomics: How to Design for Ease and Efficiency*. Prentice Hall, Englewood Cliffs, NJ.

Salvendy, G. (Ed.) 1987. *Handbook of Human Factors*. John Wiley & Sons, New York.

Salvendy, G. and Karwowski, W. (Eds.) 1994. *Design of Work and Development of Personnel in Advanced Manufacturing*. John Wiley & Sons, New York.

Wilson, J. R. and Corlett, E. N. (Eds.) 1990. *Evaluation of Human Work: A Practical Methodology*. Taylor & Francis, London.

Woodson, W. E. 1981. *Human Factors Design Handbook*. McGraw-Hill, New York.

## **Ergonomics Information Sources and Professional Societies**

International Ergonomics Association (IEA)  
Poste Restante  
Human Factors and Ergonomics Society  
P.O. Box 1369  
Santa Monica, CA  
Phone: (310) 394-1811/9793. Fax: 310.394.2410  
Crew System Ergonomics Information Analysis Center  
AL/CFH/CSERIAC  
Wright Patterson AFB  
Dayton, OH 45433-6573  
Phone: (513) 255-4842. Fax: 513.255.4823  
Ergonomics Information Analysis Centre (EIAC)  
School of Manufacturing and Mechanical Engineering  
The University of Birmingham  
Birmingham B15 2TT  
England  
Phone: +44-21-414-4239. Fax: +44-21-414-3476

## **Journals**

*Ergonomics Abstracts*. Published by Taylor & Francis, Ltd., London.

*Human Factors*. Published by the Human Factors and Ergonomics Society, Santa Monica, California.

*International Journal of Human Factors in Manufacturing*. Published by John Wiley & Sons, New York.

*Applied Ergonomics*. Published by Butterworth-Heinemann, Oxford, UK.

*International Journal of Human-Computer Interaction*. Published by Ablex, Norwood, NJ.

*International Journal of Industrial Ergonomics*. Published by Elsevier, Amsterdam.

*International Journal of Occupational Safety and Ergonomics*. Published by Ablex, Norwood, NJ.

Peter Bilotto, P., Borzileri, C., et al. "Pressure and Vacuum"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## 168.1 Pressure

Temperature • Volume • Basic Pressure Equations • Basic Design Equations (Internally Pressurized Vessels)

## 168.2 The Vacuum Environment

Methods for Measuring Subatmospheric Pressures • Methods for Reducing Pressure in a Vacuum Vessel • Vacuum System Design

### **Peter Bilotft**

*University of California  
Lawrence Livermore National Laboratory*

### **Charles Borzileri**

*University of California  
Lawrence Livermore National Laboratory*

### **Dave Holten**

*University of California  
Lawrence Livermore National Laboratory*

### **Matt Traini**

*University of California  
Lawrence Livermore National Laboratory*

## 168.1 Pressure

---

Pressure ( $P$ ) can be defined as a force ( $F$ ) acting on a given area ( $A$ );  $P = F/A$ . It is measured using metric (SI) units, pascals (Pa), or English units, pounds per square inch (psi). Pressure can also be indicated in absolute (abs) or gage (g) (as measured with a gage) units. Relationships between metric and English units of measure are as follows:

- 100 kPa = 1 bar = 14.7 psia  $\approx$  1 atmosphere (atm).
- 100 kPa gage (or 100 kPa) = 14.7 psig (or 14.7 psi).
- 1 MPa abs = 147 psia.



- 1 MPa gage (or 1 MPa) = 147 psig (or 147 psi).

These units can be used when measuring seven kinds of pressure:

- *Absolute pressure*: sum of atmospheric and gage pressures.
- *Atmospheric pressure*: ambient pressure at local elevations. The atmospheric pressure at sea level is approximately 100 kPa or 14.7 psia.
- *Burst pressure*: theoretical or actual pressure at which a vessel/component will fail.
- *Gage pressure*: pressure above atmospheric pressure as measured by a gage.
- *Maximum allowable working pressure (MAWP)*: maximum pressure at which component is safe to operate. MAWP is the maximum permissible setting for relief devices. Service, rated, design, and working pressure (WP) are the same as MAWP.
- *Maximum operating pressure (MOP)*: maximum pressure for continuous operation. MOP should be 10 to 20% below MAWP.
- *Pressure test*: a test to ensure that a vessel or system will not fail or permanently deform and will operate reliably at a specified pressure.

## Temperature

Temperature ( $T$ ) is the intensity of heat in a substance and may be measured on the graduated scale of a thermometer. It is usually measured in degrees Fahrenheit ( $^{\circ}\text{F}$ ) or degrees Celsius ( $^{\circ}\text{C}$ ). The two absolute scales are: Rankine (R) and Kelvin (K) (see [Table 168.1](#)).

**Table 168.1** Comparison of Temperature Scales<sup>1,2</sup>

Criterion	Fahrenheit ( $^{\circ}\text{F}$ )	Rankine abs (R)	Celsius or Centigrade ( $^{\circ}\text{C}$ )	Kelvin abs (K)
Water boils	212	672	100.0	373.0
Water freezes	32	492	0.0	273.0
Zero $^{\circ}\text{F}$	0	460	-17.7	255.3
Absolute zero	-460	0	-273.0	0.0

<sup>1</sup>Conversion factors:  $^{\circ}\text{F} = \frac{9}{5}^{\circ}\text{C} + 32$ ;  $^{\circ}\text{C} = \frac{5}{9} (^{\circ}\text{F} - 32)$ ;  $\text{R} = ^{\circ}\text{F} + 460$ ;  $\text{K} = ^{\circ}\text{C} + 273$ .

<sup>2</sup>Use absolute temperatures, either Rankine or Kelvin, in all calculations involving the ideal gas law, except when temperature is constant.

*Source*: This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract N. W-7405-Eng-48.

## Volume

Volume ( $V$ ) refers to a space occupied in three dimensions. It is generally measured in cubic meters ( $\text{m}^3$ ), cubic centimeters ( $\text{cm}^3$ ), cubic inches ( $\text{in}^3$ ), or cubic feet ( $\text{ft}^3$ ). Volume may also be measured in units of liters (L) or gallons (gal).

## Basic Pressure Equations

### General Gas Law

The general (or ideal) gas law is the basic equation that relates pressure, volume, and temperature of a gas. It is expressed mathematically as

$$\frac{PV}{T} = \text{constant} \quad (168.1)$$

where

$P$  = pressure in kilopascals absolute (kPa abs) or pounds per square inch absolute (psia).  
(Add atmospheric pressure of 100 kPa or 15 psia to gage pressure.)

$V$  = volume in any convenient cubic unit such as liters, cubic feet,  
etc.

$T$  = temperature in absolute units of Rankine or Kelvin.

When the conditions of a gas system change, they do so according to this general gas law, at least where moderate pressures are concerned. At higher pressures ( $P > 200$  atm) gases do not obey the general gas law exactly; this problem is discussed later.

The following checklist will assist in solving general gas law problems:

1. Sketch the problem and mark all conditions.
2. List initial and final values of  $P$ ,  $V$ , and  $T$ . Use absolute units for  $P$  and  $T$ .
3. Write the general gas law for initial and final conditions. Cross out conditions that remain constant.
4. Substitute known values and solve for unknowns.
5. Convert your answer if it is in units different from those given in the problem.

The general gas law expressed for initial (subscript 1) and final (subscript 2) conditions is

$$\frac{P_1 V_1}{T_1} = \frac{P_2 V_2}{T_2} \quad (168.2)$$

Any one of these quantities can be solved for by simply cross-multiplying. Following is an example to show how easily the general gas law can be applied and how much can be determined about the behavior of gases.

**Example.** A cylinder of dry nitrogen is received from a vendor. The temperature of the nitrogen is 70°F, and the pressure within the cylinder is 2250 psig. An employee inadvertently stores the cylinder too close to a radiator so that within 8 hours the nitrogen is heated to 180°F. What is the new pressure within the cylinder?

Initial	Final
$P_1 = 2250 \text{ psig} = 2265 \text{ psia}$	$P_2 = (\text{find})$
$V_1 = V_2 = \text{constant}$	$V_2 = V_1 = \text{constant}$
$T_1 = 70^\circ\text{F} + 460^\circ = 530 \text{ R}$	$T_2 = 180^\circ\text{F} + 460^\circ = 640 \text{ R}$

$$\frac{P_1 \cancel{V_1}}{T_1} = \frac{P_2 \cancel{V_2}}{T_2} \quad \text{cross-multiplying,} \quad P_2 = \frac{P_1 T_2}{T_1} \quad (168.3)$$

$$P_2 = \frac{2265 \times 640}{530} = 2735 \text{ psia} = 2720 \text{ psig}$$

### Pascal's Law

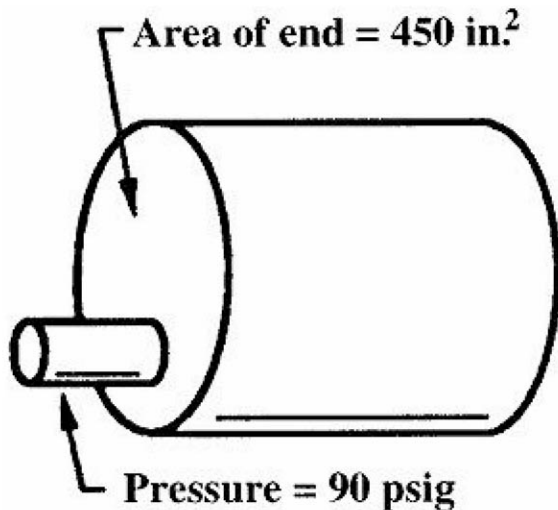
According to Pascal's law, pressure exerted on a confined fluid will act in all directions with equal force.

$$\text{Pressure} = \frac{\text{Force}}{\text{Area}} \quad \text{or} \quad \text{Force} = \text{Pressure} \times \text{Area} \quad (168.4)$$

**Example.** Find the force on the end of a cylinder (see [Fig. 168.1](#)).

$$F = AP \quad \text{or} \quad F = 450 \text{ in.}^2 \times 90 \text{ lb/in.}^2 = 40\,500 \text{ lb of force}$$

**Figure 168.1** Example problem: force on the end of a cylinder.



### Safety Factor

The ratio of the calculated failure pressure (or actual failure pressure if known) to the MAWP is

the safety factor.

$$SF = \frac{\text{Burst pressure}}{\text{MAWP}} \quad (168.5)$$

A safety factor related to a value other than the failure pressure should be so identified with an appropriate subscript, for example,  $SF_y$  for a safety factor based on the yield strength or  $SF_u$  for a safety factor based on the ultimate strength of the material.

### Stored Energy in a Pressurized System

Gas confined under high pressure is like a hungry tiger in a cage. If it should suddenly escape in a populated area, someone would probably be injured. When a gas-pressure vessel fails, it propels jagged vessel fragments in all directions. The energy of the gas, assuming isentropic expansion, is expressed as

$$U_{\text{gas}} = \frac{P_1 V}{k - 1} \left[ 1 - \left( \frac{P_2}{P_1} \right)^{(k-1)/k} \right] \quad (168.6)$$

where

$U_{\text{gas}}$  = stored energy

$P_1$  = container pressure

$P_2$  = atmospheric pressure

$V$  = volume of the pressure vessel

$k = c_p/c_v = 1.41$  for  $N_2$ ,  $H_2$ ,  $O_2$ , and air, and 1.66 for He [see [Baumeister, 1958, chap. 4](#)]

$c_p$  = specific heat at constant pressure

$c_v$  = specific heat at constant volume

This expression as written will yield an energy value in joules if pressure and volume units are in megapascals (MPa) and cubic centimeters (cc), respectively.

A liquid confined under high pressure is also a potential safety hazard. However, for systems of comparable volume and pressure, the amount of stored energy in the liquid case will be considerably less than that contained in the gas because liquids are essentially noncompressible. The energy involved in the sudden failure of a liquid-filled vessel may be conservatively determined from

$$U_{\text{liq}} = \frac{1}{2} \left( \frac{P_1^2 V}{B} \right) \quad (168.7)$$

where  $B$  is the liquid bulk modulus. Some typical bulk moduli ( $B$ ) are 300 000 psi for water, 225 000 psi for oil, and 630 000 psi for glycerin.

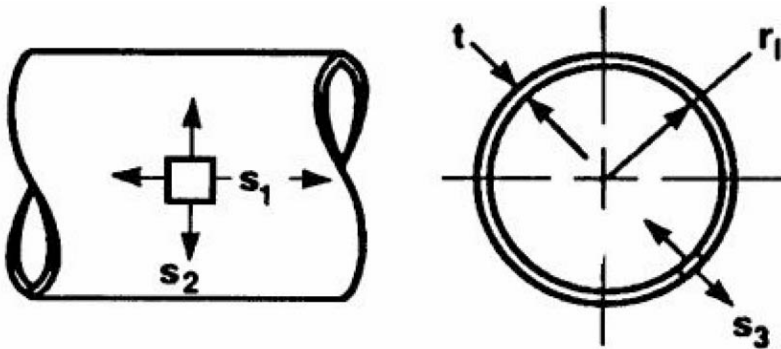
An additional and significant potential source of released energy for confined gases is chemical reactions, for example, hydrogen and oxygen. For these special situations a separate analysis

addressing any potential chemical energy release should also be made [see [Baumeister, 1958, chap. 4](#)].

## Basic Design Equations (Internally Pressurized Vessels)

The following equations, when properly applied, give stresses that can be expected to exist in a design—longitudinal or axial stress ( $s_1$ ), circumferential or hoop stress ( $s_2$ ), and radial stress ( $s_3$ ) (see [Fig. 168.2](#)). These stresses must be compared to those permitted for the vessel material after application of the appropriate safety factor.

**Figure 168.2** Stress in thin cylinders.



### Thin Cylinders

A cylinder (or any vessel) is said to be "thin" when its thickness ( $t$ ) and internal radius ( $r_i$ ) satisfy the following relationship:

$$\frac{t}{r_i} \leq 0.1 \quad (168.8)$$

The following stresses exist in a thin cylinder:

$$s_1 = \frac{P}{2} \times \frac{r_i}{t}, \quad s_2 = P \times \frac{r_i}{t}, \quad s_3 = -P \quad (168.9)$$

Given  $P = 800$  psig,  $r_i = 5$  in.,  $t = 0.25$ , then  $s_1 = 8000$  psi,  $s_2 = 16\,000$  psi, and  $s_3 = -800$  psi.

### Thin Spheres

The following stresses exist in a thin sphere (see [Fig. 168.3](#)):

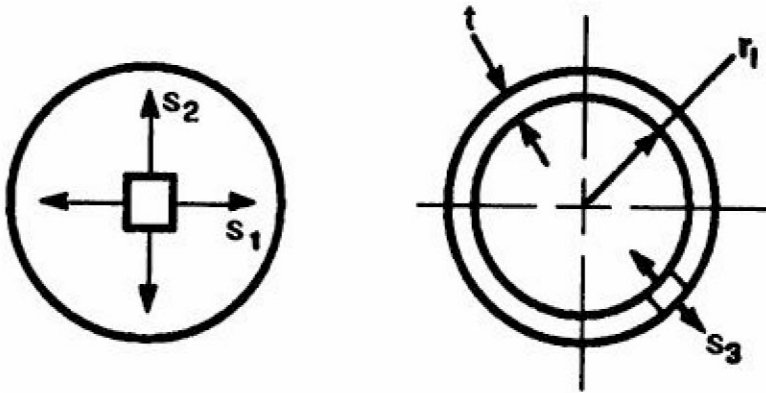
$$s_1 = s_2, \quad s_2 = \frac{P}{2} \times \frac{r_i}{t}, \quad s_3 = -P \quad (168.10)$$

Given:

$$\frac{t}{r_i} = 0.05 (\leq 0.1), \quad \frac{r_i}{t} = 20, \quad s_1 = s_2 \quad (168.11)$$

Then  $s_1 = 800/2 \times 20 = 8000$  psi and  $s_3 = -800$  psi. *Note that the hoop stress of a thin sphere is half the hoop stress of a corresponding thin cylinder.*

**Figure 168.3** Stress in thin spheres.



### Thick Cylinders (Roark)

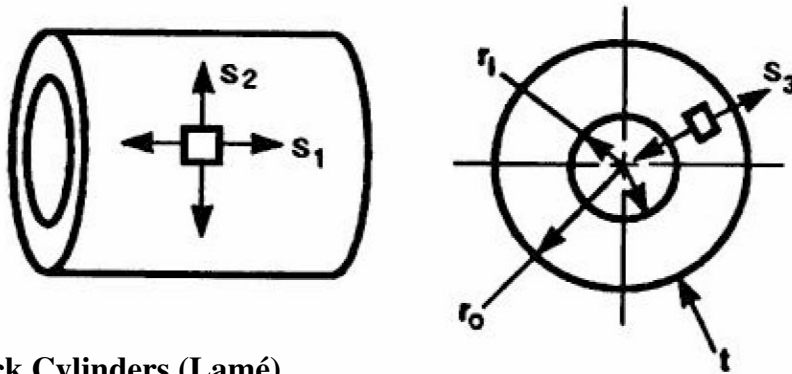
A cylinder or sphere is "thick" when its dimensions satisfy the following relationship [Roark, 1989]:

$$\frac{t}{r_i} > 0.1 \quad (168.12)$$

The following stresses exist in a thick cylinder (see Fig. 168.4):

$$\begin{aligned} S_1 &= P \frac{r_i^2}{(r_o^2 - r_i^2)}, & S_2 &= P \frac{r_o^2 + r_i^2}{(r_o^2 - r_i^2)} \text{ (max. inner surface),} \\ S_3 &= -P \text{ (max. inner surface)} \end{aligned} \quad (168.13)$$

**Figure 168.4** Stress in thick cylinders.



### Thick Cylinders (Lamé)

Compare Roark's equation with that of Lamé [Baumeister, 1958, chap. 5]. Use the same vessel and conditions as above, with hoop stress maximum at the inner surface.

$$S_1 = P \frac{R^2 + 1}{R^2 - 1}, \quad R = \frac{r_o}{r_i} \quad (168.14)$$

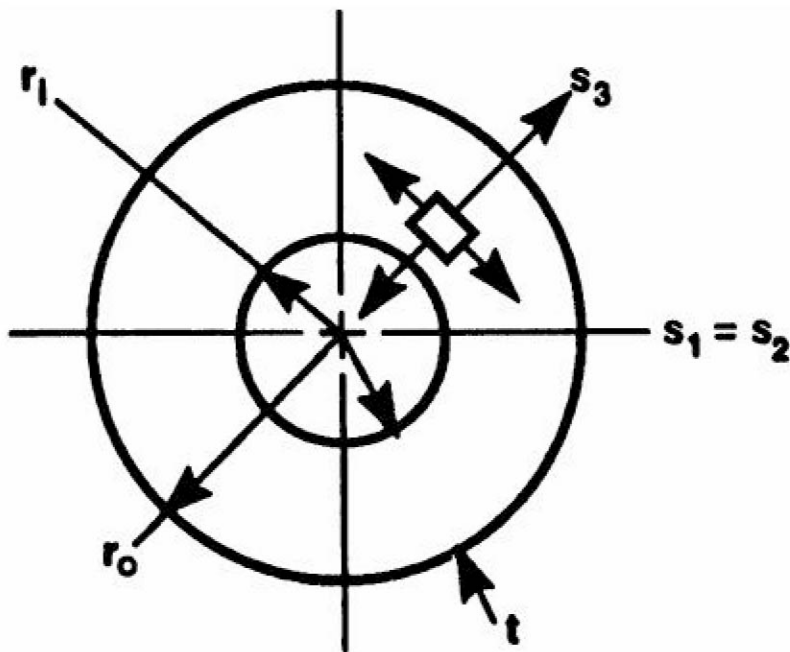
Note that the Lamé equation agrees with Roark's. Actually, it is the same equation.

### Thick Spheres

Figure 168.5 shows the stress relationship in thick spheres, where  $t/r_i > 0.1$  .

$$\begin{aligned} s_1 \text{ (max. inner surface)} &= s_2, \\ s_2 &= \frac{P(r_o^3 + 2r_i^3)}{2E(r_o^3 - r_i^3)} \quad \text{where } E \text{ is the joint efficiency} \quad (168.15) \\ s_3 \text{ (max. inner surface)} &= -P \end{aligned}$$

Figure 168.5 Stresses in thick spheres.



## 168.2 The Vacuum Environment

As a practical definition, any environment in which the pressure is lower than atmospheric may be referred to as a vacuum. It is useful to express this subatmospheric pressure condition in units which reference zero pressure. Atmospheric pressure is approximately equal to 760 torr, 101 300 Pa, or 14.7 psi. Table 168.2 provides conversion factors for some basic units.

Table 168.2 Conversion Factors for Selected Pressure Units Used in Vacuum Technology

	Pa	torr	atm	mbar	psi
Pa	1	0.0075	$9.78 \cdot 10^{-6}$	0.01	$1.45 \cdot 10^{-4}$
torr	133	1	$1.32 \cdot 10^{-3}$	1.333	0.0193
atm	$1.01 \cdot 10^5$	760	1	1013	14.7
mbar	100	0.75	$9.87 \cdot 10^{-4}$	1	0.0145
psi	6890	51.71	$6.80 \cdot 10^{-2}$	68.9	1

The *ideal gas law* is generally valid when pressure is below atmospheric.

$$PV = nRT \quad (168.16)$$

where

$P$  = pressure [atm]

$V$  = volume [liters]

$n$  = amount of material [**moles**]

$R$  = gas law constant [0.082 liter-atm/K-mole]

$T$  = temperature [K]

The rate at which molecules strike a surface of unit area per unit time is referred to as the *impingement rate* and is given by

$$I = 3.5 \times 10^{22} \left( \frac{P}{MW \times T} \right) \quad (168.17)$$

where

$P$  = pressure [torr]

$MW$  = molecular weight [g/mole]

$T$  = temperature [K]

The distance, on average, that a molecule can travel without colliding with another molecule is the *mean free path* and is given by

$$\lambda = \frac{1}{\sqrt{2}\pi d_o^2 n} \quad (168.18)$$

where

$\lambda$  = mean free path [m]

$d_o$  = molecular diameter [m]

$n$  = gas density (molecules per cubic meter) [ $\text{m}^{-3}$ ]

The *average velocity* of molecules in the gas phase is a function of the molecular weight of the gaseous species and the average temperature.

$$\overline{V} = 145.5 \sqrt{\frac{T}{MW}} \quad (168.19)$$

where  $V$  = average velocity [m/s].

## Methods for Measuring Subatmospheric Pressures

A wide variety of gages are used in vacuum applications. Gage selection is based upon the vacuum system operating pressure range, the presence of corrosive gases, the requirement for computer interface, and unit cost.

Subatmospheric pressure gages ([Table 168.3](#)) which infer pressure (gas density) from measurement of some physical property of the rarefied gas, have the unfortunate characteristic of



being gas species-sensitive. Most commercial gages are calibrated for nitrogen; calibration curves for other gas species are available. Other factors which influence the readings of pressure gages used in vacuum technology include: location of the gage on the system, cleanliness of the gage elements, orientation of the gage, and environment of the gage (temperature, stray electric and magnetic fields). Gage readings are typically accurate to  $\pm 10\%$  of the reported value.

**Table 168.3** Subatmospheric Pressure Gages

Type	Principle of Operation	Range [torr]
U tube manometer	Liquid surface displacement	1–760
Bourdon tube	Solid surface displacement	1–760
McLeod gage	Liquid surface displacement	$10^{-4}$ – $10^{-1}$
Diaphragm gage	Solid surface displacement	$10^{-1}$ –760
Capacitance manometer	Solid surface displacement	$10^{-3}$ –760
Thermocouple gage	Thermal conductivity of gas	$10^{-3}$ –1
Piraini gage	Thermal conductivity of gas	$10^{-3}$ –760
Spinning rotor gage	Viscosity of gas	$10^{-7}$ – $10^{-2}$
Hot cathode ion gage	Gas ionization cross section	$10^{-12}$ – $10^{-4}$

## Methods for Reducing Pressure in a Vacuum Vessel

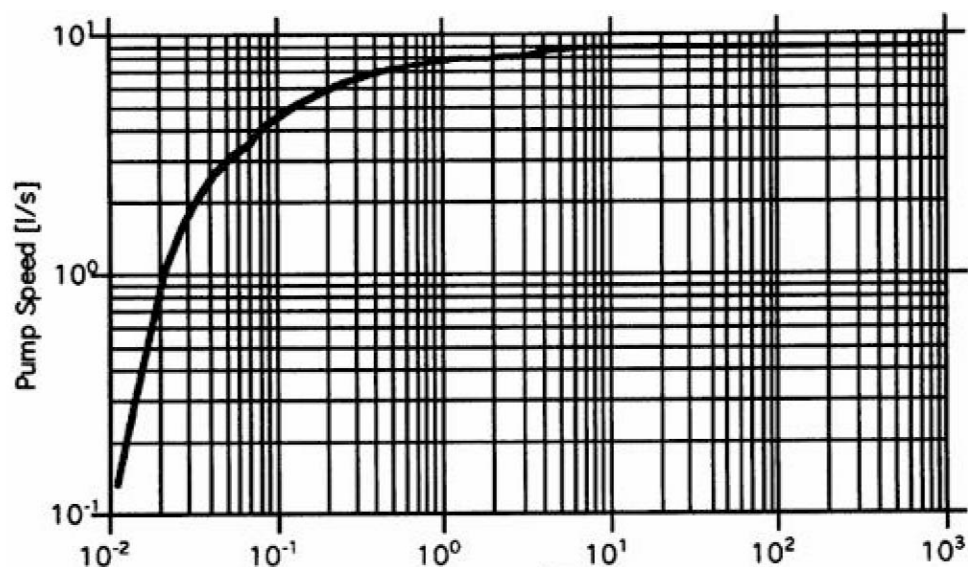
Vacuum pumps (Table 168.4) generally fall into one of two categories: primary (roughing) pumps and secondary (high-vacuum) pumps. Primary pumps are used to evacuate a vacuum vessel from atmospheric pressure to a pressure of about 10 mtorr. Secondary pumps typically operate in the pressure range of from  $10^{-10}$  to  $10^{-3}$  torr.

**Table 168.4** Vacuum Pumps

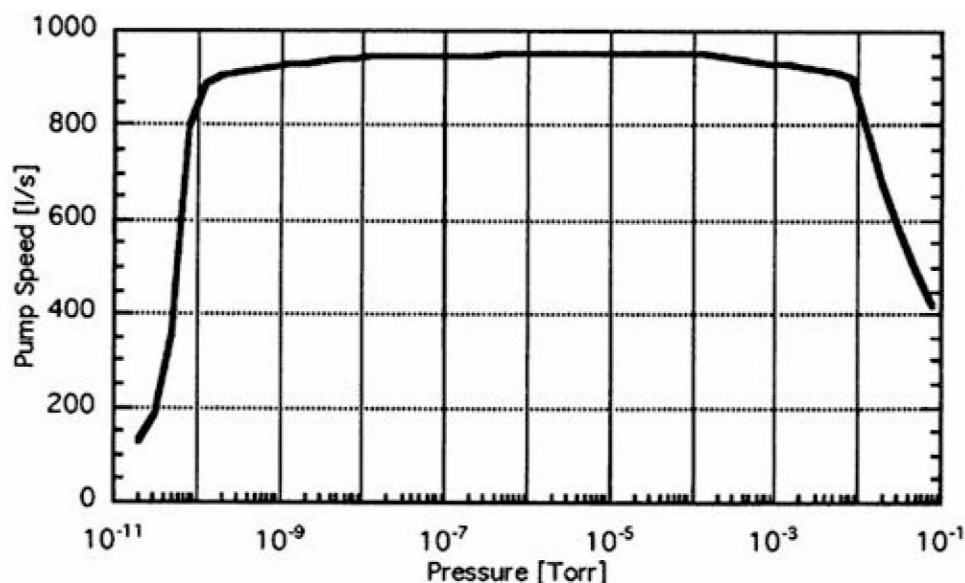
Type	Principle of Operation	Range [torr]
Rotary vane	Positive gas displacement	$10^{-2}$ –760
Rotary piston	Positive gas displacement	$10^{-2}$ –760
Cryo-sorption	Gas capture	$10^{-3}$ –760
Oil vapor diffusion	Momentum transfer	$10^{-9}$ – $10^{-3}$
Turbomolecular	Momentum transfer	$10^{-9}$ – $10^{-3}$
Sputter-ion	Gas capture	$10^{-10}$ – $10^{-3}$
He cryogenic	Gas capture	$10^{-10}$ – $10^{-3}$

Pump speed curves such as those presented in Figs. 168.6 and 168.7 are generally supplied by the pump manufacturer and reflect optimal performance for pumping air.

**Figure 168.6** Pump speed curve typical of a two-stage oil-sealed rotary vane pump.



**Figure 168.7** Pump speed curve typical of a fractionating oil-vapor diffusion pump equipped with a liquid nitrogen cold trap.



## Vacuum System Design

The following relationships provide a first-order method for calculating **conductances**, delivered pump speed, and time to achieve a specified pressure. These relationships assume that conductance elements are of circular cross section and have smooth internal surfaces, and that the gas being pumped is air.

The conductance of tubes and orifices is a function of the flow mode (see Fig. 168.8), which in turn is determined by the average pressure and the conductance element inside diameter. Flow modes may be determined by the following method:

$$D\bar{P} > 0.18 \Rightarrow \text{turbulent flow}$$

$$D\bar{P} < 0.004 \Rightarrow \text{molecular flow} \quad (168.20)$$

where

$D$  = tube inside diameter [in.]

$\bar{P}$  = average pressure [torr]

$$C = \frac{3000\bar{P}D^4 + 80D^3}{L} \quad (168.21)$$

where

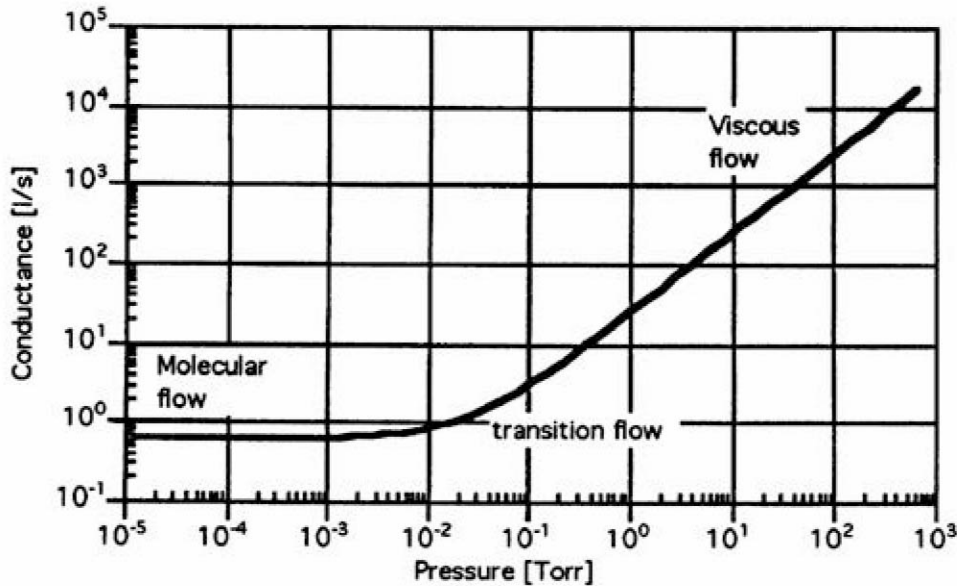
$C$  = conductance [L/s]

$D$  = tube inside diameter [in.]

$\bar{P}$  = average pressure [torr]

$L$  = tube length [in.]

**Figure 168.8** Conductance as a function of pressure for a 1.0" I.D. tube 10 feet long.



For conductance elements connected in series, the total conductance is given by

$$\frac{1}{C_{\text{tot}}} = \sum_{i=1}^n \frac{1}{C_i} \quad (168.22)$$

where

$C_{\text{tot}}$  = total conductance [L/s]

$C_i$  = conductance of  $i$ th tube [L/s]

Manufacturers of vacuum components (valves, traps, etc.) generally publish the conductance of the unit in the molecular flow regime. If this information is unavailable, the maximum conductance of a valve or trap may be estimated using the formula for aperture conductance in molecular flow:

$$C_m^{\text{ap}} = 75A \quad (168.23)$$

where

$C_m^{\text{ap}}$  = molecular flow conductance [L/s]

$A$  = area of aperture [ $\text{in.}^2$ ]

The effect of conductance on the pump speed delivered to a vessel may be evaluated using the following relationship:

$$\frac{1}{S_t} = \frac{1}{C_{\text{tot}}} + \frac{1}{S_p} \quad (168.24)$$

where

$S_t$  = delivered pump speed [L/s]

$S_p$  = speed measured at pump inlet [L/s]

The amount of time required to evacuate a vessel using a primary pump is given by:

$$t = \frac{V}{S_t} \ln \left( \frac{P_1}{P_2} \right) \quad (168.25)$$

where

$t$  = time to pump from  $P_1$  to  $P_2$  [s]

$V$  = volume of vessel [L]

$S_t$  = delivered pump speed [L/s]

$P_1$  = initial pressure [torr]

$P_2$  = final pressure [torr]

The lowest pressure (ultimate pressure) a vacuum system can attain is a function of the total gas load [which is the algebraic sum of gas loads due to leaks, **outgassing** (see [Table 168.5](#)),

permeation, and process gas] and the pumping speed of the secondary pump.

$$Q_{\text{tot}} = S_i P_u \quad (168.26)$$

where

$Q_{\text{tot}}$  = total gas load [torr-L/s]

$P_u$  = ultimate pressure [torr]

**Table 168.5** Outgassing Data for Materials Used in Vacuum Vessels [O'Hanlon, 1989]

Treatment		$q$ [W/m <sup>2</sup> ]	$q$ [T · 1/s·cm <sup>2</sup> ]
Metals			
Aluminum	15 h at 250°C	$5.3 \cdot 10^{-10}$	$3.98 \cdot 10^{-13}$
Aluminum	100 h at 100°C	$5.3 \cdot 10^{-11}$	$3.98 \cdot 10^{-14}$
6061 aluminum	Glow discharge and 200°C bake	$1.3 \cdot 10^{-11}$	$9.75 \cdot 10^{-15}$
Copper	20 h at 100°C	$1.46 \cdot 10^{-9}$	$1.10 \cdot 10^{-12}$
304 stainless steel	30 h at 250°C	$4.0 \cdot 10^{-9}$	$3.00 \cdot 10^{-12}$
316L stainless steel	2 h at 800°C	$4.6 \cdot 10^{-10}$	$3.45 \cdot 10^{-13}$
U15C stainless steel	3 h at 1000°C and 25 h at 360°C	$2.1 \cdot 10^{-11}$	$1.58 \cdot 10^{-14}$
Glasses			
Pyrex glass	Fresh	$9.8 \cdot 10^{-7}$	$7.35 \cdot 10^{-9}$
Pyrex glass	1 month in air	$1.55 \cdot 10^{-6}$	$1.16 \cdot 10^{-9}$
Elastomers			
Viton-A	Fresh	$1.52 \cdot 10^{-3}$	$1.14 \cdot 10^{-6}$
Neoprene	—	$4.0 \cdot 10^{-2}$	$3.00 \cdot 10^{-5}$

## Defining Terms

**Conductance:** Mass throughput divided by the pressure drop across a tube.

**Mole:** Amount of material,  $6.02\text{E}+23$  particles; 1 mole of carbon weighs 12 grams.

**Outgassing:** Evolution of gas from a solid or liquid in a vacuum environment.

## References

- Baumeister, T. 1958. *Marks' Mechanical Engineers' Handbook*, 9th ed. McGraw-Hill, New York.
- Borzileri, C. V. 1993. DOE pressure safety: Pressure calculations. Rev. 10, Appendix C. Lawrence Livermore National Laboratory, Livermore, CA.
- O'Hanlon, J. F. 1989. *A User's Guide to Vacuum Technology*, 2nd ed. John Wiley & Sons, New York.
- Roark, R. J. 1989. Pressure vessels: Pipes. In *Formulas for Stress and Strain*, 6th ed. p. 636. McGraw-Hill, New York.

## Further Information

- Beavis, L. C. and Harwood, V. J. 1979. *Vacuum Hazards Manual*. American Vacuum Society, New York.
- Drinkwine, M. J. and Lichtman, D. 1979. *Partial Pressure Analyzers and Analysis*. American Vacuum Society, New York.
- Dushman, S. and Lafferty, J. M. 1976. *The Scientific Foundations of Vacuum Technique*. John Wiley & Sons, New York.
- Faupel, J. H. 1956. *Trans. Am. Soc. Mech. Eng.* 78(5):1031–64,
- Hablanian, M. H. 1990. *High Vacuum Technology: A Practical Guide*. Marcel Dekker, Inc., New York.
- Harris, N. S. 1989. *Modern Vacuum Practice*. McGraw-Hill, New York.
- Madey, T. E. and Brown, W. C. 1984. *History of Vacuum Science and Technology*. American Vacuum Society, New York.
- Roth, A. 1966. *Vacuum Sealing Techniques*. Pergamon Press, New York.
- Roth, A. 1982. *Vacuum Technology*. Elsevier Publishing Co., Inc., New York.
- Sec. VIII: Pressure vessels, division-1. In *ASME Boiler and Pressure Vessel Code*. ASME, New York (current issue).
- Vossen J. L. and Kern, W. 1978. *Thin Film Processes*. Academic Press, Inc., New York.
- Weissler, G. L. and Carlson, R. W. 1979. *Vacuum Physics and Technology*. Academic Press, Inc., New York.
- Wilson, N. G. and Beavis, L. C. 1979. *Handbook of Leak Detection*. American Vacuum Society, New York.

R. Paul Singh. "Food Engineering"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

### 169.1 Liquid Transport Systems

#### 169.2 Heat Transfer

**R. Paul Singh**

*University of California, Davis*

The food industry relies on a wide range of unit operations to manufacture a myriad of processed foods. In order to design food processes, a practitioner in the food industry uses fundamental principles of chemistry, microbiology, and engineering. During the last 30 years, the food engineering discipline has evolved to encompass several aspects of food processing. The diversity of processes typically employed in a food processing plant is illustrated in [Fig. 169.1](#). Typical food processes may include sorting and size reduction, transport of liquid foods in pipes, heat transfer processes carried out using heat exchangers, separation processes using membranes, simultaneous heat and mass transfer processes important in drying, and processes that may involve a phase change such as freezing. A food engineer uses the concepts common to the fields of chemical, mechanical, civil, and electrical engineering in addition to food sciences to design engineering systems that interact with foods. When foods are used as raw materials they offer unique challenges. Perhaps the most important concern in food processing is the variability in the raw material. To achieve consistency in the final quality of a processed food, the processes must be carefully designed to minimize variations caused by processing. In this chapter, some of the unique engineering issues encountered in food processing will be presented. For more details on the engineering design of food processing systems, the reader is referred to books by Singh and Heldman [1993], Toledo [1991], Brennan *et al.* [1990], Heldman and Singh [1981], Loncin and Merson [1979], and Charm [1978].

## 169.1 Liquid Transport Systems

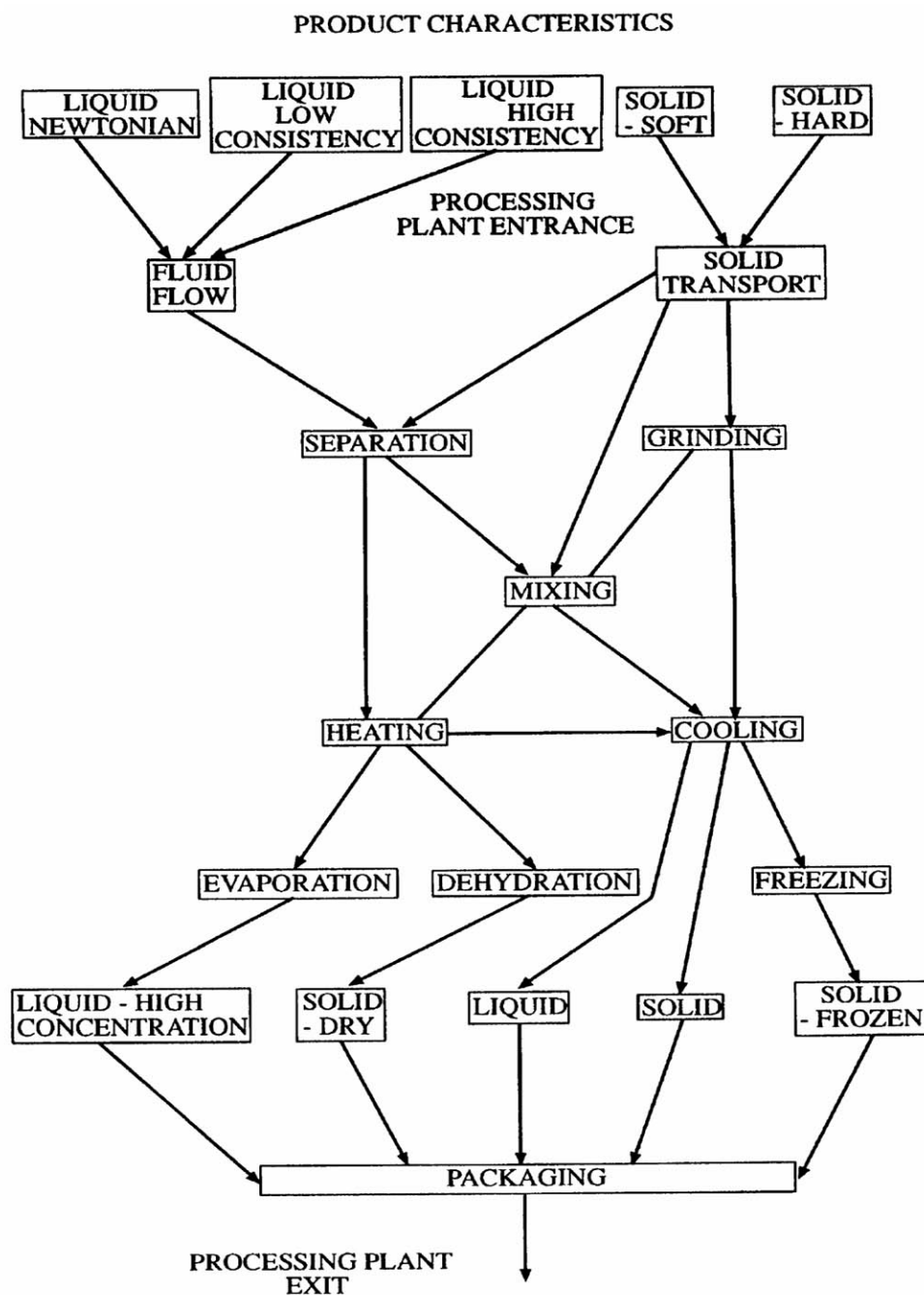
---

In a food processing plant, one of the common operations is the transport of liquid foods from one piece of processing equipment to the next. The characteristics of the liquid food must be known before a liquid transport system can be designed. As seen in [Fig. 169.2](#), a linear relationship exists between the shear stress and shear rate for a **Newtonian liquid**, such as water, orange juice, milk, and honey. The viscosity of a Newtonian liquid is determined from the slope of the straight line. Viscosity is an important property needed in many flow-related calculations. For example, viscosity of a liquid must be known before the volumetric flow rate in a pipe, under laminar conditions, can be calculated from the following equation:

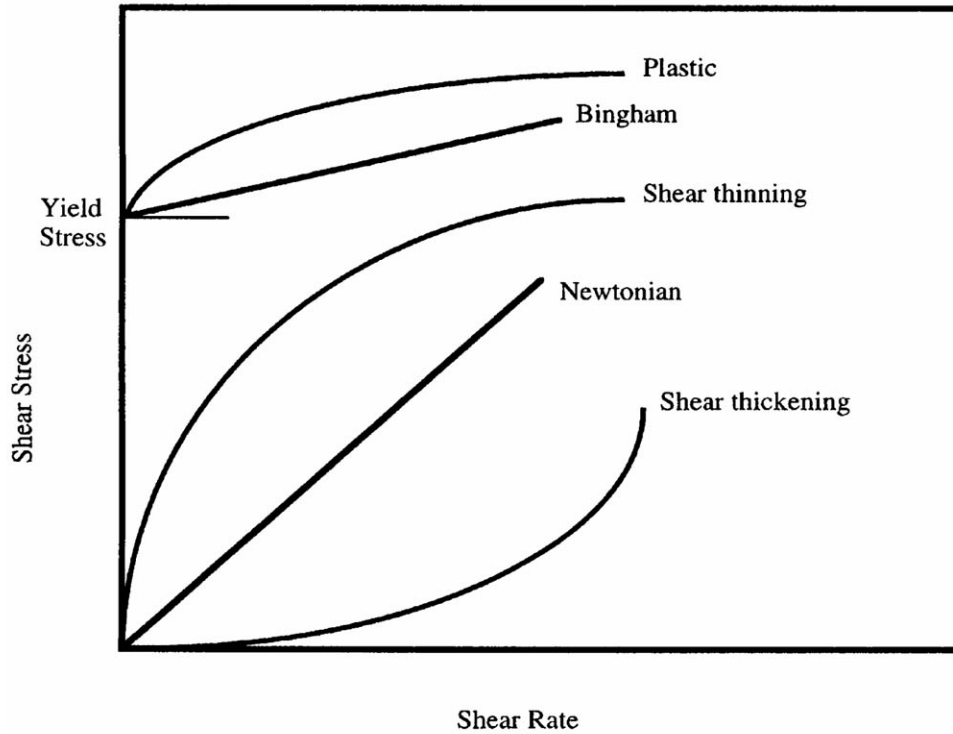
$$\dot{V} = \frac{\pi \Delta P R^4}{8 \mu L} \quad (169.1)$$



**Figure 169.1** Generalized flow in a typical food manufacturing plant. (Source: Heldman, D. R. and Singh, R. P. 1981. *Food Process Engineering*, 2nd ed. AVI, Westport, CT.)



**Figure 169.2** Shear stress–shear rate relationships for Newtonian and non-Newtonian liquids. (Source: Singh, R. P. and Heldman, D. R. 1993. *Introduction to Food Engineering*, 2nd ed. Academic Press, San Diego, CA.)



For **non-Newtonian liquids**, the relationship between shear stress and shear rate is nonlinear. Non-Newtonian liquids are classified as shear thinning, shear thickening, Bingham, and plastic fluids. Both Newtonian and non-Newtonian characteristics can be described by the Herschel–Bulkley model [[Herschel and Bulkley, 1926](#)].

$$\sigma = K \left[ \frac{du}{dy} \right]^n + \sigma_0 \quad (169.2)$$

The values of the consistency coefficient,  $K$ , and the flow behavior index,  $n$ , in Eq. (169.2) are used to differentiate between different types of non-Newtonian liquids ([Table 169.1](#)). A few measured values of  $K$  and  $n$  using capillary and rotational viscometers are shown in [Table 169.2](#).

**Table 169.1** Values of Coefficients in the Herschel–Bulkley Fluid Model

Fluid	$K$	$n$	$s_0$	Typical Examples
Herschel–Bulkley	$>0$	$0 < n < 1$	$>0$	Minced fish paste, raisin paste
Newtonian	$>0$	1	0	Water, fruit juice, honey, milk, vegetable oil
Shear-thinning (pseudoplastic)	$>0$	$0 < n < 1$	0	Applesauce, banana puree, orange juice concentrate
Shear-thickening	$>0$	$1 < n < \infty$	0	Some types of honey, 40% raw corn starch solution
Bingham plastic	$>0$	1	$>0$	Toothpaste, tomato paste

Source: Steffe, J. F. 1992. *Rheological Methods in Food Process Engineering*. Freeman Press, East Lansing, MI.

**Table 169.2** Rheological Properties of Selected Liquid Foods

Product	Temperature(°C)	Composition	Consistency Coefficient (m) (Pa · s <sup>n</sup> )	Flow Behavior Index (n)	Measurement Method
Apple juice	27	20° Brix	0.0021	1.0	Capillary tube
Apple juice	27	60° Brix	0.03	1.0	Capillary tube
Applesauce	25	31.7% T.S.	22.0	0.4	Coaxial cylinder
Applesauce	27	11.6% T.S.	12.7	0.28	Capillary tube
Apricot puree	21	17.7% T.S.	5.4	0.29	Coaxial cylinder
Apricot puree	25	19% T.S.	20.0	0.3	Coaxial cylinder
Banana puree	24	Unknown	6.5	0.458	Coaxial cylinder
Banana puree	24	Unknown	10.7	0.333	Capillary tube
Corn syrup	27	48.4% T.S.	0.053	1.0	Coaxial cylinder
Cream	3	20% fat	0.0062	1.0	Unknown
Cream	3	30% fat	0.0138	1.0	Unknown
Grape juice	27	20° Brix	0.0025	1.0	Capillary tube
Grape juice	27	60° Brix	0.11	1.0	Capillary tube
Honey	24	Normal	5.6	1.0	Capillary tube
Honey	24	Normal	6.18	1.0	Single cylinder
Olive oil	20	Normal	0.084	1.0	Unknown
Peach puree	27	10.0% T.S.	4.5	0.34	Capillary tube
Pear puree	27	14.6% T.S.	5.3	0.38	Capillary tube
Pear puree	27	15.2% T.S.	4.25	0.35	Coaxial cylinder
Pear puree	32	18.31% T.S.	2.25	0.486	Coaxial cylinder
Pear puree	32	45.75% T.S.	35.5	0.479	Coaxial

					cylinder
Skim milk	25	Normal	0.0014	1.0	Unknown
Soy bean oil	30	Normal	0.04	1.0	Unknown
Tomato concentrate	32	5.8% T.S.	0.223	0.59	Coaxial cylinder
Tomato concentrate	32	30% T.S.	18.7	0.4	Coaxial cylinder
Whole milk	20	Normal	0.0212	1.0	Unknown

Adapted from Heldman, D. R. and Singh, R.P. 1981. *Food Process Engineering*, 2nd ed. AVI, Westport, CT.

When designing transport systems for water and other Newtonian liquids, the flow regime is determined by calculating the Reynolds number. For example, a flow with Reynolds number smaller than 2100 is considered to be laminar, whereas a Reynolds number greater than 10 000 indicates turbulent flow. In the case of non-Newtonian liquid foods, a generalized Reynolds number is calculated using both the consistency coefficient  $K$  and the flow behavior index  $n$ . The generalized Reynolds number ( $N_{\text{GRe}}$ ) is defined as

$$N_{\text{GRe}} = \frac{\rho \bar{u}^{2-n} D^n}{2^{n-3} K \left[ \frac{3n+1}{n} \right]^n} \quad (169.3)$$

It is well known that for Newtonian liquids flowing in a pipe, the ratio between the mean velocity,  $\bar{u}$ , and the maximum velocity,  $u_{\text{max}}$ , under laminar conditions is given by

$$\frac{\bar{u}}{u_{\text{max}}} = 0.5 \quad (169.4)$$

Similarly, in the case of turbulent flow, for a Newtonian liquid the velocity ratio is

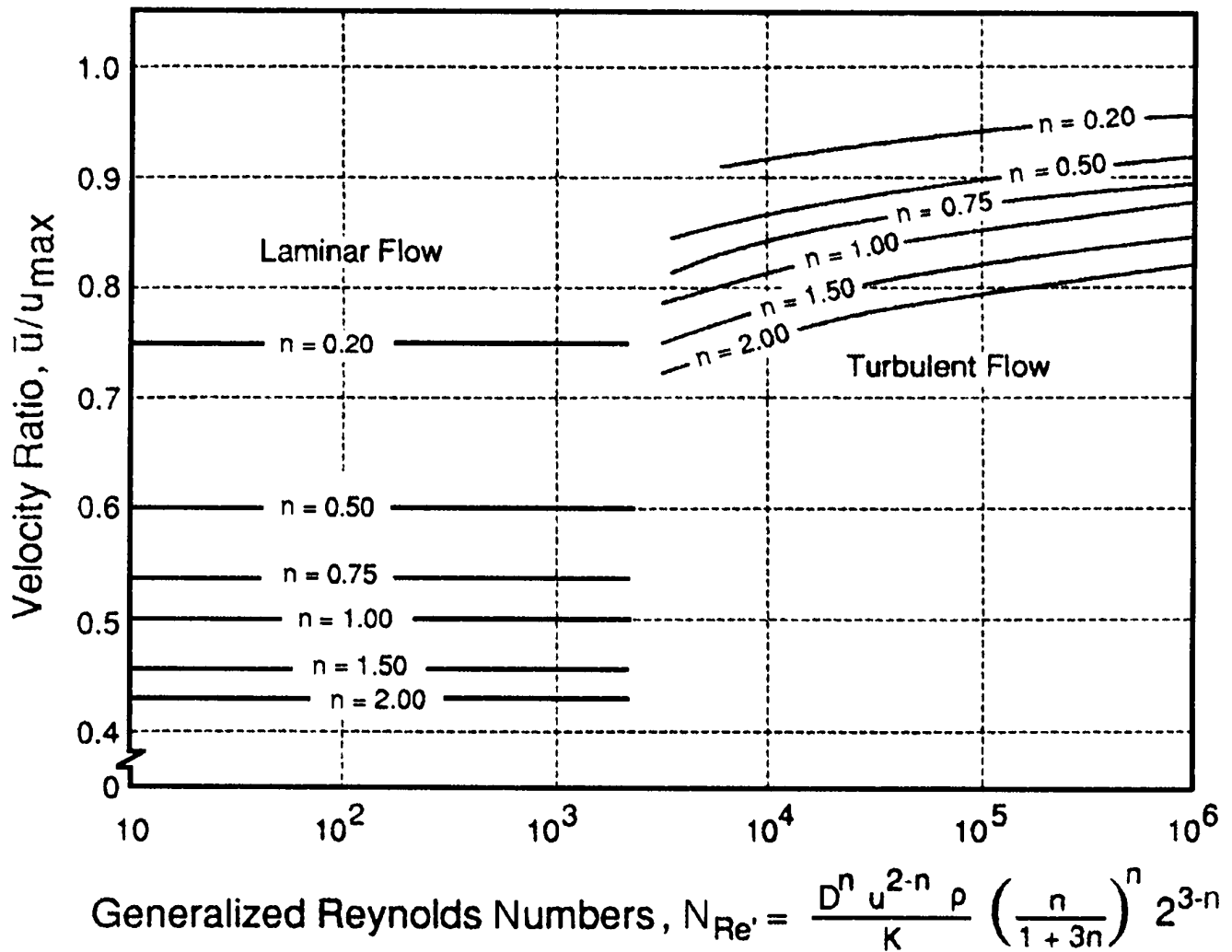
$$\frac{\bar{u}}{u_{\text{max}}} = 0.82 \quad (169.5)$$

In the case of non-Newtonian liquids, the velocity ratio is influenced by the flow behavior index,  $n$ , and is determined as

$$\frac{\bar{u}}{u_{\text{max}}} = \frac{n+1}{3n+1} \quad (169.6)$$

The velocity ratio for non-Newtonian liquids plotted as a function of generalized Reynolds number is shown in Fig. 169.3 [Palmer and Jones, 1976]. Knowledge of such velocity profiles is important in the design of aseptic processing systems for liquid foods.

**Figure 169.3** Plot of velocity ratio versus generalized Reynolds numbers. (Source: Palmer, J. and Jones, V. 1976. Prediction of holding times for continuous thermal processing of power law fluids. *J. Food Sci.* 41(5): 1233.)



## 169.2 Heat Transfer

Heating and cooling processes are widely used in the food industry. The three common modes of heat transfer—conduction, convection, and radiation—play an important role in the processing of foods.

The governing equation for heat transfer in the conductive mode is

$$\frac{\partial T}{\partial t} = \frac{\partial}{\partial x} \left[ \alpha \frac{\partial T}{\partial x} \right] \quad (169.7)$$

Under steady state conditions, Eq. (169.7) reduces to the well-known Fourier law:

$$q_x = -kA \frac{dT}{dx} \quad (169.8)$$

Solution of both transient and steady state equations [Eqs. (169.7) and (169.8)] requires a knowledge of the **physical and thermal properties** of foods. Considerable research has been done to measure food properties such as thermal conductivity, density, specific heat, and thermal diffusivity [Rao and Rizvi, 1995]. A computerized database developed by Singh [1993] contains over 2500 food-property combinations, along with literature citations. Some selected food property values are shown in Tables 169.3, 169.4, and 169.5. It is evident that for most high-moisture foods, the property values are greatly influenced by the presence of water. In fact, many of the empirical models used for predicting thermal properties are based on the amount of water present in a food material.

**Table 169.3** Thermal Diffusivity of Selected Foodstuffs

Product	Water Content(% wt.)	Temperature <sup>a</sup> (°C)	Thermal Diffusivity (×10 <sup>-7</sup> m <sup>2</sup> /s)
Fruits, vegetables, and by-products			
Apple, whole, Red Delicious	85	0–30	1.37
Applesauce	37	5	1.05
	37	65	1.12
	80	5	1.22
	80	65	1.40
	—	26–129	1.67
Avocado, flesh	—	24, 0	1.24
Seed	—	24, 0	1.29
Whole	—	41, 0	1.54
Banana, flesh	76	5	1.18
	76	65	1.42
Beans, baked	—	4–122	1.68
Cherries, tart, flesh	—	30, 0	1.32
Grapefruit, Marsh, flesh	88.8	—	1.27
Grapefruit, Marsh, albedo	72.2	—	1.09
Lemon, whole	—	40, 0	1.07
Lima beans, pureed	—	26–122	1.80
Pea, pureed	—	26–128	1.82
Peach, whole	—	27, 4	1.39
Potato, flesh	—	25	1.70
Potato, mashed, cooked	78	5	1.23
	78	65	1.45
Rutabaga	—	48, 0	1.34
Squash, whole	—	47, 0	1.71
Strawberry, flesh	92	5	1.27
Sugarbeet	—	14, 60	1.26

Sweet potato, whole	—	35	1.06
	—	55	1.39
	—	70	1.91
Tomato, pulp	—	4, 26	1.48
Fish and meat products			
Codfish	81	5	1.22
	81	65	1.42
Corned beef	65	5	1.32
	65	65	1.18
Beef, chuck <sup>b</sup>	66	40–65	1.23
Beef, round <sup>b</sup>	71	40–65	1.33
Beef, tongue <sup>b</sup>	68	40–65	1.32
Halibut	76	40–65	1.47
Ham, smoked	64	5	1.18
Ham, smoked	64	40–65	1.38
Water	—	30	1.48
	—	65	1.60
Ice	—	0	11.82

Source: Singh, R. P. 1982. Thermal diffusivity in food processing. *Food Technology* 36(2):87–91.  
Copyright©1982 by Institute of Food Technologists.

<sup>a</sup>Where two temperatures separated by a comma are given, the first is the initial temperature of the sample and the second is that of the surroundings.

<sup>b</sup>Data are applicable only where juices that exuded during heating remain in the food samples.

**Table 169.4** Thermal Conductivity of Selected Food Products

Product	Moisture Content (%)	Temperature (°C)	Thermal Conductivity (W/m·K)
Apple	85.6	2–36	0.393
Applesauce	78.8	2–36	0.516
Beef, freeze dried			
1000 mm Hg pressure	—	0	0.065
0.001 mm Hg pressure	—	0	0.037
Beef, lean			
Perpendicular to fibers	78.9	7	0.476
Perpendicular to fibers	78.9	62	0.485
Parallel to fibers	78.7	8	0.431
Parallel to fibers	78.7	61	0.447
Butter	15	46	0.197
Cod	83	2.8	0.544
Egg, white	—	36	0.577

Egg, yolk	—	33	0.338
Fish muscle	—	0–10	0.557
Grapefruit, whole	—	30	0.45
Honey	12.6	2	0.502
	80	2	0.344
Juice, apple	87.4	20	0.559
	36.0	20	0.389
Milk	—	37	0.530
Milk, condensed	90	24	0.571
Milk, skimmed	—	1.5	0.538
Milk, nonfat dry	4.2	39	0.419
Olive oil	—	15	0.189
	—	100	0.163
Pork			
Perpendicular to fibers	75.1	6	0.488
		60	0.54
Parallel to fibers	75.9	4	0.443
		61	0.489
Pork fat	—	25	0.152
Potato, raw flesh	81.5	1–32	0.554
Potato, starch gel	—	1–67	0.04
Poultry, broiler muscle	69.1–74.9	4–27	0.412
Salmon			
Perpendicular	73	4	0.502
Salt	—	87	0.247
Strawberries	—	14 to 25	0.675
Sugars	—	29–62	0.087–0.22
Turkey, breast			
Perpendicular to fibers	74	3	0.502
Parallel to fibers	74	3	0.523
Vegetable and animal oils	—	4–187	0.169
Wheat flour	8.8	43	0.45
		65.5	0.689
		1.7	0.542
Whey		80	0.641

Source: Reidy, G. A. 1968. *Thermal Properties of Foods and Methods of Their Determination*. M.S. thesis, Michigan State University, East Lansing, MI.

**Table 169.5** Specific Heats of Foods

Product	Water	Experimental
Beef	68.3	3.52
Butter	15.5	2.051–2.135



Milk, whole	87.0	3.852
Skim milk	90.5	3.977–4.019
Beef, lean	71.7	3.433
Potato	79.8	3.517
Apple, raw	84.4	3.726–4.019
Bacon	49.9	2.01
Cucumber	96.1	4.103
Blackberry, syrup pack	76.0	
Potato	75.0	3.517
Veal	68.0	3.223
Fish	80.0	3.60
Cheese, cottage	65.0	3.265
Shrimp	66.2	3.014
Sardines	57.4	3.014
Beef, roast	60.0	3.056
Carrot, fresh	88.2	3.81–3.935

Adapted from Reidy, G. A. 1968. *Thermal Properties of Foods and Methods of Their Determination*. M.S. thesis, Michigan State University, East Lansing, MI.

Thermal conductivity of foods can be estimated from an empirical equation developed by Sweat [1995] using 430 data points.

$$k = 0.25m_c + 0.155m_p + 0.16m_f + 0.135m_a + 0.58m_m \quad (169.9)$$

A more rigorous model of thermal conductivity has been suggested by Kopelman [1966] to account for the nonisotropic nature of foods. For example, for a food material containing fibers, when the fibers in a food are parallel to the applied heat transfer, the thermal conductivity may be calculated as

$$k_{\parallel} = k_L [1 - N^2(1 - k_S/k_L)] \quad (169.10)$$

If the food fibers are perpendicular to the direction of heat transfer,

$$k_{\perp} = k_L \frac{1 - Q''}{1 - Q''(1 - N)} \quad (169.11)$$

$$Q'' = \frac{N}{(1 - k_S/k_L)} \quad (169.12)$$

Specific heat of a food material is mainly influenced by the components of the food. Thus, specific heat may be estimated from the knowledge of food composition using the following equation:

$$c_p = 1.424m_c + 1.549m_p + 1.675m_f + 0.837m_a + 4.187m_m \quad (169.13)$$

Food composition values are available in a computerized database [Singh, 1993] and *Agricultural Handbook Number 8* [Watt and Merrill, 1975]. Composition of selected foods is given in Table 169.6.

**Table 169.6** Composition of Selected Foods

Food	Water (%)	Protein (%)	Fat (%)	Carbohydrate (%)	Ash (%)
Apples, fresh	84.4	0.2	0.6	14.5	0.3
Applesauce	88.5	0.2	0.2	10.8	0.6
Asparagus	91.7	2.5	0.2	5.0	0.6
Beans, lima	67.5	8.4	0.5	22.1	1.5
Beef, hamburger, raw	68.3	20.7	10.0	0.0	1.0
Bread, white	35.8	8.7	3.2	50.4	1.9
Butter	15.5	0.6	81.0	0.4	2.5
Cod	81.2	17.6	0.3	0.0	1.2
Corn, sweet, raw	72.7	3.5	1.0	22.1	0.7
Cream, half-and-half	79.7	3.2	11.7	4.6	0.6
Eggs	73.7	12.9	11.5	0.9	1.0
Garlic	61.3	6.2	0.2	30.8	1.5
Lettuce, Iceburg	95.5	0.9	0.1	2.9	0.6
Milk, whole	87.4	3.5	3.5	4.9	0.7
Orange juice	88.3	0.7	0.2	10.4	0.4
Peaches	89.1	0.6	0.1	9.7	0.5
Peanuts, raw	5.6	26.0	47.5	18.6	2.3
Peas, raw	78.0	6.3	0.4	14.4	0.9
Pineapple, raw	85.3	0.4	0.2	13.7	0.4
Potatoes, raw	79.8	2.1	0.1	17.1	0.9
Rice, white	12.0	6.7	0.4	80.4	0.5
Spinach	90.7	3.2	0.3	4.3	1.5
Tomatoes	93.5	1.1	0.2	4.7	0.5
Turkey	64.2	20.1	14.7	0.0	1.0
Turnips	91.5	1.0	0.2	6.6	0.7
Yogurt, whole milk	88.0	3.0	3.4	4.9	0.7

Newton's law of cooling is used when the mode of heat transfer is by convection. Thus,

$$q = hA(T_p - T_\infty) \quad (169.14)$$

The convective heat transfer coefficient is determined using the lumped heat capacity method, as illustrated by Singh and Heldman [1993]. One of the key pieces of processing equipment used in the heating and cooling of foods is a heat exchanger. In the case of Newtonian liquids, both plate- and tubular-type heat exchangers are commonly used. For example, milk pasteurization is done using plate-type heat exchangers. When working with non-Newtonian liquids such as fruit purees

and pastes, or liquid foods that contain food particulates such as beef chunks in a gravy, the scraped-surface heat exchangers are used. To calculate the rate of heat transfer in Eq. (169.14) a value for the convective heat transfer coefficient must be known. For a scraped-surface heat exchanger the following relationship is recommended:

$$h = 1.2 \left( \frac{k}{D} \right) N_{\text{Re}}^{0.5} N_{\text{Pr}}^{0.33} N_m^{0.26} \quad (169.15)$$

A more complete and rigorous analysis of non-Newtonian heat transfer in scraped-surface heat exchangers is given by Harrod [1987].

Unsteady state heat transfer is important in many food processing applications. Reactions occurring in the initial stages of a heating or cooling process can significantly affect the quality attributes of foods. For example, in the food canning process, knowledge of temperature and time inside the can during the heating and cooling processes is important when designing thermal processes that ensure required microbial sterilization. Standard methods involving Heisler charts may be used to determine the temperature history of conduction heating foods [Holman, 1990]. Due to the low thermal diffusivities of foods, obtaining values of Fourier number and temperature ratios from the Heisler charts is cumbersome. Numerical methods are often employed to solve the governing heat transfer equation [Heldman and Singh, 1981].

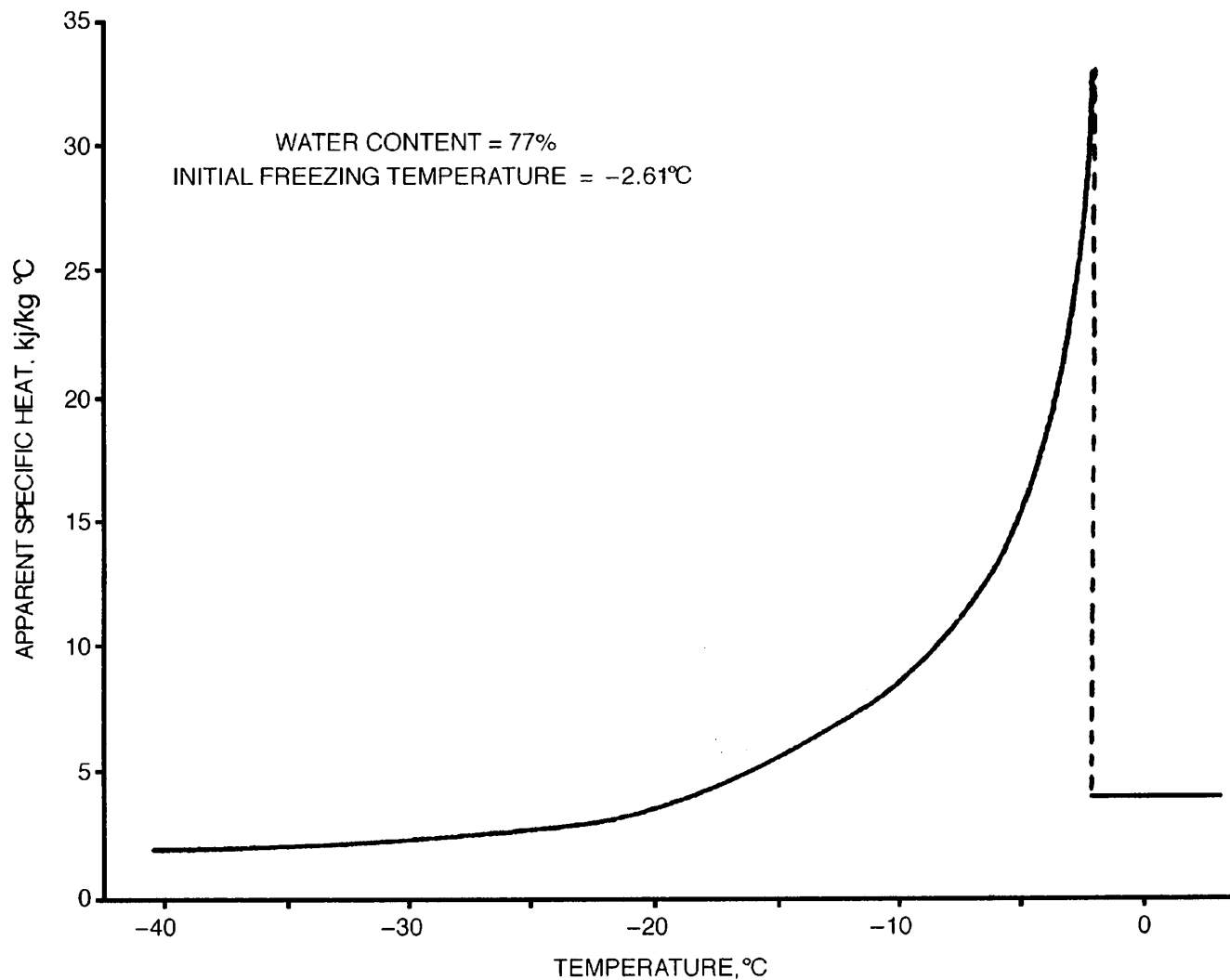
For heating or cooling processes without phase change, the thermal diffusivity is usually a weak function of temperature. However, when a phase change occurs during a process, such as in freezing or boiling, the properties of the food material change, often dramatically. In such cases the thermal diffusivity is a strong function of temperature.

$$\alpha(T) = \frac{k(T)}{\rho(T)c_{pa}(T)} \quad (169.16)$$

Due to the nonlinearities introduced by the variable properties, more rigorous approaches are used to calculate heat transfer. In the case of freezing, food properties are influenced by the state of water. For example, a dramatic change in the specific heat of sweet cherries as a function of temperature in the freezing zone is shown in Fig. 169.4. The freezing time,  $t_F$ , may be estimated using the following equation:

$$t_F = \frac{\rho H_L}{T_F - T_\infty} \left( \frac{\text{Pa}}{h_c} + \frac{\text{Ra}^2}{k} \right) \quad (169.17)$$

**Figure 169.4** Predicted apparent specific heat of frozen sweet cherries as a function of temperature.  
(Source: Heldman, D. R. and Lund, D. B. 1992. *Handbook of Food Engineering*. Marcel Dekker, New York.)



In freezing of foods there is a phase change of water into ice. Other processes where such complexities are encountered include thawing, dehydration, and evaporation. Design considerations of such processes are elaborated in Heldman and Lund [1992] and Singh and Mannapperuma [1990].

The preceding discussion provides ample evidence that food processing offers complex situations that require the use of advanced mathematical skills to develop appropriate solutions. The data on food properties presented in this chapter show the diversity of input conditions prevalent in the design and analysis of food processing operations.

## Defining Terms

**Newtonian liquids:** Newtonian liquids exhibit a linear relationship between shear stress and rate of shear. The slope of the straight line passing through the origin of the axis is used to calculate viscosity of the liquid. Examples of Newtonian liquids include water, milk, and orange juice.

**Non-Newtonian liquids:** Non-Newtonian liquids exhibit a nonlinear relationship between shear stress and rate of shear. Furthermore, these liquids may be classified as time-independent or time-dependent non-Newtonian liquids. The rheological properties that describe non-Newtonian liquids are consistency coefficient, flow behavior index, and yield stress. Examples of non-Newtonian liquids are minced fish paste, apple sauce, banana puree, and tomato paste.

**Physical and thermal properties:** Physical and thermal properties of foods include thermal diffusivity, thermal conductivity, specific heat, density, and viscosity. A knowledge of these properties is essential in designing food processes and equipment.

## Nomenclature

Symbol	Quantity	Unit
$A$	area (normal to $x$ direction) through which heat flows	(m <sup>2</sup> )
$a$	product thickness for an infinite slab, diameter for an infinite cylinder, and diameter for a sphere	
$\alpha$	thermal diffusivity	(m <sup>2</sup> /s)
$c_{pa}$	specific heat capacity	(kJ/kg°C)
$D$	pipe diameter	(m)
$\frac{du}{dy}$	strain	s <sup>-1</sup>
$\Delta P$	pressure drop	(Pa)
$h_c$	convective heat-transfer coefficient	
$H_L$	latent heat of fusion	
$K$	consistency coefficient	(Pa·s <sup><math>n</math></sup> )
$k$	thermal conductivity	(W/m·°C)
$k_L$	thermal conductivity of liquid component	
$k_S$	thermal conductivity of solid component	
$L$	length of pipe	(m)
$\mu$	viscosity of liquid	(Pa·s)
$m_a$	mass fraction of ash	
$m_c$	mass fraction of carbohydrate	
$m_f$	mass fraction of fat	
$m_m$	mass fraction of water	
$m_p$	mass fraction of protein	
$N^2$	volume fraction of solids or discontinuous phase in the fibrous product	
$N_m$	number of blades on the mutator	
$N_{Nu}$	Nusselt number	
$N_{Pr}$	Prandtl number	
$N_{Re}$	Reynolds number	
$n$	flow behavior index	(dimensionless)
$P$	1/2, $R = 1/8$ for infinite plate; $P = 1/4$ , $R = 1/16$ for infinite cylinder $P = 1/6$ , $R = 1/24$ for sphere	

$q$	the rate of heat transfer	(W)
$q_x$	rate of heat flow in $x$ direction by conduction	(W)
$R$	radius of pipe	(m)
$\rho$	density	(kg/m <sup>3</sup> )
$\sigma$	shear stress	
$\sigma_0$	yield stress	
$T$	temperature	(°C)
$T_\infty$	surrounding temperature	(°C)
$T_F$	freezing point	(°C)
$T_p$	surface temperature	(°C)
$t$	time	(s)
$\bar{u}$	mean velocity	(m/s)
$\dot{V}$	volumetric flow rate	(m <sup>3</sup> /s)
$x$	length	(m)

---

## References

- Brennan, J. G., Butters, J. R., Cowell, N. D., and Lilly, A. E. V. 1990. *Food Engineering Operations*, 3rd ed. Elsevier, New York.
- Chandra, P. K. and Singh R. P., 1994. *Applied Numerical Methods in Food and Agricultural Engineering*. CRC Press, Boca Raton, FL.
- Charm, S. E. 1978. *The Fundamentals of Food Engineering*, 3rd ed. AVI, Westport, CT.
- Harrod, M. 1987. Scraped surface heat exchanger: A literature survey of flow patterns, mixing effects, residence time distribution, heat transfer and power requirements. *J Food Proc. Eng.* 9(1):1–62.
- Heldman, D. R. and Lund, D. B. 1992 *Handbook of Food Engineering*. Marcel Dekker, New York.
- Heldman, D. R. and Singh, R. P. 1981. *Food Process Engineering*, 2nd ed. AVI, Westport, CT.
- Herschel, W. H. and Bulkley, R. 1926. Konsistenzmessungen von gummi-benzollösungen. *Kolloid-Zeitschr.* 39:291.
- Holman, J. P. 1990. *Heat Transfer*, 7th ed. McGraw-Hill, New York.
- Kopelman, I. J. 1966. *Transient Heat Transfer and Thermal Properties in Food Systems*. Ph.D. thesis, Michigan State University, East Lansing, MI.
- Loncin, M. and Merson, R. L. 1979. *Food Engineering: Principles and Selected Applications*. Academic Press, New York.
- Palmer, J. and Jones, V. 1976. Prediction of holding times for continuous thermal processing of power law fluids. *J. Food Sci.* 41(5):1233.
- Rao, M. A. and Rizvi, S. S. H. 1995. *Engineering Properties of Foods*, 2nd ed. Marcel Dekker, New York.
- Reidy, G. A. 1968. *Thermal Properties of Foods and Methods of their Determination*. M.S. thesis, Michigan State University, East Lansing, MI.
- Singh, R. P. 1993. *A Computerized Database of Food Properties*. CRC Press, Boca Raton, FL.

- Singh, R. P. and Heldman, D. R. 1993. *Introduction to Food Engineering*, 2nd ed. Academic Press, San Diego, CA.
- Singh, R. P. and Mannapperuma, J. D. 1990. Developments in food freezing. In *Biotechnology and Food Process Engineering*, ed. H. G. Schwartzberg and M. A. Rao, pp.309–358. Marcel Dekker, New York.
- Steffe, J. F. 1992. *Rheological Methods in Food Process Engineering*. Freeman Press, East Lansing, MI.
- Sweat, V. E. 1995. Thermal properties of foods. In *Engineering Properties of Foods*, ed. M. A. Rao and S. S. H. Rizvi, 2nd ed, pp. 99–138. Marcel Dekker, New York.
- Toledo, R. T. 1991. *Fundamentals of Food Process Engineering*, 2nd ed. Van Nostrand Reinhold, New York.
- Watt, B. K. and Merrill, A. L. 1975. *Composition of Foods. Agriculture Handbook No. 8*. United States Department of Agriculture, Washington, D. C.

## Further Information

The following references contain a wealth of information from proceedings of recent conferences and collaborative projects on engineering and food:

- Jowitt, R., Escher, F., Hallstrom, B., Meffert, H. F. T., Spiess, W. E. L., and Gilbert, V. (Eds.) 1983. *Physical Properties of Foods*. Applied Science, England.
- Jowitt, R., Escher, F., Kent, M., McKenna, B., and Roques, M. (Eds.) 1987. *Physical Properties of Foods* 3/42. Elsevier Applied Science, London.
- Le Maguer, M. and Jelen, P. (Eds.) 1986. *Food Engineering and Process Applications. Vol 1. Transport Phenomena*. Elsevier Applied Science, London.
- Le Maguer, M. and Jelen, P. (Eds.) 1986. *Food Engineering and Process Applications. Vol 2. Unit Operations*. Elsevier Applied Science, London.
- Singh, R. P. and Medina, A. G. (Eds.) 1989. *Food Properties and Computer-Aided Engineering of Food Processing Systems*. Kluwer Academic, Dordrecht, The Netherlands.
- Singh, R. P. and Oliveira, F. (Eds.) 1994. *Minimal Processing of Foods and Process Optimization*. CRC Press, Boca Raton, FL.
- Singh R. P. and Wirakartakusumah, M. A. (Eds.) 1992. *Advances in Food Engineering*. CRC Press, Boca Raton, FL.
- Singh, R. P. (Ed.) 1986. *Energy in Food Processing*. Elsevier, Amsterdam, The Netherlands.
- Spiess, W. E. L. and Schubert, H. (Eds.) 1990. *Engineering and Food. Vol.1. Physical Properties and Process Control*. Elsevier Applied Science, London.
- Spiess, W. E. L. and Schubert, H. (Eds.) 1990. *Engineering and Food. Vol.2. Preservation Processes and Related Techniques*. Elsevier Applied Science, London.
- Spiess, W. E. L. and Schubert, H. (Eds.) 1990. *Engineering and Food. Vol.3. Advanced Techniques*. Elsevier Applied Science. London.
- Yano, T., Matsuno, R., and Nakamura, K. (Eds.) 1994. *Developments in Food Engineering. Part 1*. Blackie Academic and Professional, London.
- Yano, T., Matsuno, R., and Nakamura, K. (Eds.) 1994. *Developments in Food Engineering. Part 2*. Blackie Academic and Professional, London.

In addition to the above references, the following journals contain research papers on food engineering topics: *Journal of Food Engineering*, *Journal of Food Process Engineering*, *Journal of Food Processing and Preservation*, *Journal of Texture Studies*, and *Journal of Food Science*.



Hills, D. J. "Agricultural Engineering"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## 170.1 Equipment Sizing Criteria

Field Capacity • Power Requirements

## 170.2 Equipment Selection

Soil Tillage • Crop Planting • Crop Harvest

### David J. Hills

*University of California, Davis*

Agricultural engineering is the application of engineering principles and animal/plant biology to the production, handling, processing, packaging, and use of agricultural materials. Although agricultural engineering encompasses a broad area, this chapter focuses only on crop production and specifically on equipment associated with cultural practices. The three paramount operations in crop production are soil tillage, planting, and harvesting. Additionally, judicious attention paid to water, nutriment, and plant protection can maximize crop yields.

Selection of agricultural equipment requires the consideration of three parameters that can lead to high performance: machine, power, and operation. Measures of machine performance are the rate and quality at which the operations are accomplished. Rate is an important measure because agriculture is sensitive to changeable weather and crop ripeness. Most biological materials are fragile and many are perishable, so the amount of product damage or reduction in product quality caused by a machine's operation is important. The second performance parameter that must be considered is the effectiveness with which power is applied to accomplish an agricultural operation. Selecting the proper implement for a certain task and matching the implement to the engine power are critical for obtaining high power efficiency. The final performance parameter refers to the machine operator. Knowledge of the machine's mechanisms and of the overall agricultural enterprise is needed by the operator. Sensors and automatic controls are incorporated in modern equipment for additionally improving operator efficiency.

## 170.1 Equipment Sizing Criteria

---

### Field Capacity

**Field capacity** refers to the amount of processing that a machine can accomplish per hour on either an area or a material basis. These two capacities are expressed as follows:

$$C_f = \frac{SWE}{10} \quad (170.1)$$

$$C_m = \frac{SWYE}{10} \quad (170.2)$$

where

$C_f$  = field capacity on an area basis (ha/h)  
 $C_m$  = field capacity on a material basis (t/h)  
 =  
 $S$  = travel speed (km/h)  
 $W$  = machine working width (m)  
 $Y$  = crop yield (t/ha)  
 $E$  = field efficiency (decimal)

Theoretical field capacity is used to describe a machine's capacity when the field efficiency term is equal to 1.0. This capacity implies that the machine is utilizing its full width and assumes no interruption for turns or other idle time. For cultivators and many harvesters, which work in rows, the machine working width is equal to the row spacing times the number of rows processed in each pass. Operator performance is not perfect, however, so less than the full width of such machines is used in order to ensure coverage of the entire land area; that is, there is some overlapping on each pass. A range of efficiency values for various field operations is provided in [Table 170.1](#). The actual values depend on operator skill, equipment condition, and field, crop, and environmental conditions.

**Table 170.1** Typical Field Equipment Efficiencies and Operating Speeds

Machine	Field Efficiency (%)		Speed (km/h)	
	Range	Typical	Range	Typical
Tillage				
Cultivator (field)	70–90	85	5.0–13.0	9.0
Cultivator (row crop)	70–90	80	4.0–8.0	5.5
Harrow (disk)	70–90	80	5.0–9.9	6.5
Harrow (spiketooth)	70–90	85	5.0–9.5	8.0
Harrow (springtooth)	70–90	85	5.0–9.5	8.0
Landplane	70–90	85	3.0–8.0	6.5
Plow (chisel)	70–90	85	6.5–10.5	10.5
Plow (disk)	70–90	85	3.5–9.5	7.0
Plow (moldboard)	70–90	80	5.0–9.5	7.0
Rotary hoe	70–85	80	8.0–15.5	11.0
Rotary tiller	70–90	85	1.5–7.0	5.0
Subsoiler/ripper	75–85	80	3.0–5.0	4.0
Planters				

Grain drill	65–80	70	4.0–9.0	6.5
Row crop planter	50–75	60	3.0–6.5	4.0
Harvesters				
Combine	65–80	70	3.0–11.0	5.0
Corn picker	60–75	65	3.0–6.5	4.0
Harvester—cotton	60–75	70	3.0–6.5	5.0
Harvester—potato	55–70	60	2.5–6.5	3.0
Harvester—tomato	55–70	60	2.5–6.5	4.0
Hay baler	60–85	75	4.0–8.0	5.5
Mower	75–85	80	6.5–11.0	8.0
Miscellaneous				
Mower—flail	75–90	85	5.0–7.0	5.5
Sprayer—boom type	50–80	65	5.0–11.0	10.5
Spreader—fertilizer	70–70	70	5.0–8.0	7.0

Data from *Fundamentals of Machine Operation*[Anonymous, 1976]; Kepner *et al.*[1978]; Srivastava *et al.* [1993]; and *Standards 1993*[Anonymous, 1993].

Typical operating speeds for various machines are listed in Table 170.1. Travel speeds for harvesters and other machines that process a product are limited by their materials-handling capacity. For machines that do not process a product, such as tillage machines, the speed is limited by other factors, such as available power, quality of the work, and safety.

## Power Requirements

The power required for an agricultural machine is determined by its intended use. Tractors, for example, typically provide power to implements in three forms: drawbar, rotary, and hydraulic. Pulled or towed implements are powered through the traction of drive wheels and the pull, or **draft**, from the drawbar. Rotary power is obtained from the power takeoff (PTO) shaft. Either linear or rotary power can be produced by a tractor's hydraulic system. These three power terms are defined in Eq. (170.3), (170.4), (170.5), respectively.

$$P_{db} = \frac{D_u W S}{3.6} \quad (170.3)$$

where

- $P_{db}$  = **drawbar power** (kW)
- $D_u$  = unit draft of the implement (kN)
- $W$  = machine working width (m)
- $S$  = travel speed (km/h)

$$P_{pto} = \frac{TN}{9.5} \quad (170.4)$$

where

$P_{pto}$  = **PTO power**(kW)

$T$  = torque (N-m)

$N$  = revolutions per minute (rpm)

$$P_{hyd} = \frac{pQ}{1000} \quad (170.5)$$

where

$P_{hyd}$  = hydraulic power (kW)

$p$  = pressure of pumped oil (kP)

$Q$  = oil flow rate (L/s)

Total power requirement for operating implements is the sum of implement power components converted to equivalent PTO power.

$$P_T = 1.15 \left[ \frac{P_{db}}{K_t} + P_{pto} + P_{hyd} \right] \quad (170.6)$$

where

$P_T$  = total implement power requirement (kW)

1.15 = factor that adds 15% to total power for acceleration, slope, and so on

$K_t$  = tractive and transmission efficiency (decimal)

Values for the equation variables can be obtained directly from the implement manufacturer or can be estimated from typical values as listed in [Table 170.2](#). The major factors influencing draft on tillage tools are soil characteristics, forward speed, and crop resistance. The draft for most pull-type, nontillage implements is in the form of rolling resistance.

**Table 170.2** Draft and Energy Requirements for Selected Field Equipment Operated at 5 km/h

Machine	Unit Draft (kN/m)	Energy or Work (kW-h/ha)
Tillage		
Cultivator (field)	0.9–4.4	2.4–12.0
Cultivator (row crop)	0.6–1.2	1.6–3.3
Harrow (disk)	0.7–1.5	2.0–4.0
Harrow (spiketooth)	0.3–0.9	0.7–2.4
Harrow (springtooth)	1.0–4.4	2.1–12.2
Landplane	4.4–11.7	12.2–31.3
Plow (chisel—18 to 23 cm)	2.9–13.1	8.1–36.9
Plow (moldboard or disk)		

light soils—18 cm depth	3.2–6.3	8.7–17.5
medium soils—18 cm depth	5.3–9.5	14.6–25.8
heavy soils—18 cm depth	8.5–16.6	22.1–46.1
Rotary hoe	0.4–0.9	1.3–2.4
Rotary tiller—10 cm forward slice	12.2–24.5	25.8–51.6
Subsoiler/Ripper 2 m spacing		
light soils—40 cm depth	16.0–26.3/unit	7.2–12.0
medium soils—40 cm depth	23.3–36.5/unit	10.1–15.7
Planters		
Grain drill	0.4–1.5	1.1–3.9
Row crop planter—1 m spacing	0.5–0.8/row	1.1–2.4
PTO Power (kW/m or kW/row)		
Harvesters		
Combine—small grain	3.6–11.0	7.2–22
Corn picker	1.5–3.7/row	4.4–8.8
Harvester—cotton (spindle)	7.5–11.2/row	12.9–18.4
Harvester—potato	0.7–1.5/row	8.3–13.8
Hay baler		1.2–2.0
Rotary mower (grass, legumes)	7.3–19.6	9.2–24
Miscellaneous		
Mower—flail		0.9–2.0
Sprayer—boom type	0.2 kW	0.02–0.4
Spreader—fertilizer	0.3–1.2	0.9–3.1

Data from *Fundamentals of Machine Operation*[[Anonymous, 1976](#)]; Kepner *et al.*[[1978](#)]; Srivastava *et al.* [[1993](#)]; and *Standards 1993*[[Anonymous, 1993](#)].

Approximate values for tractive and transmission efficiency for two- and four-wheel-drive tractors and crawler tractors are provided in [Table 170.3](#). As shown, tractive efficiencies for four-wheel-drive tractors are somewhat higher than those for two-wheel-drive tractors, especially for soft soils. Crawler-type tractors seldom have more than 5% slip, even on soft soils.

**Table 170.3** Tractor Tractive and Transmission Coefficients,  $K_t$

Soil Condition/Tractive Condition	Two-Wheel Drive	Four-Wheel Drive	Crawler
Concrete	0.87	0.88	—
Firm, untilled	0.72	0.78	0.82
Tilled, reasonably firm	0.67	0.75	0.80
Freshly plowed, soft	0.55	0.70	0.78

Data from Zoz [[1987](#)].

Fuel requirements for tractors can be estimated from [Table 170.4](#), provided the maximum PTO power rating and the actual PTO power requirement [as calculated by Eq. (170.6)] are known.

The drawbar power is always less than PTO power because of drive-wheel slippage, tractor rolling resistance, and friction losses in the drive between the engine and the wheels. The sum of these losses forms the basis of the tractive and transmission coefficients listed in [Table 170.3](#). These coefficients are essentially the ratios of drawbar power to PTO power.

**Table 170.4** Tractor Fuel Conversion, PTO kW-h/L

Loading, % of Maximum PTO Power	Gasoline	Diesel	LP Gas
100	1.90	2.57	1.57
80	1.74	2.50	1.49
60	1.50	2.26	1.34
40	1.16	1.87	1.09
20	0.76	1.27	0.74

Data from Kepner *et al.* [1978] and Hunt [1973].

## 170.2 Equipment Selection

### Soil Tillage

Tillage is used for seedbed preparation, weed control, incorporation of crop residues and fertilizer materials, breaking soil crusts and hardpans to improve water penetration and aeration, and shaping the soil for irrigation and erosion control. The tillage requirement is determined by the type of crop, the soil type, and the field conditions. A tillage implement consists of a single tool or a group of tools, together with the associated frame, wheels, hitch, control and protection devices, and power transmission components.

Tillage operations for seedbed preparation are often classified as primary or secondary. A primary tillage operation constitutes the initial, major soil-working operation after harvest of the previous crop. It is normally designed to reduce soil strength, cover plant materials, and rearrange soil aggregates. The main objective of secondary tillage is to break down large clods and to prepare a seedbed ideal for planting. An ideal seedbed provides for good seed-to-soil contact, conserves moisture needed for germination, and allows for vigorous and uninhibited root and shoot growth.

Implements used for primary tillage are moldboard plows, disk plows and tillers, heavy disk harrows, chisel plows, subsoilers, rotary plows, listers, and bedders. Moldboard plows and heavy disk harrows are the most commonly used primary tillage tools. Implements used for secondary tillage are disk harrows, cultivators, spike and spring tooth harrows, and rotary hoes and cultivators. The most common implement used for secondary cultivation is the disk harrow. Generally, several tillage operations are performed before the field is ready for planting. In dry climates, culti-packers are also used as the final tillage operation before planting. Increasing the soil density in the top few centimeters helps retain soil moisture.

**Example 170.1.** What is the engine power requirement for a four-wheel drive tractor pulling a three-bottom 400 mm moldboard plow to a depth of 18 cm and at a speed of 6.0 km/h on medium textured soil?

**Solution.** Data from [Tables 170.2](#) and [170.3](#) are incorporated into Eqs. (170.3) and (170.6), respectively.

$$P_{db} = \frac{(7.4)(3 \cdot 400/1000)(6)}{3.6} = 14.8 \text{ kW}$$

$$P_T = 1.15 \frac{14.8}{0.75} = 22.7 \text{ kW}$$

## Crop Planting

Agricultural plants usually begin from either seeds or seedling transplants. Important factors affecting seed germination and emergence include seed viability, soil temperature, availability of moisture and air to the seeds, and soil strength and resistance to seedling emergence. The planter can exert a strong influence on the rate of germination and emergence of seeds through control of planting depth and firming of soil around the seeds or roots of seedlings. In addition, the planter must meter seeds at the proper rate and, in some cases, must control the horizontal down-the-row placement of seeds in a desired pattern.

Equipment is available for three seeding practices: broadcasting, drilling, and precision planting. Broadcasting refers to random placement of seeds on the soil surface. Seed is metered from a hopper through a variable orifice onto a spinning disk, which accelerates the seed and distributes it. Drilling is the random down-the-row or horizontal placement of seeds in furrows that are then covered. In a seed drill, seeds are metered from a series of hoppers, typically by fluted wheels, into small furrows that are subsequently covered and pressed. High-density plantings, high costs of hand thinning, and erratic performance of mechanical thinners have resulted in the development of precision seeding techniques. In precision planting, the seeds are planted in rows at uniform spacing. Precision planters are similar in operation to press drills; however, the metering hardware is more complex, allowing for seed placement at precise depths and locations.

[Table 170.5](#) provides data on typical seeding rates and practices for selected crops. The wide ranges of seed spacing reflect the dependence on climate, season, and type of market. For agronomic crops, the lower seeding rates and wider spacings are more typical for nonirrigated conditions. For vegetable crops, spacing is determined by type of market (fresh versus processor), desired fruit size, harvesting equipment dimensions, and time of year. Depth of seeding is determined by antecedent soil moisture, soil type, and soil temperature. These data can be used to establish general criteria for planting and cultural equipment.

**Example 170.2.** A 3.4 m wide grain drill is used to plant a 50 ha field in wheat. Approximately how many hours are required for this planting operation?

**Solution.** Data from [Table 170.1](#) are used in Eq. (170.1) for determining the field capacity, and



then the operation time is calculated.

$$C_f = \frac{(6.5)(3.4)(0.7)}{10} = 1.55 \text{ ha/h}$$

$$\text{Time} = \frac{50 \text{ ha}}{1.55 \text{ ha/h}} = 32 \text{ h or four 8 h days}$$

**Table 170.5** Traditional Seed Depth and Rate for Selected Crops

Crop	Depth to Plant Seed(cm)	Spacing between Plants in Row (cm)	Spacing between Rows(cm)	No. Seeds per Gram	Seeding Rate (kg/ha)
Alfalfa	0.6–1.3	Drilled	Drilled	480	9–22
Barley	1.3–2.5	Drilled	Drilled	28	81–108
Bean, snap	2.5–3.8	10–30	46–107	4–5	75–100
Broccoli	0.6	36–91	61–91	320	½–1½
Cabbage	0.6	36–91	61–91	320	½–1½
Carrot	0.6	3–8	38–61	820	2–4
Cauliflower	0.6	36–61	61–122	320	½–1½
Celery	0.6	15–30	46–91	2500	1–2
Corn, field	2.5–5.0	13–30	61–91	4–6	12–18
Corn, sweet	2.5–5.0	15–25	61–91	4–6	10–17
Cotton	2.5–5.0	3–8	61–97	2–4	18–25
Cucumber	2.5	30–91	91–122	40	3–6
Lettuce, head	0.6–1.3	15–25	46–91	900	1–3
Oat	1.3–2.5	Drilled	Drilled	28	54–143
Onion	1.3	8–10	38–91	305	3–5
Pea, English	5.0	3–5	46–91	3–6	100–250
Potato (tubers)	10.0	23–38	76–107	—	1000–2000
Rice	0.6–1.9	Drilled	Drilled	25	75–179
Sorghum	1.3–5.0	Drilled	Drilled	62	17–50
Soybean	2.5–3.8	3–15	30–91	6–12	20–50
Squash, bush	2.5	46–122	91–152	4–14	2–7
Sugar beet	0.7–1.5	10–15	60–91	120	½–1
Tomato	1.3	30–91	61–152	250–430	½–1
Watermelon	2.5	61–244	183–244	1–3	1–3
Wheat	1.5–2.5	Drilled	Drilled	26	67–101

Data from *Facts and Figures*[[Anonymous, 1990](#)]; Lorenz and Maynard [[1988](#)]; and Treadgill [[1988](#)].

## Crop Harvest

The final crop production operation is the harvesting of the plant parts that have economic value to the grower. In some cases, more than one plant part may have economic value. Many crops are highly perishable products that must be harvested within a very narrow time range, handled carefully, and either processed, properly stored, or consumed fresh soon after harvesting.

Mechanical harvesters are available for a number of crops. The general groups are hay and forage harvesters, grain harvesters, and fruit, nut, and vegetable harvesters.

Hay and forage harvesters are used in producing animal feed as ensilage or hay. Ensilage involves cutting the forage at 70 to 80% moisture, allowing it to field dry to 50 to 60% moisture, chopping it into short lengths to obtain adequate packing, and preserving it by fermentation in an airtight chamber. Equipment for the following steps are required: cut/condition, windrow, wilt, chop, transport, and store. For hay production, the forage is cut and allowed to dry to a moisture content of 15 to 23% before storage. Hay production requires equipment for the following steps: cut/condition/swath, rake into windrows, dry, bale or chop, transport, and store.

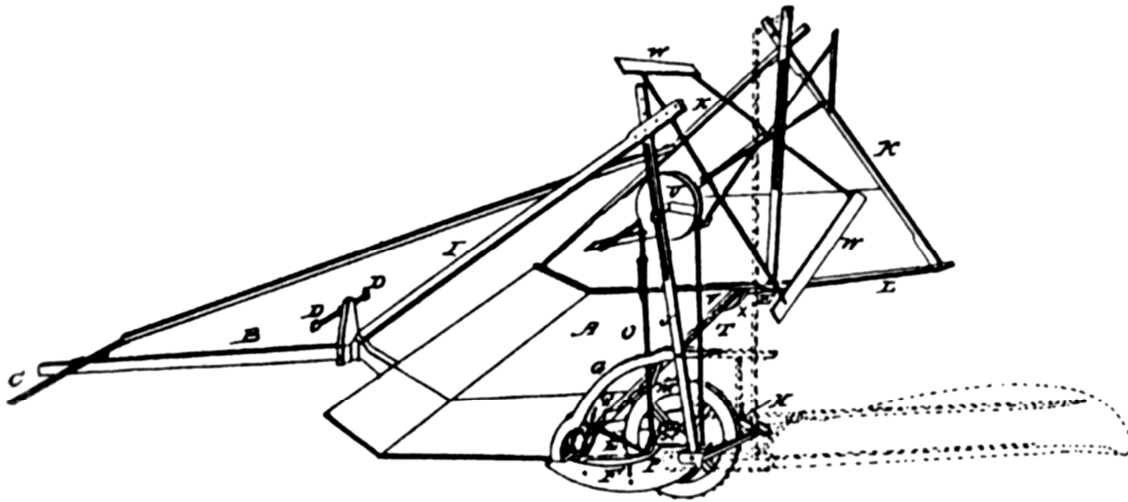
The harvesting of cereal grains, grasses, and legumes is accomplished almost entirely with the combine. A combine has five general mechanical functions: cutting, feeding, threshing, separating, and cleaning. The cutting operation is accomplished with a cutter bar and reel. The feeding mechanism distributes and delivers the crop material to the threshing cylinder in a steady uniform flow. The threshing operation is accomplished with a cylinder working against iron bar concaves. The separating mechanism extracts the straw. Cleaning of the grain, the final step, is accomplished with screening devices and blowers.

Harvesters for fruits, nuts, and vegetables are usually crop specific and may be equipped with special sensors for product selection according to maturity and size. Although these harvesters are designed for a broad range of specialty crops, they may be broadly classified according to the physical location in which the harvestable portion of the crop is located. These four crop zones are root (e.g., for sugar beets, potatoes), surface (e.g., for beans, tomatoes), bush and trellis (e.g., for boysenberries, grapes), and tree (e.g., for olives, almonds). Successful harvest mechanization requires a total systems approach, which includes varietal breeding, cultural practices, materials handling, grading and sorting, and ultimate processing. Harvester selection is therefore based on these factors, as well as such factors as economics and available labor.

**Example 170.3.** A tomato harvester is observed to travel at 4.0 km/h with a design width of 1.5 m. The average yield for the field is 80 t/ha. What is the material capacity of the machine?

**Solution.** Data from [Table 170.1](#) are used in Eq. 2 for determining the material handling capacity.

$$C_m = \frac{(4.0)(1.5)(80)(0.60)}{10} = 28.8 \text{ t/h}$$



## REAPER

*Cyrus H. McCormick Patented June 21, 1834*

Britain had seen several partially successful grain reapers invented in the 1820s. The McCormick reaper was first successfully used in the 1831 harvest on farms in his own Rockbridge County, Virginia. Cyrus's father, Robert, had built several partially successful reapers in his blacksmith shop. Father and son together built the reaper patented by Cyrus in 1834.

The traction wheel was directly behind the horse, and the grain was cut at the side, avoiding trampling. It had a vibrating cutter bar ("either smooth or with teeth"). Important, too, was a revolving reel of slats ("movable to any height required to suit the grain") to bend the grain toward the cutter and to move the cut grain to the side platform for gathering by a farmer following behind. Competition (by Hussey and others) was strong, but McCormick became very wealthy and was even elected to the French Academy in 1878 for his invention. (Courtesy of DewRay Products, Inc.)

## Defining Terms

**Draft:** The horizontal force required to propel an implement in the direction of travel.

**Drawbar power:** The power to pull or move an implement at a uniform speed. It is chiefly a function of implement draft and forward speed.

**Field capacity:** The amount of processing that a machine can accomplish per hour, expressed on either an area or a material basis.

**PTO power:** The power to operate an implement from the power-takeoff shaft. It is chiefly a function of torque and rotational speed.

## References

- Anonymous. 1976. *Fundamentals of Machine Operation<sup>3</sup>/<sub>4</sub>Tillage*. John Deere Service Publications, Moline, IL.
- Anonymous. 1990. *Facts and Figures*, 3rd ed. Doane Information Services, St. Louis, MO.
- Anonymous. 1993. *Standards 1993<sup>3</sup>/<sub>4</sub>Standards, Engineering Practices and Data*, 40th ed. ASAE Press, St. Joseph, MI.
- Hunt, D. R. 1973. *Farm Power and Machinery Management*, 7th ed. Iowa State University Press, Ames, IA.
- Kepner, R. A., Bainer, R., and Barger, E. L. 1978. *Principles of Farm Machinery*, 3rd ed. AVI, Westport, CT.
- Lorenz, O. A. and Maynard, D. N. 1988. *Knott's Handbook for Vegetable Growers*, 3rd ed. John Wiley & Sons, New York.
- Srivastava, A. K., Goering, C. E., and Rohrbach, R. P. 1993. *Engineering Principles of Agricultural Machines*. ASAE Press, St. Joseph, MI.
- Treadgill, E. D. 1988. Plant growth data. In *Handbook of Engineering in Agriculture*, ed. R. H. Brown, pp. 129–131. CRC Press, Boca Raton, FL.
- Zoz, F. M. 1987. *Predicting Tractor Field Performance* (updated). ASAE Paper No. 87.1623, ASAE, St. Joseph, MI.

## Further Information

- Applied Engineering in Agriculture*. Published bimonthly by ASAE—the Society for Engineering in Agricultural, Food, and Biological Systems.
- Brown, R. H. (Ed.) 1988. *Handbook of Engineering in Agriculture*. CRC Press, Boca Raton, FL.
- Culpin, C. 1992. *Farm Machinery*, 12th ed. Blackwell Scientific, London.
- Transactions of the ASAE*. Published bimonthly by ASAE—the Society for Engineering in Agricultural, Food, and Biological Systems.

Ramakumar, R. "System Reliability"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

- 171.1 Catastrophic Failure Models
- 171.2 The Bathtub Curve
- 171.3 Mean Time to Failure
- 171.4 Average Failure Rate
- 171.5 A Posteriori Failure Probability
- 171.6 Units for Failure Rates
- 171.7 Application of the Binomial Distribution
- 171.8 Application of the Poisson Distribution
- 171.9 The Exponential Distribution
- 171.10 The Weibull Distribution
- 171.11 Combinatorial Aspects
- 171.12 Modeling Maintenance
- 171.13 Markov Models
- 171.14 Binary Model for a Repairable Component
- 171.15 Two Dissimilar Repairable Components
- 171.16 Two Identical Repairable Components
- 171.17 Frequency and Duration Techniques
- 171.18 Applications of Markov Process
- 171.19 Some Useful Approximations

**Rama Ramakumar**

*Oklahoma State University*

Application of system reliability evaluation techniques is gaining importance because of its effectiveness in the detection, prevention, and correction of failures in the design, manufacturing, and operational phases of a product. Increasing emphasis on the reliability and quality of products and systems, coupled with pressures to minimize cost, further emphasize the need to study and quantify reliability and arrive at innovative designs.

Reliability engineering has grown significantly during the past five decades (since World War II) to encompass many subareas, such as reliability analysis, failure theory and modeling, reliability allocation and optimization, reliability growth and modeling, reliability testing (including accelerated testing), data analysis and plotting, quality control and acceptance sampling, maintenance engineering, software reliability, system safety analysis, Bayesian analysis, reliability management, simulation, Monte Carlo techniques, and economic aspects of reliability.

The objectives of this chapter are to introduce the reader to the fundamentals and applications of classical reliability concepts and bring out the importance and benefits of reliability considerations.

## 171.1 Catastrophic Failure Models

Catastrophic failure refers to the case in which repair of the component is not possible, not available, or of no value to the successful completion of the mission originally planned. Modeling such failures is typically based on life test results. We can consider the "lifetime" or "time to failure"  $T$  as a continuous random variable. Then,

$$P(\text{survival up to time } t) = P(T > t) \equiv R(t) \quad (171.1)$$

where  $R(t)$  is the *reliability* function. Obviously, as  $t \rightarrow \infty$ ,  $R(t) \rightarrow 0$  because the probability of failure increases with time of operation. Moreover,

$$P(\text{failure at } t) = P(T \leq t) \equiv Q(t) \quad (171.2)$$

where  $Q(t)$  is the unreliability function. From the definition of the distribution function of a continuous random variable, it is clear that  $Q(t)$  is indeed the distribution function for  $T$ . Therefore, the failure density function  $f(t)$  can be obtained as

$$f(t) = \frac{d}{dt}Q(t) \quad (171.3)$$

The **hazard rate function**  $\lambda(t)$  is defined as

$$\lambda(t) \equiv \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \left[ \text{probability of failure in } (t, t + \Delta t], \right. \\ \left. \text{given survival up to } t \right] \quad (171.4)$$

It can be shown that

$$\lambda(t) = \frac{f(t)}{R(t)} \quad (171.5)$$

The four functions  $f(t)$ ,  $Q(t)$ ,  $R(t)$ , and  $\lambda(t)$  constitute the set of functions used in basic reliability analysis. The relationships between these functions are given in [Table 171.1](#).

**Table 171.1** Relationships between Different Reliability Functions

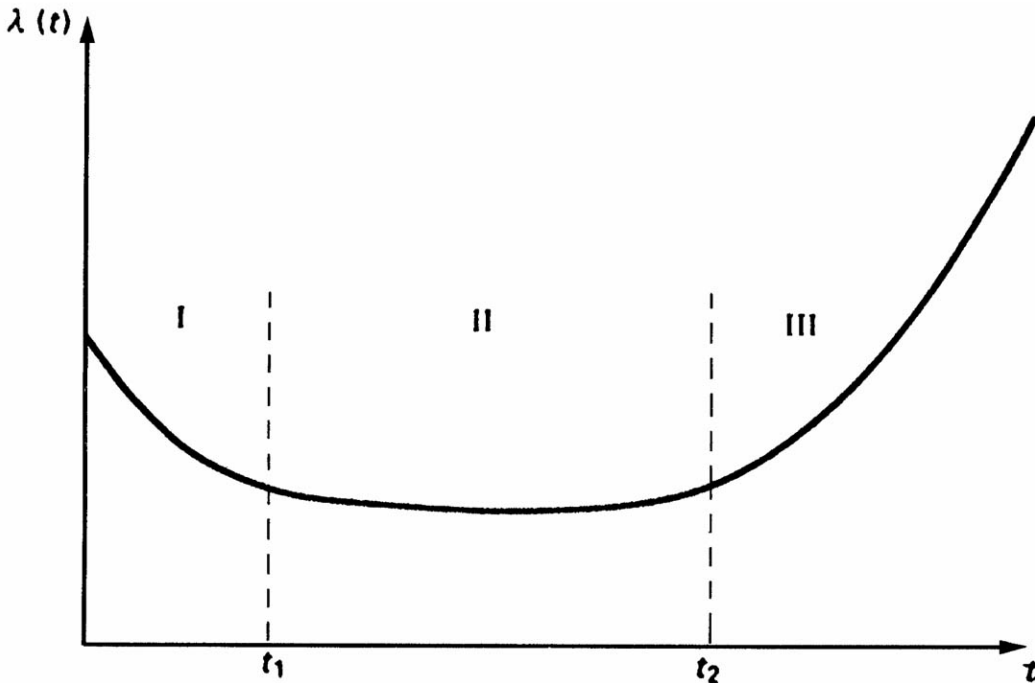
	$f(t)$	$\lambda(t)$	$Q(t)$	$R(t)$
$f(t) =$	$f(t)$	$\lambda(t) \exp \left[ - \int_0^t \lambda(\xi) d\xi \right]$	$\frac{d}{dt}Q(t)$	$-\frac{d}{dt}R(t)$
$(t) =$	$\frac{f(t)}{1 - \int_0^t f(\xi) d\xi}$	$\lambda(t)$	$\frac{1}{1 - Q(t)} \frac{d}{dt}(Q(t))$	$-\frac{d}{dt}[\ln R(t)]$
$Q(t) =$	$\int_0^t f(\xi) d\xi$	$1 - \exp \left[ - \int_0^t \lambda(\xi) d\xi \right]$	$Q(t)$	$1 - R(t)$
$R(t) =$	$1 - \int_0^t f(\xi) d\xi$	$\exp \left[ - \int_0^t \lambda(\xi) d\xi \right]$	$1 - Q(t)$	$R(t)$

Source: Ramakumar, R. 1993. *Engineering Reliability: Fundamentals and Applications*. Prentice Hall, Englewood Cliffs, NJ. With permission.

## 171.2 The Bathtub Curve

Of the four functions discussed, the hazard rate function  $\lambda(t)$  displays the different stages during the lifetime of a component most clearly. In fact, typical  $\lambda(t)$  plots have the general shape of a **bathtub curve** as shown in Fig. 171.1. The first region corresponds to *wear-in* (infant mortality) or early failures during debugging. The hazard rate goes down as debugging continues. The second region corresponds to an essentially constant and low failure rate—failures can be considered to be nearly random. This is the useful lifetime of the component. The third region corresponds to the *wear-out* or fatigue phase with a sharply increased hazard rate.

**Figure 171.1** Bathtub-shaped hazard function (Source: Ramakumar, R. 1993. *Engineering Reliability: Fundamentals and Applications*. Prentice Hall, Englewood Cliffs, NJ. With permission.)



*Burn-in* refers to the practice of subjecting components to an initial operating period of  $t_1$  (see Fig. 171.1) before delivering them to the customer. This eliminates all the initial failures from occurring after delivery to customers requiring high-reliability components. Moreover, it is prudent to replace a component as it approaches the wear-out region (i.e., after an operating period of  $(t_2 - t_1)$ ). Electronic components tend to have a long useful life (constant hazard) period. The wear-out region tends to dominate in the case of mechanical components.

## 171.3 Mean Time to Failure

The mean or expected value of the continuous random variable *time to failure* is the **mean time to failure** (MTTF). This is a very useful parameter which is often used to assess the suitability of components. It can be obtained using either the failure density function  $f(t)$  or the reliability function  $R(t)$  as follows:

$$\text{MTTF} = \int_0^{\infty} t f(t) dt \quad \text{or} \quad \int_0^{\infty} R(t) dt \quad (171.6)$$



In the case of repairable components, the repair time can also be considered as a continuous random variable with an expected value of MTTR. The mean time between failures, MTBF, is the sum of MTTF and MTTR. For well-designed components,  $\text{MTTR} \ll \text{MTTF}$ . Thus, MTBF and MTTF are often used interchangeably.

## 171.4 Average Failure Rate

---

The average failure rate over the time interval 0 to  $T$  is defined as

$$\text{AFR}(0, T) \equiv \text{AFR}(T) = -\frac{\ln R(T)}{T} \quad (171.7)$$

## 171.5 A Posteriori Failure Probability

---

When components are subjected to a burn-in (or wear-in) period of duration  $T$ , and if the component survives during  $(0, T)$ , the probability of failure during  $(T, T + t)$  is called the *a posteriori failure probability*  $Q_c(t)$ . It can be found using

$$Q_c(t) = \frac{\int_T^{T+t} f(\xi) d\xi}{\int_T^{\infty} f(\xi) d\xi} \quad (171.8)$$

The probability of survival during  $(T, T + t)$  is

$$\begin{aligned} R(t|T) &= 1 - Q_c(t) = \frac{\int_{T+t}^{\infty} f(\xi) d\xi}{\int_T^{\infty} f(\xi) d\xi} \\ &= \frac{R(T+t)}{R(T)} = \exp \left[ -\int_T^{T+t} \lambda(\xi) d\xi \right] \end{aligned} \quad (171.9)$$

## 171.6 Units for Failure Rates

---

Several units are used to express failure rates. In addition to  $\lambda(t)$ , which is usually in number per hour,  $\%/K$  is used to denote failure rates in percent per thousand hours, and  $PPM/K$  is used to express failure rate in parts per million per thousand hours. The last unit is also known as FIT for "fails in time." The relationships between these units are given in [Table 171.2](#).

**Table 171.2** Relationships between Different Failure Rate Units

	$I$ (#/hr)	%/K	PPM/K (FIT)
$\lambda =$	$I$	$10^{-5}$ (%/K)	$10^{-9}$ (PPM/K)
%/K =	$10^5 I$	%/K	$10^{-4}$ (PPM/K)
PPM/K (FIT) =	$10^9 I$	$10^4$ (%/K)	PPM/K

Source: Ramakumar, R. 1993. *Engineering Reliability: Fundamentals and Applications*. Prentice Hall, Englewood Cliffs, NJ. With permission.

## 171.7 Application of the Binomial Distribution

In an experiment consisting of  $n$  identical independent trials, with each trial resulting in success or failure with probabilities of  $p$  and  $q$ , the probability  $P_r$  of  $r$  successes and  $(n - r)$  failures is

$$P_r = {}_n C_r p^r (1 - p)^{n-r} \quad (171.10)$$

If  $X$  denotes the number of successes in  $n$  trials, then it is a discrete random variable with a mean value of  $(np)$  and variance of  $(npq)$ .

In a system consisting of a collection of  $n$  identical components with a probability  $p$  that a component is defective, the probability of finding  $r$  defects out of  $n$  is given by the  $P_r$  in Eq. (171.10). If  $p$  is the probability of success of one component and if at least  $r$  of them must be good for system success, then the system reliability (probability of system success) is given by

$$R = \sum_{k=r}^n {}_n C_k p^k (1 - p)^{n-k} \quad (171.11)$$

For systems with redundancy,  $r < n$ .

## 171.8 Application of the Poisson Distribution

For events that occur *in time* at an average rate of  $\lambda$  occurrences per unit of time, the probability  $P_x(t)$  of exactly  $x$  occurrences during the time interval  $(0, t)$  is given by

$$P_x(t) = \frac{(\lambda t)^x e^{-\lambda t}}{x!} \quad (171.12)$$

The number of occurrences  $X$  in  $(0, t)$  is a discrete random variable with a mean value of  $\mu$  of  $(\lambda t)$ , and a standard deviation  $\sigma$  of  $\sqrt{\lambda t}$ . By setting  $X = 0$  in Eq. (171.12), we obtain the probability of no occurrence in  $(0, t)$  as  $e^{-\lambda t}$ . If the event is failure, then no occurrence means success and  $e^{-\lambda t}$  is the probability of success or system reliability. This is the well-known and often used exponential distribution, also known as the constant hazard model.

## 171.9 The Exponential Distribution

---

A constant hazard rate (constant  $\lambda$ ) corresponding to the useful lifetime of components leads to the single parameter exponential distribution. The functions of interest associated with a constant  $\lambda$  are

$$f(t) = \lambda e^{-\lambda t}, t > 0 \quad (171.13)$$

$$R(t) = e^{-\lambda t} \quad (171.14)$$

$$Q(t) = Q_c(t) = 1 - e^{-\lambda t} \quad (171.15)$$

The a posteriori failure probability  $Q_c(t)$  is independent of the prior operating time  $T$ , indicating that the component does not degrade no matter how long it operates. Obviously, such a scenario is valid only during the useful lifetime (horizontal portion of the bathtub curve) of the component.

The mean and standard deviation of the random variable *lifetime* are

$$\mu \equiv \text{MTTF} = \frac{1}{\lambda} \quad \text{and} \quad \sigma = \frac{1}{\lambda} \quad (171.16)$$

## 171.10 The Weibull Distribution

---

The Weibull distribution has two parameters—a scale parameter  $\alpha$  and a shape parameter  $\beta$ . By adjusting these two parameters, a wide range of experimental data can be modeled in system reliability studies. The associated functions are

$$\lambda(t) = \frac{\beta t^{\beta-1}}{\alpha^\beta}; \quad \alpha > 0, \beta > 0, t \geq 0 \quad (171.17)$$

$$f(t) = \frac{\beta t^{\beta-1}}{\alpha^\beta} \exp \left[ - \left( \frac{t}{\alpha} \right)^\beta \right] \quad (171.18)$$

$$R(t) = \exp \left[ - \left( \frac{t}{\alpha} \right)^\beta \right] \quad (171.19)$$

With  $\beta = 1$ , the Weibull distribution reduces to the constant hazard model with  $\lambda = (1/\alpha)$ . With  $\beta = 2$ , the Weibull distribution reduces to the Rayleigh distribution.

The associated MTTF is

$$\text{MTTF} = \mu = \alpha \Gamma \left( 1 + \frac{1}{\beta} \right) \quad (171.20)$$

where  $\Gamma$  denotes the gamma function.

## 171.11 Combinatorial Aspects

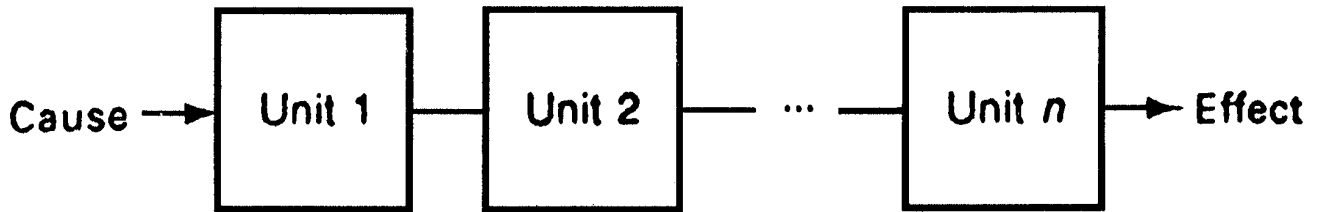
Analysis of complex systems is facilitated by decomposition into functional entities consisting of subsystems or units and by the application of combinatorial considerations and network modeling techniques.

A **series structure** (or chain structure) consisting of  $n$  units is shown in Fig. 171.2. From the reliability point of view, the system will succeed only if all the units succeed. The units may or may not be physically in series. If  $R_i$  is the probability of success of the  $i$ th unit, then the series system reliability  $R_s$  is given as

$$R_s = \prod_{i=1}^n R_i \quad (171.21)$$

if the units do not interact with each other. If they do, then the conditional probabilities must be carefully evaluated.

**Figure 171.2** Series or chain structure. (Source: Ramakumar, R. 1993. *Engineering Reliability: Fundamentals and Applications*. Prentice Hall, Englewood Cliffs, NJ. With permission.)



If each of the units has a constant hazard, then

$$R_s(t) = \prod_{i=1}^n \exp(-\lambda_i t) \quad (171.22)$$

where  $\lambda_i$  is the constant failure rate for the  $i$ th unit or component. This enables us to replace the  $n$  components in series by an equivalent component with a constant hazard  $\lambda_s$  where

$$\lambda_s = \sum_{i=1}^n \lambda_i \quad (171.23)$$

If the components are identical, then  $\lambda_s = n\lambda$  and the MTTF for the equivalent component is  $(1/n)$  of the MTTF of one component.

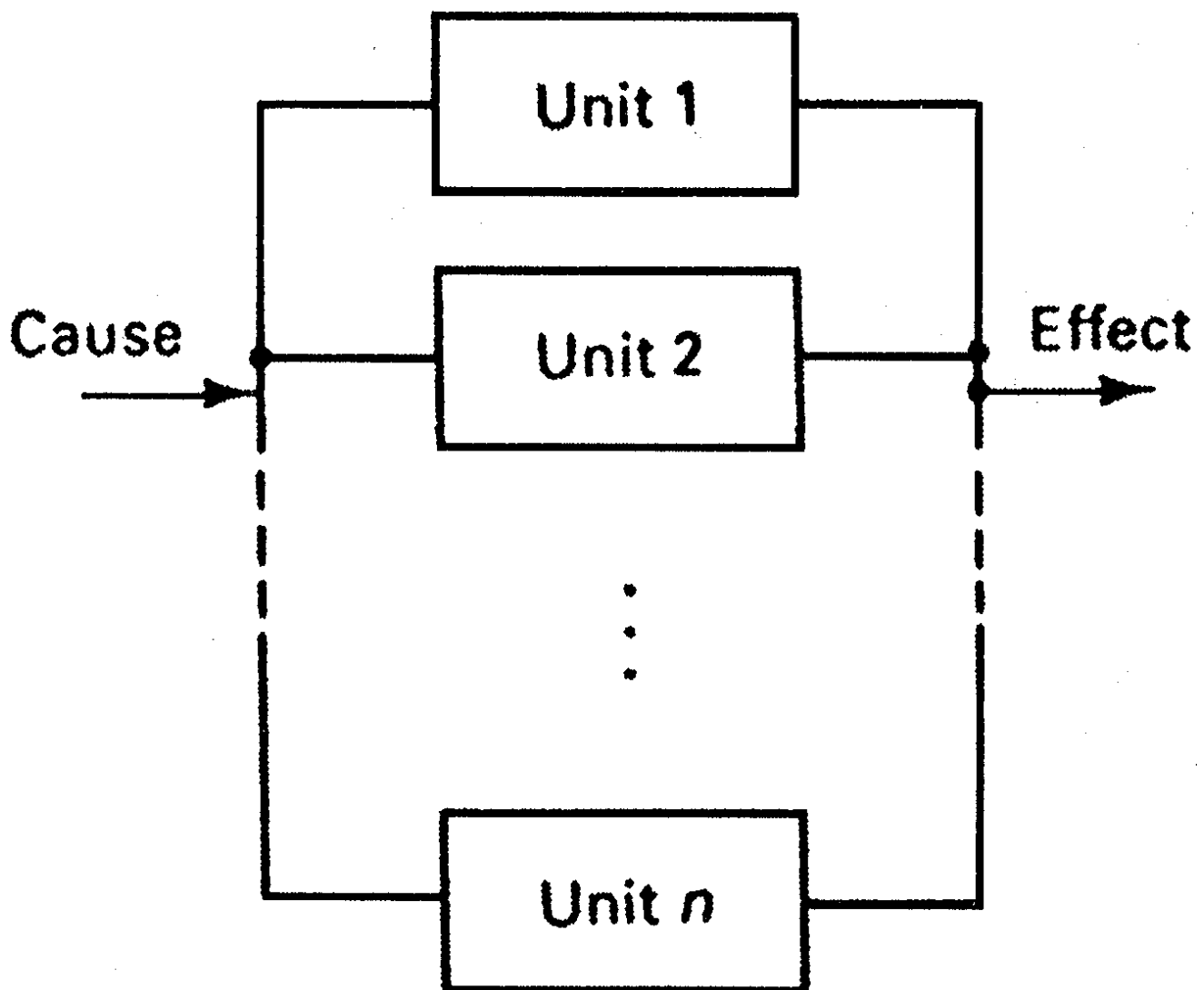
A **parallel structure** consisting of  $n$  units is shown in Fig. 171.3. From the reliability point of view, the system will succeed if any one of the  $n$  units succeeds. Once again, the units may or may

not be physically or topologically in parallel. If  $Q_i$  is the probability of failure of the  $i$ th unit, then the parallel system reliability  $R_p$  is given as

$$R_p = 1 - \prod_{i=1}^n Q_i \quad (171.24)$$

if the units do not interact with each other (i.e., are independent).

**Figure 171.3** Parallel structure. (Source: Ramakumar, R. 1993. *Engineering Reliability: Fundamentals and Applications*. Prentice Hall, Englewood Cliffs, NJ. With permission.)



If each of the units has a constant hazard, then

$$R_p(t) = 1 - \prod_{i=1}^n [1 - \exp(-\lambda_i t)] \quad (171.25)$$

and we do not have the luxury of being able to replace the parallel system by an equivalent component with a constant hazard. The parallel system does not exhibit constant hazard even though each of the units has constant hazard.

The MTTF of the parallel system can be obtained by using Eq. (171.25) in Eq. (171.6). The results for the case of components with identical hazards  $\lambda$  are:  $(1.5/\lambda)$ ,  $(1.833/\lambda)$ , and  $(2.083/\lambda)$  for  $n = 2, 3$ , and  $4$ , respectively. The largest gain in MTTF is obtained by going from one component to two components in parallel. It is uncommon to have more than two or three components in a truly parallel configuration because of the cost involved. For two nonidentical components in parallel with hazard rates  $\lambda_1$  and  $\lambda_2$ , the MTTF is given as

$$\text{MTTF} = \frac{1}{\lambda_1} + \frac{1}{\lambda_2} - \frac{1}{\lambda_1 + \lambda_2} \quad (171.26)$$

An  $r$ -out-of- $n$  structure, also known as a partially redundant system, can be evaluated using Eq. (171.11). If all the components are identical, independent, and have a constant hazard  $\lambda$ , then the system reliability can be expressed as

$$R(t) = \sum_{k=r}^n {}_n C_k e^{-k\lambda t} (1 - e^{-\lambda t})^{n-k} \quad (171.27)$$

For  $r = 1$ , the structure becomes a parallel system. For  $r = n$ , it becomes a series system.

Series-parallel systems are evaluated by repeated application of the expressions derived for series and parallel configurations by employing the well-known network reduction techniques.

Several general techniques are available for evaluating the reliability of complex structures that do not come under purely series or parallel or series-parallel. They range from inspection to cutset and tieset methods and connection matrix techniques that are amenable to computer programming.

## 171.12 Modeling Maintenance

---

Maintenance of a component could be a scheduled (preventative) one or a forced (corrective) one. The latter follows in-service failures and can be handled using Markov models discussed later. Scheduled maintenance is conducted at fixed intervals of time, irrespective of the system continuing to operate satisfactorily.

Scheduled maintenance, under ideal conditions, takes very little time (compared to the time between maintenances) and the component is restored to an "as new" condition. Even if the component is irreparable, scheduled maintenance postpones failure and prolongs the life of the component. Scheduled maintenance makes sense only for those components with increasing

hazard rates. Most mechanical systems come under this category. It can be shown that the density function  $f_T^*(t)$ , with scheduled maintenance included, can be expressed as

$$f_T^*(t) = \sum_{k=0}^{\infty} f_1(t - kT_M) R^k(T_M) \quad (171.28)$$

where

$$f_1(t) = \begin{cases} f_T(t) & \text{for } 0 < t \leq T_M \\ 0 & \text{otherwise} \end{cases} \quad (171.29)$$

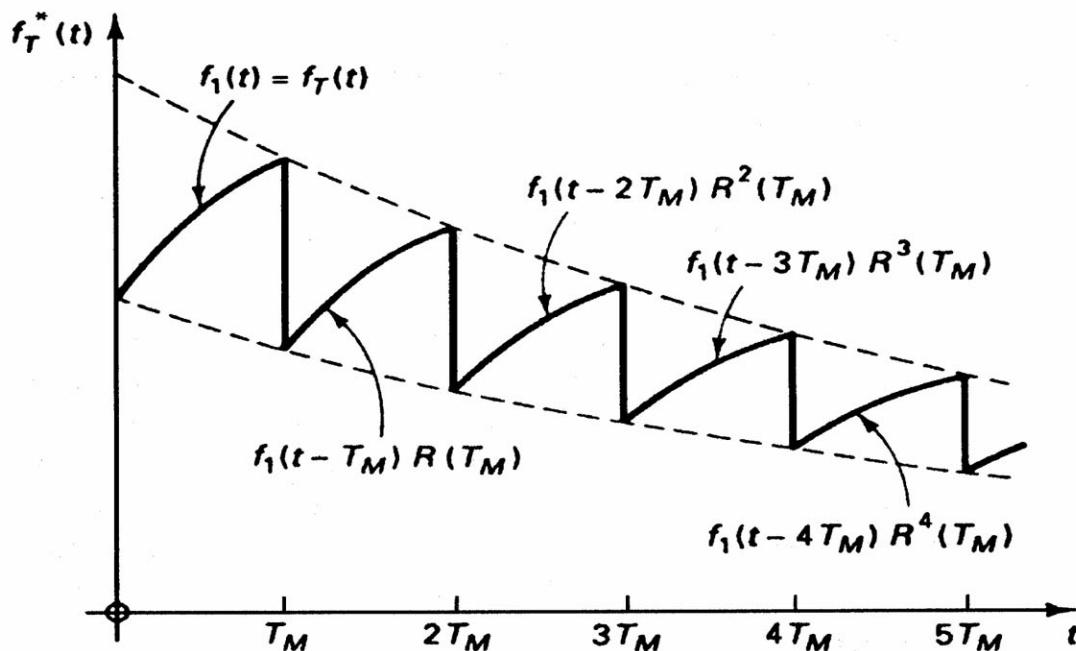
and

$R(t)$  = component reliability function  
 $T_M$  = time between maintenance, constant  
 $f_T(t)$  = original failure density function

In Eq. (171.28),  $k = 0$  is used only between  $t = 0$  and  $t = T_M$ , and  $k = 1$  is used only between  $t = T_M$  and  $t = 2T_M$  and so on.

A typical  $f_T^*(t)$  is shown in Fig. 171.4. The time scale is divided into equal intervals of  $T_M$  each. The function in each segment is a scaled-down version of the one in the previous segment, the scaling factor being equal to  $R(T_M)$ . Irrespective of the nature of the original failure density function, scheduled maintenance gives it an exponential tendency. This is another justification for the widespread use of exponential distribution in system reliability evaluations.

**Figure 171.4** Density function with ideal scheduled maintenance incorporated. (Source: Ramakumar, R. 1993. *Engineering Reliability: Fundamentals and Applications*. Prentice Hall, Englewood Cliffs, NJ. With permission.)



## 171.13 Markov Models

---

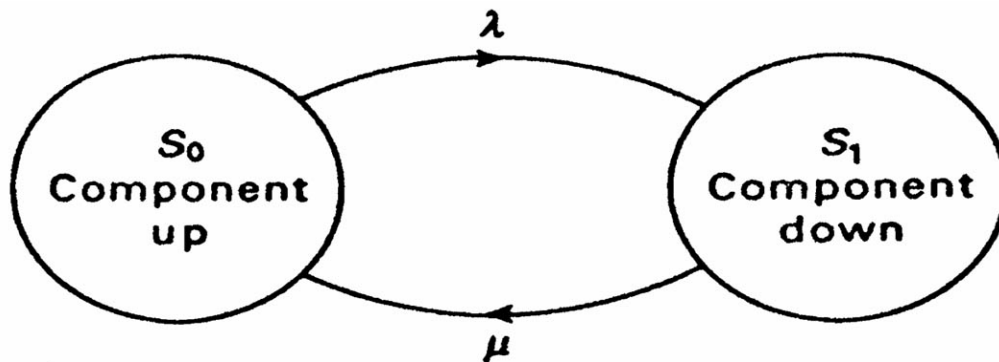
Of the different Markov models available, the discrete-state, continuous-time Markov process has found many applications in system reliability evaluation, including the modeling of repairable systems. The model consists of a set of discrete states (called the state space) in which the system can reside and a set of transition rates between appropriate states. Using these, a set of first-order differential equations is derived in the standard vector-matrix form for the time-dependent probabilities of the various states. Solution of these equations incorporating proper initial conditions gives the probabilities of the system residing in different states as functions of time. Several useful results can be gleaned from these functions.

## 171.14 Binary Model for a Repairable Component

---

The binary model for a repairable component assumes that the component can exist in one of two states—the *up* state or the *down* state. The transition rates between these two states,  $S_0$  and  $S_1$ , are assumed to be constant and equal to  $\lambda$  and  $\mu$ . These transition rates are the constant failure and repair rates implied in the modeling process and their reciprocals are the MTTF and MTTR, respectively. Figure 171.5 illustrates the binary model.

**Figure 171.5** State space diagram for a single repairable component. (Source: Ramakumar, R. 1993. *Engineering Reliability: Fundamentals and Applications*. Prentice Hall, Englewood Cliffs, NJ. With permission.)





The associated Markov differential equations are

$$\begin{bmatrix} P_0'(t) \\ P_1'(t) \end{bmatrix} = \begin{bmatrix} -\lambda & \mu \\ \lambda & -\mu \end{bmatrix} \begin{bmatrix} P_0(t) \\ P_1(t) \end{bmatrix} \quad (171.30)$$

with the initial conditions

$$\begin{bmatrix} P_0(0) \\ P_1(0) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad (171.31)$$

The coefficient matrix of Markov differential equations, namely

$$\begin{bmatrix} -\lambda & \mu \\ \lambda & -\mu \end{bmatrix},$$

is obtained by transposing the matrix of rates of departures

$$\begin{bmatrix} 0 & \lambda \\ \mu & 0 \end{bmatrix}$$

and replacing the diagonal entries by the negative of the sum of all the other entries in their respective columns. Solution of Eq. (171.30) with initial conditions as given by Eq. (171.31) yields

$$P_0(t) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} e^{-(\lambda + \mu)t} \quad (171.32)$$

$$P_1(t) = \frac{\lambda}{\lambda + \mu} \left[ 1 - e^{-(\lambda + \mu)t} \right] \quad (171.33)$$

The limiting, or steady state, probabilities are found by letting  $t \rightarrow \infty$ . They are also known as limiting **availability**  $A$  and limiting unavailability  $U$  and they are

$$P_0 = \frac{\mu}{\lambda + \mu} \equiv A \quad \text{and} \quad P_1 = \frac{\lambda}{\lambda + \mu} \equiv U \quad (171.34)$$

The time-dependent  $A(t)$  and  $U(t)$  are simply  $P_0(t)$  and  $P_1(t)$ , respectively.

Referring back to Eq. (171.14) for a constant hazard component and comparing it with Eq. (171.32) which incorporates repair, the difference between  $R(t)$  and  $A(t)$  becomes obvious. Availability  $A(t)$  is the probability that the component is up at time  $t$ , and reliability  $R(t)$  is the probability that the system has continuously operated from 0 to  $t$ . Thus,  $R(t)$  is much more stringent than  $A(t)$ . While both  $R(0)$  and  $A(0)$  are unity,  $R(t)$  drops off rapidly as compared to  $A(t)$  as time progresses. With a small value of MTTR (or large value of  $\mu$ ), it is possible to realize a very high availability for a repairable component.

## 171.15 Two Dissimilar Repairable Components

Irrespective of whether the two components are in series or in parallel, the state space consists of four possible states:  $S_1$  (1 up, 2 up),  $S_2$  (1 down, 2 up),  $S_3$  (1 up, 2 down), and  $S_4$  (1 down, 2 down). The actual system configuration will determine which of these four states correspond to system success and failure. The associated state space diagram is shown in Fig. 171.6. Analysis of this system results in the following steady state probabilities:

$$P_1 = \frac{\mu_1 \mu_2}{\text{Denom}}; \quad P_2 = \frac{\lambda_1 \mu_2}{\text{Denom}}; \quad P_3 = \frac{\lambda_2 \mu_1}{\text{Denom}}; \quad P_4 = \frac{\lambda_1 \lambda_2}{\text{Denom}} \quad (171.35)$$

where

$$\text{Denom} \equiv (\lambda_1 + \mu_1)(\lambda_2 + \mu_2) \quad (171.36)$$

For components in series,  $A = P_1$ ,  $U = (P_2 + P_3 + P_4)$ , and the two components can be replaced by an equivalent component with a failure rate of  $\lambda_s = (\lambda_1 + \lambda_2)$  and a mean repair duration of  $r_s$ , where

$$r_s \cong \frac{\lambda_1 r_1 + \lambda_2 r_2}{\lambda_s} \quad (171.37)$$

Extending this to  $n$  components in series, the equivalent system will have

$$\lambda_s = \sum_{i=1}^n \lambda_i \quad \text{and} \quad r_s \cong \frac{1}{\lambda_s} \sum_{i=1}^n \lambda_i r_i \quad (171.38)$$

$$\text{and system unavailability} = U_s \cong \lambda_s r_s = \sum_{i=1}^n \lambda_i r_i \quad (171.39)$$

For components in parallel,  $A = (P_1 + P_2 + P_3)$ ,  $U = P_4$ , and the two components can be replaced by an equivalent component with

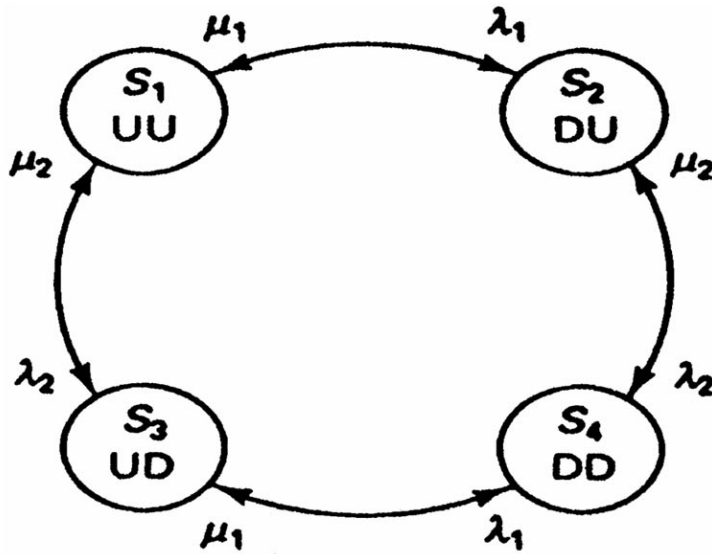
$$\lambda_p \cong \lambda_1(\lambda_2 r_1) + \lambda_2(\lambda_1 r_2) \quad \text{and} \quad \mu_p = \mu_1 + \mu_2 \quad (171.40)$$

$$\text{and system unavailability} = U_p = \lambda_p(1/\mu_p) \quad (171.41)$$

Extension to more than two components in parallel follows similar lines. For three components in parallel,

$$\mu_p = (\mu_1 + \mu_2 + \mu_3) \quad \text{and} \quad U_p = \lambda_1 \lambda_2 \lambda_3 r_1 r_2 r_3 \quad (171.42)$$

**Figure 171.6** State space diagram for two dissimilar reparable components. (Source: Ramakumar, R. 1993. *Engineering Reliability: Fundamentals and Applications*. Prentice Hall, Englewood Cliffs, NJ. With permission.)

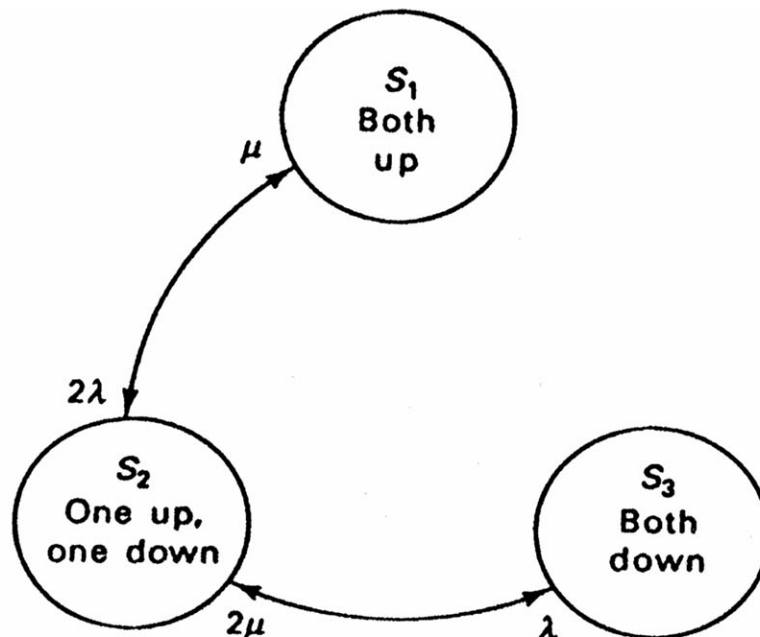


### 171.16 Two Identical Reparable Components

In this case, only three states are needed to complete the state space:  $S_1$  (both up),  $S_2$  (one up and one down), and  $S_3$  (both down). The corresponding state space diagram is shown in Fig. 171.7. Analysis of this system results in the following steady state probabilities:

$$P_1 = \left( \frac{\mu}{\lambda + \mu} \right)^2 ; \quad P_2 = \frac{2\lambda}{\mu} \left( \frac{\mu}{\lambda + \mu} \right)^2 ; \quad P_3 = \left( \frac{\lambda}{\lambda + \mu} \right)^2 \quad (171.43)$$

**Figure 171.7** State space diagram for two identical reparable components. (Source: Ramakumar, R. 1993. *Engineering Reliability: Fundamentals and Applications*. Prentice Hall, Englewood Cliffs, NJ. With permission.)



## 171.17 Frequency and Duration Techniques

---

The expected residence time in a state is the mean value of the passage time from the state in question to any other state. Cycle time is the time required to complete an *in* and *not in* cycle for that state. Frequency of occurrence (or encounter) for a state is the reciprocal of its cycle time. It can be shown that the frequency of occurrence of a state is equal to the steady state probability of being in that state multiplied by the total rate of departure from it. Also, the expected value of the residence time is equal to the reciprocal of the total rate of departure from that state.

Under steady state conditions, the expected frequency of entering a state must be equal to the expected frequency of leaving that state (this assumes that the system is *ergodic*, which will not be elaborated for lack of space). Using this principle, frequency balance equations can be easily written (one for each state) and solved in conjunction with the fact that the sum of the steady state probabilities of all the states must be equal to unity to obtain the steady state probabilities. This procedure is much simpler than solving the Markov differential equations and letting  $t \rightarrow \infty$ .

## 171.18 Applications of Markov Process

---

Once the different states are identified and a state space diagram is developed, Markov analysis can proceed systematically (probably with the help of a computer in the case of large systems) to yield a wealth of results used in system reliability evaluation. Inclusion of installation time after repair, maintenance, spare, and stand-by systems, and limitations imposed by restricted repair facilities are some of the many problems that can be studied.

## 171.19 Some Useful Approximations

---

For an  $r$ -out-of- $n$  structure with failure and repair rates of  $\lambda$  and  $\mu$  for each, the equivalent MTTR and MTTF can be approximated as

$$\text{MTTR}_{\text{eq}} = \frac{\text{MTTR of one component}}{n - r + 1} \quad (171.44)$$

$$\text{MTTF}_{\text{eq}} = \left( \begin{matrix} \text{MTTF} \\ \text{of one component} \end{matrix} \right) \left( \frac{\text{MTTF}}{\text{MTTR}} \right)^{n-r} \left[ \frac{(n-r)!(r-1)!}{n!} \right] \quad (171.45)$$

The influence of weather must be considered for components operating in an outdoor environment. If  $\lambda$  and  $\lambda'$  are the normal weather and stormy weather failure rates,  $\lambda'$  will be much greater than  $\lambda$ , and the average failure rate  $\lambda_f$  can be approximated as

$$\lambda_f \cong \left( \frac{N}{N+S} \right) \lambda + \left( \frac{S}{N+S} \right) \lambda' \quad (171.46)$$

where  $N$  and  $S$  are the expected durations of normal and stormy weather. For well-designed, high-reliability components, the failure rate  $\lambda$  will be very small and  $\lambda t \ll 1$ . Then, for a single component,

$$R(t) \cong 1 - \lambda t \quad \text{and} \quad Q(t) \cong \lambda t \quad (171.47)$$

and for  $n$  dissimilar components in series,

$$R(t) \cong 1 - \sum_{i=1}^n \lambda_i t \quad \text{and} \quad Q(t) \cong \sum_{i=1}^n \lambda_i t \quad (171.48)$$

For the case of  $n$  identical components in parallel,

$$R(t) \cong 1 - (\lambda t)^n \quad \text{and} \quad Q(t) \cong (\lambda t)^n \quad (171.49)$$

For the case of an  $r$ -out-of- $n$  configuration,

$$Q(t) \cong {}_nC_{n-r+1} (\lambda t)^{n-r+1} \quad (171.50)$$

Equations (171.47)–(171.50) are called rare-event approximations.

## Defining Terms

**Availability:** The availability  $A(t)$  is the probability that a system is performing its required function successfully at time  $t$ . The steady state availability  $A$  is the fraction of time that an item, system, or component is able to perform its specified or required function.

**Bathtub curve:** For most physical components and living entities, the plot of failure (or hazard) rate versus time has the shape of the longitudinal cross section of a bathtub.

**Hazard rate function:** The plot of instantaneous failure rate versus time is called the hazard function. It clearly and distinctly exhibits the different life cycles of the component.

**Mean time to failure:** The mean time to failure (MTTF) is the mean or expected value of time to failure.

**Parallel structure:** Also known as a completely redundant system, it describes a system that can succeed when at least one of two or more components succeeds.

**Redundancy:** Refers to the existence of more than one means, identical or otherwise, for accomplishing a task or mission.

**Reliability:** The reliability  $R(t)$  of an item or system is the probability that it has performed successfully over the time interval from 0 to  $t$ . In the case of irreparable systems,  $R(t) = A(t)$ . With repair,  $R(t) \leq A(t)$ .

**Series structure:** Also known as a chain structure or nonredundant system, it describes a system whose success depends on the success of all of its components.

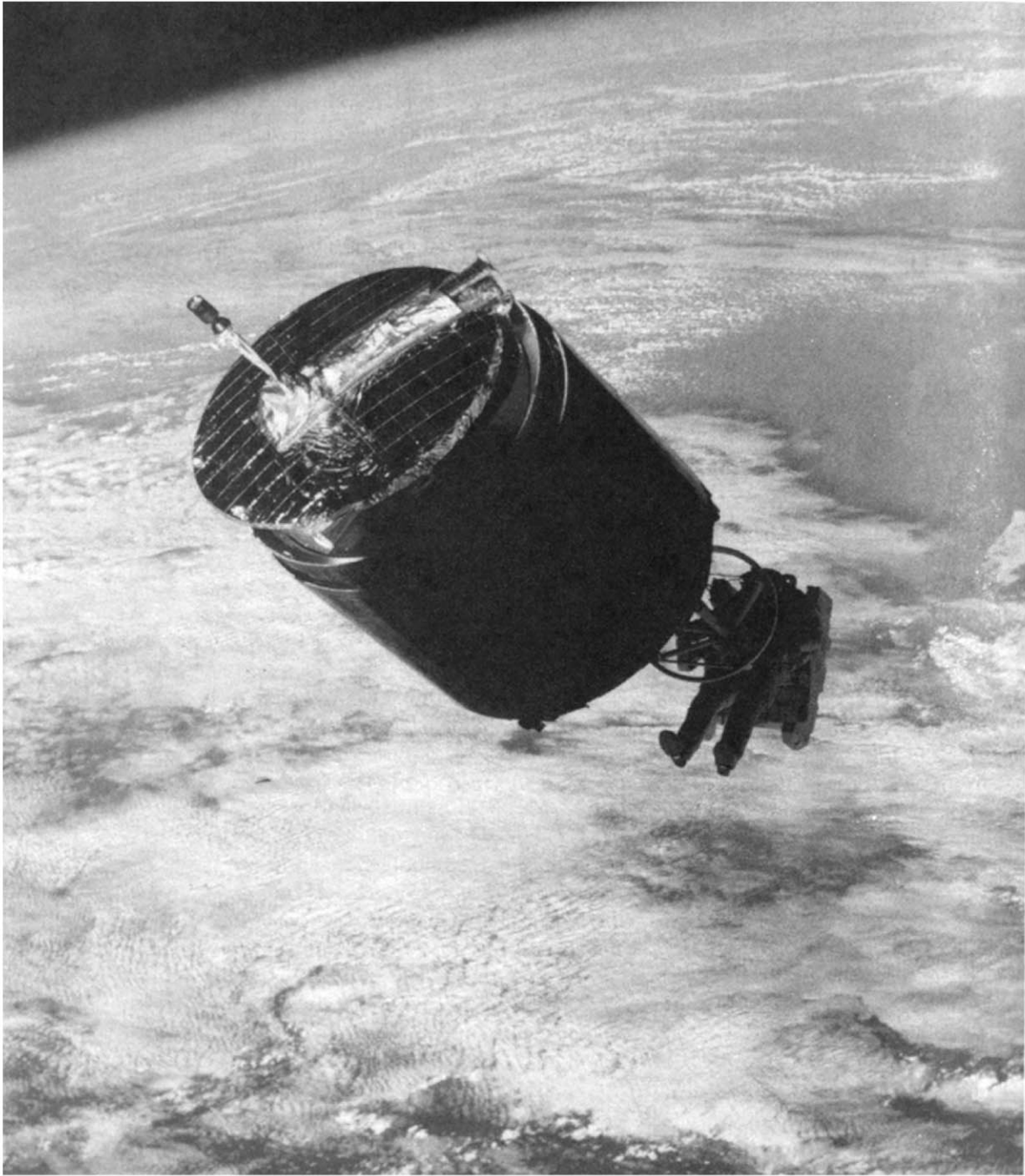
## References

- Billinton, R. and Allan, R. N. 1992. *Reliability Evaluation of Engineering Systems: Concepts and Techniques*, 2nd ed. Plenum, New York.
- Lewis, E. E. 1987. *Introduction to Reliability Engineering*. John Wiley & Sons, New York.
- Ramakumar, R. 1993. *Engineering Reliability: Fundamentals and Applications*. Prentice Hall, Englewood Cliffs, NJ.
- Shooman, M. L. 1990. *Probabilistic Reliability: An Engineering Approach*, 2nd ed. R.E. Krieger, Malabar, FL.

## Further Information

- Green, A. E. and Bourne, A. J. 1972. *Reliability Technology*. Wiley-Interscience, New York.
- Henley, E. J. and Kumamoto, H. 1991. *Probabilistic Risk Assessment<sup>3/4</sup>Reliability Engineering, Design, and Analysis*. IEEE Press, New York.
- IEEE Transactions on Reliability*. Institute of Electrical and Electronics Engineers, New York.
- O'Connor, P. D. T. 1985. *Practical Reliability Engineering*, 3rd ed. John Wiley & Sons, New York.
- Proceedings: Annual Reliability and Maintainability Symposium*. Institute of Electrical and Electronics Engineers, New York.
- Siewiorek, D. P. and Swarz, R. S. 1982. *The Theory and Practice of Reliable System Design*. Digital Press, Bedford, MA.
- Trivedi, K. S. 1982. *Probability and Statistics with Reliability, Queuing, and Computer Science Applications*. Prentice Hall, Englewood Cliffs, NJ.
- Villemeur, A. 1992. *Reliability, Availability, Maintainability and Safety Assessment, Volumes I and II*. John Wiley & Sons, New York.

Spitzer, C. R. "Aeronautical and Aerospace"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



Astronaut Dale A. Gardner, wearing the Manned Maneuvering Unit (MMU), prepares to dock with the spinning *Westar VI* satellite. Gardner made a hard dock with the stinger over the Bahama Banks at 6:32 A.M. (CST) on November 14, 1984.

Gardner used the Apogee kick motor Capture Device (ACD) to enter the nozzle of the *Westar* engine and stabilize the communications spacecraft sufficiently to capture it for the return to earth in the cargo bay of the *Discovery*.

This photograph represents the pinnacle of aeronautical and aerospace engineering for its time in 1984, more than ten years ago. For more information on the Manned Maneuvering Unit, see page 1866. (Photo courtesy of National Aeronautics and Space Administration.)



# XXVI

## Aeronautical and Aerospace

---

**Cary R. Spitzer**

*AvioniCon, Inc.*

**172 Aerodynamics** *J. F. Donovan*

Background • Flow about a Body • Two-Dimensional Airfoils • Finite Wing Effects • Effects of Compressibility

**173 Stability and Turbulence** *R. A. Hess*

Descriptions of Atmospheric Movement • Turbulence and Aircraft Dynamics • An Example • Other Applications

**174 Computational Fluid Dynamics** *R. K. Agarwal*

Geometry Modeling and Grid Generation • Flow Simulation Algorithms • Turbulence Modeling • Flow Simulation Examples • Future Directions and Challenges

**175 Aerospace Materials** *R. R. June*

System Requirements and Materials Selection • Design Considerations • Nonstructural Materials • The Future

**176 Propulsion Systems** *J. C. Monk*

Performance Characteristics • Liquid Rocket Engine Cycles • Major Components • System Preliminary Design Process • Conclusion

**177 Aircraft Performance and Design** *F. J. Hale*

Aircraft Forces and Subsystems • Level Flight • Climbing Flight • Turning Flight

**178 Spacecraft and Mission Design** *W. T. Fowler*

Spacecraft Environments • Fundamental Principles • Spacecraft/Mission Categories • Spacecraft Subsystems • Spacecraft/Mission Design Process

AERONAUTICAL AND AEROSPACE ENGINEERING offers extraordinary challenges. Aircraft and spacecraft missions can range from medical evacuation to air superiority to robotic exploration of distant planets. Flight distances can be semi-infinite—launched from earth but never returning—and speeds can range from zero to over 50,000 km/h. And all of this has to be accomplished with aircraft or spacecraft that ideally should have no weight and require no fuel.

Aerodynamics, a subdiscipline of fluid dynamics, is a concern for any object that moves through an atmosphere. For aircraft, it is the principal concern. **Chapter 172** lays the groundwork for understanding aerodynamics, beginning with Bernoulli's equation and continuing to more complex topics, such as supersonic flight.

Traditional aerodynamics has relied on wind tunnel testing to confirm analytical results; however, in recent times the emergence of computational fluid dynamics (CFD) has sharply reduced the need for such testing. CFD is a computationally demanding task that requires

supercomputers, but the speed and accuracy with which candidate aerodynamic configurations can be analyzed often makes CFD the preferred approach. **Chapter 174** examines the "mathematical wind tunnel."

No one has ever built a weightless aircraft or spacecraft. However, modern materials—specifically, advanced composites—have enabled ever lighter vehicles to accomplish the same or more demanding missions than their predecessors. As the capabilities of advanced materials have grown, their use in aircraft and spacecraft has more than kept pace. The broad spectrum of advanced materials properties, combined with the equally broad spectrum of aircraft and spacecraft missions and operating environments, dictates careful selection of materials.

**Chapter 175** reviews advanced materials for aerospace applications and describes a proven procedure developed by the Boeing Commercial Airplane Co. to ensure selection of the correct one.

Regardless of the mission, every aircraft and spacecraft must have mobility—mobility derived from propulsion. The fundamentals of propulsion, with emphasis on rocket examples, are found in **Chapter 176**.

The design of aircraft and spacecraft begins with a clear statement and understanding of the mission requirements; otherwise the result will be the wrong product performing the wrong functions. Because of the wide speed range and diversity of operating environments, noted in the first paragraph, aircraft design is a series of compromises, typically trading off maneuverability for weight or safety. For spacecraft, particularly unmanned ones, weight, operational flexibility, and reliability are the principal design factors. **Chapters 177** and **178** capture the salient points of the design process for their respective vehicles.

Donovan, J. F. "Aerodynamics"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 172

## Aerodynamics

---

### 172.1 Background

### 172.2 Flow about a Body

Lift • Drag

### 172.3 Two-Dimensional Airfoils

### 172.4 Finite Wing Effects

### 172.5 Effects of Compressibility

### John F. Donovan

*McDonnell Douglas Corporation*

*Aerodynamics* is a subset of fluid dynamics that deals with the flow of air about objects, typically aircraft, missiles, or their components. Much of the work in aerodynamics focuses on the generation of forces and moments on a body due to the air flowing over and through the body. Aerodynamics deals with theoretical and numerical predictions of performance characteristics, experimental determination of performance characteristics, and the design of new and improved geometries using this information. This chapter provides the basic understanding and equations to enable the engineer to calculate the performance of many aerodynamic configurations and interpret aerodynamic data.

## 172.1 Background

---

Understanding aerodynamics requires a knowledge of the basic principles of fluid mechanics. These principles are covered in Section VI. However, several key points required for the understanding of aerodynamics are provided in this section.

To understand how lift is generated, it is important to know the relationship between velocity and pressure in a fluid flow. For flows where the effects of viscosity are negligible, the density is constant, and gravity does not play a role, *Bernoulli's equation* describes this relationship:

$$p + \rho \frac{V^2}{2} = p_0 = \text{constant along a streamline} \quad (172.1)$$

where  $p$  is the static pressure,  $\rho$  is the density of the fluid medium, and  $V$  is the fluid velocity. The constant in Eq. (172.1) is the total pressure, denoted  $p_0$ , and is the sum of the static pressure,  $p$ , and the dynamic pressure,  $\rho V^2/2$ . A *streamline* is the path a fluid particle makes in the flow assuming the flow is steady, and Eq. (172.1) indicates that the total pressure is constant along this line. If the

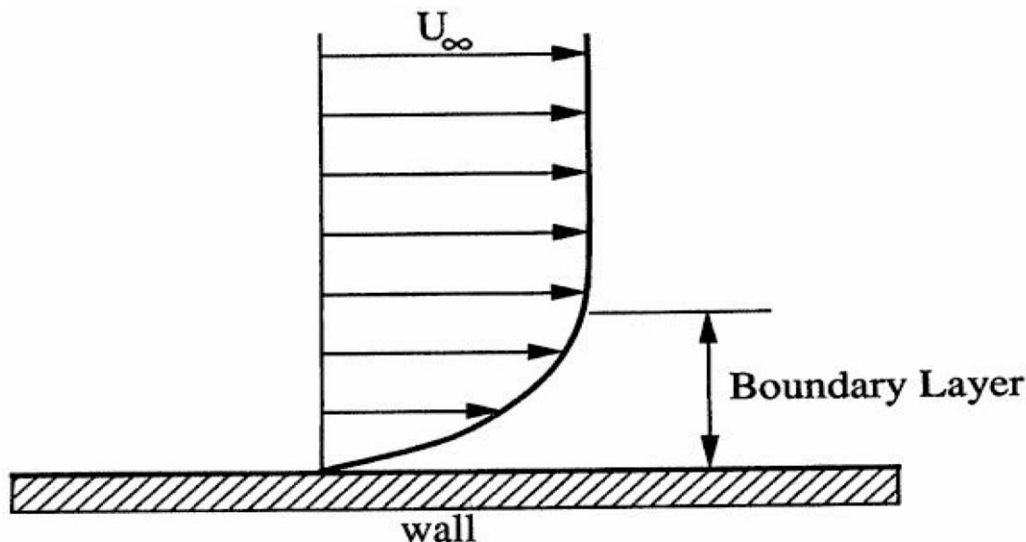
velocity and pressure are known at one point on the streamline, the flow quantities can be computed at another point on the streamline by rewriting Eq. (172.1) as

$$p_2 + \rho \frac{V_2^2}{2} = p_1 + \rho \frac{V_1^2}{2} \quad (172.2)$$

where the subscripts refer to two points along the streamline. The total pressure is a measure of the energy in the flow and Eq. (172.1) indicates that this energy is shared between the static pressure and the dynamic pressure (kinetic energy). If the velocity goes to zero at a point in the flowfield, this point is referred to as a *stagnation point* and the pressure there is the stagnation pressure, which is equal to the total pressure.

The effects of *viscosity* and the *no-slip* condition are also important concepts for the understanding of aerodynamics. When a fluid moves over the surface of a body, it actually "sticks" to the surface of the body so that there is no relative velocity between the fluid and the surface. This is called the no-slip condition and is caused by intermolecular forces and molecular-scale surface roughness. Due to the effects of viscosity, a *boundary layer* is formed near the surface where the velocity increases from zero at the surface to the freestream value,  $U_\infty$ , far away (see Fig. 172.1). Viscous effects are important only in the boundary layer for most streamlined bodies. Since viscous effects are important in the boundary layer, Bernoulli's equation cannot be applied there. In fact, in the absence of strong curvature, the pressure at the wall where the velocity is zero is equal to the static pressure in the outer flow and not the total pressure as predicted by Bernoulli's equation.

**Figure 172.1** Velocity profile in a boundary layer.



*Reynolds number* is a measure of the importance of viscous effects and is defined as the ratio of

momentum forces to viscous forces:

$$\text{Re} = \frac{U_{\infty} \rho l}{\mu} \quad (172.3)$$

where  $l$  is a length scale of the flowfield and  $\mu$  is the viscosity of the fluid. As the Reynolds number is increased, the effects of viscosity become less important.

## 172.2 Flow about a Body

---

This section describes the forces developed in a constant-density flow about general bodies. [Figure 172.2](#) illustrates the flow about an arbitrary two-dimensional body. The body is traveling through still air at a velocity of  $U_{\infty}$  and the figure is drawn in the reference frame of the body so the fluid appears to move past the body. Two forces are generated—**lift**,  $L$ , and **drag**,  $D$ . Lift is the force perpendicular to the incoming velocity and drag is the force parallel to the incoming velocity. A *moment*,  $M$ , is also exerted on the body. As the freestream velocity changes, the forces on the body change even if the flowfield remains approximately similar. It is therefore useful to define nondimensional force coefficients which vary much less with flow speed than the forces themselves. The forces are nondimensionalized by the dynamic pressure times an area associated with the body,  $S$  (in the case of a wing, it is usually the wing area). The lift and drag coefficients are defined as

$$C_L \equiv \frac{L}{\frac{1}{2} \rho U_{\infty}^2 S} \quad (172.4)$$

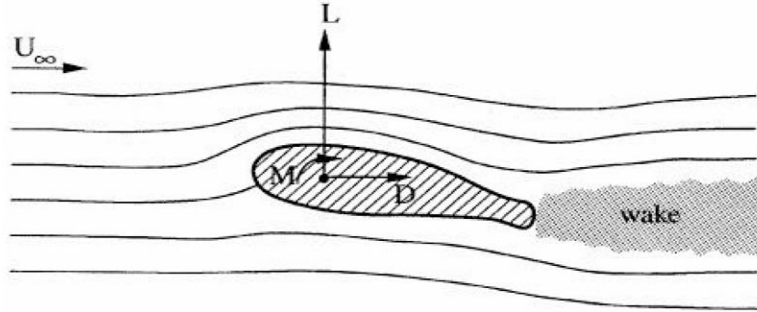
$$C_D \equiv \frac{D}{\frac{1}{2} \rho U_{\infty}^2 S} \quad (172.5)$$

The moment coefficient is defined using an additional parameter,  $l$ , a length scale of the body (in the case of a wing, it is usually the streamwise length of the wing), as

$$C_M \equiv \frac{M}{\frac{1}{2} \rho U_{\infty}^2 S l} \quad (172.6)$$

The particular choice of the reference area and length is not critical, but when using published data it is important to know which reference quantities the coefficients are based upon.

**Figure 172.2** Flow about an arbitrary two-dimensional body illustrating the aerodynamic forces and moment exerted by the flowfield. The reference frame is fixed to the body.



Two types of forces act on the surface of a body due to the motion of the fluid—*pressure forces* and *viscous forces*. Pressure forces are simply due to the pressure,  $p$ , and act perpendicular to the surface. Viscous forces, or friction forces, are caused by the flow "rubbing" against the surface as it passes over the body. This force acting on a unit area of the surface is termed the *wall shear stress*,  $\tau_w$ , and acts tangentially to the surface. It is proportional to the velocity gradient normal to the wall at the wall, and is defined as

$$\tau_w = \mu \left. \frac{\partial U}{\partial y} \right|_w \quad (172.7)$$

where  $\mu$  is the viscosity and the subscript  $w$  denotes that the derivative is evaluated at the wall. No matter how complex the surface geometry or the flowfield, the only way aerodynamic forces are transmitted to a body is through these two mechanisms.

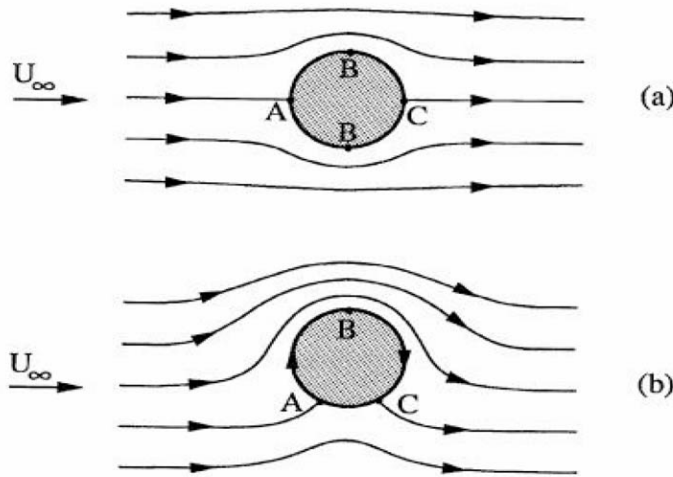
## Lift

A lift force is generated on a body by pressure differences between the top and bottom of the body. If the average pressure over the bottom of the body is higher than the average pressure over the top of the body, the lift force points upward. If we consider the flow over a two-dimensional cylinder without viscosity, the streamline pattern is symmetric, as shown in Fig. 172.3(a). The presence of the cylinder in the flow requires the same amount of fluid to pass in a smaller area. Thus, the fluid must speed up in some regions of the flowfield. A fluid element moving along the centerline slows down to zero velocity at the stagnation point  $A$ , where the pressure is equal to the total pressure. As the fluid moves away from the center of the cylinder, it speeds up, and a minimum pressure occurs in the two regions marked  $B$ . This reduction in pressure can be seen using Bernoulli's equation [Eq. (172.1)] because, as the flow speeds up, the static pressure drops in order for the total pressure to remain constant. The fluid slows down again towards the rear of the cylinder as the streamlines come together and another stagnation point is formed at  $C$ . The total lift force is the integral of the pressure over the entire surface in a direction perpendicular to the freestream flow:

$$L = \int p dA_y \quad (172.8)$$

where  $A_y$  is the component of the area perpendicular to the freestream flow. Thus, on the upper surface of the cylinder, there is an upward force, but there is an equal and opposite downward force on the lower surface because the flow is symmetric, and the net lift is zero.

**Figure 172.3** Flow about a cylinder without spin (a) and with spin (b).



If the cylinder is spun about its axis, as shown in Fig. 172.3(b), the flow pattern becomes asymmetric. The stagnation points at A and C have moved below the centerline and the minimum pressure occurs only on the upper surface. In this case, it can be seen that the average pressure over the lower half of the cylinder will be higher than that over the top because the stagnation points are both on the bottom. In this case, there is a net lift force pointing up. The same effect is used in the aerodynamics of sport balls. For example, in tennis, a top spin on the ball causes it to drop much more quickly than a ball without top spin because of the downward-pointing lift vector.

By spinning the cylinder, a *circulation* has been introduced about the body. Circulation,  $\Gamma$ , is related to the average velocity tending to move fluid elements around the body. This average is made up of velocities on a path encircling the body. However, a nonzero value of circulation does not indicate that the fluid elements are actually moving around the cylinder. As the cylinder is spun faster, and  $\Gamma$  is thus increased, the lift force increases.

The relationship between lift and circulation is termed the *Kutta-Joukowski theorem* after the two people who independently discovered it, and is expressed as

$$L' = \rho U_\infty \Gamma \quad (172.9)$$

where  $L'$  is the lift per unit span of the cylinder. This theorem applies to any two-dimensional body



about which a circulation exists, not just a cylinder. Circulation about a body can be represented as a **vortex**. A vortex is a circular flowfield in which the fluid rotates about a common axis, and the farther the fluid element is from the axis, the slower it moves around the axis. In the case of lift generation, the vortex is not attached to the same fluid elements, but is "bound" to the body and is often called a *bound vortex*.

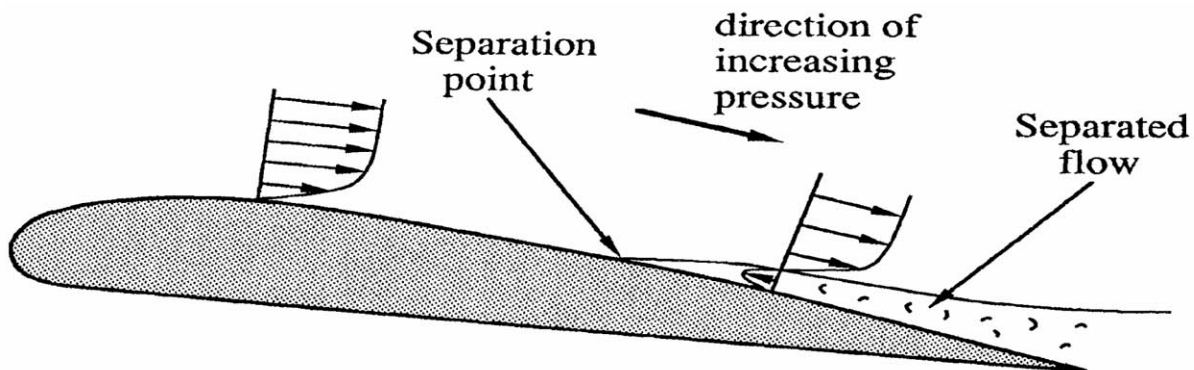
## Drag

For a general body, the total drag force results from two sources—*pressure drag* and *viscous drag*. Pressure drag is simply the component of the total pressure force on the body that is in the streamwise direction (parallel to the direction of the freestream). In many flows, *boundary layer separation* causes significant pressure drag. Boundary layer separation can occur when a boundary layer flows into a region of increasing pressure. For fluid elements outside the boundary layer, this increase in pressure is traded off as a decrease in velocity, as can be seen by Bernoulli's equation. Because the pressure is approximately constant across the boundary layer, the fluid near the wall must also undergo an increase in pressure. The fluid near the wall has less energy than the flow at the edge of the boundary layer because it is moving slower. If the pressure increases too rapidly in the downstream direction, the velocity near the wall can reverse. If this happens, the boundary layer is said to have separated, as shown in Fig. 172.4. The boundary layer leaves the surface and a recirculating region is developed between the airfoil surface and the boundary layer fluid that left the surface. As the flow is not attached to the surface, it does not continue on a path which leads to a higher pressure in that region as it would if the flow had not separated. The result is a lower pressure acting over the surface. This surface usually has a component of area facing downstream and the low pressure acting on this area is seen as drag (pressure drag). Pressure drag,  $D_p$ , is calculated as

$$D_p = \int p dA_x \quad (172.10)$$

where  $A_x$  is the streamwise component of the area. For this reason, boundary layer separation is usually considered to have a negative effect on a flow.

**Figure 172.4** Diagram illustrating boundary layer separation.



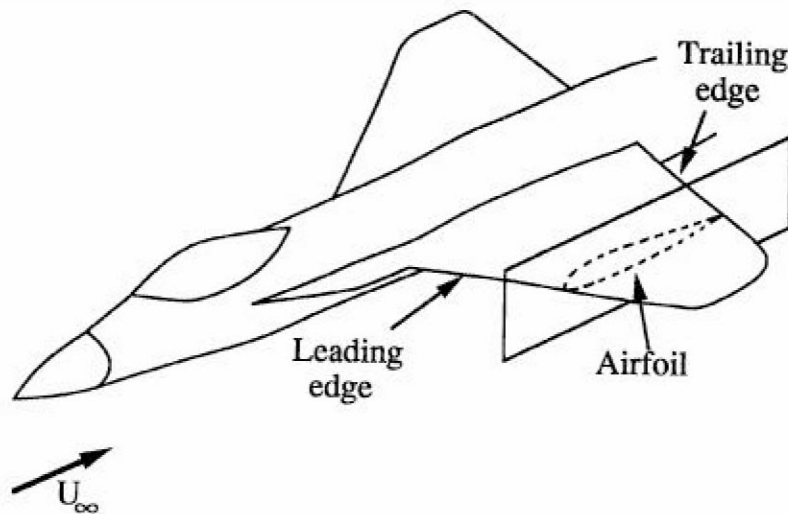
Viscous drag,  $D_v$ , is present in all vehicles which move through a fluid. It is caused by the wall shear stress, which was discussed earlier in this section. Wall shear stress acts tangentially to the surface, so the total viscous drag is determined by integrating the streamwise component of  $\tau_w$  over the surface:

$$D_v = \int \tau_w dA_x \quad (172.11)$$

### 172.3 Two-Dimensional Airfoils

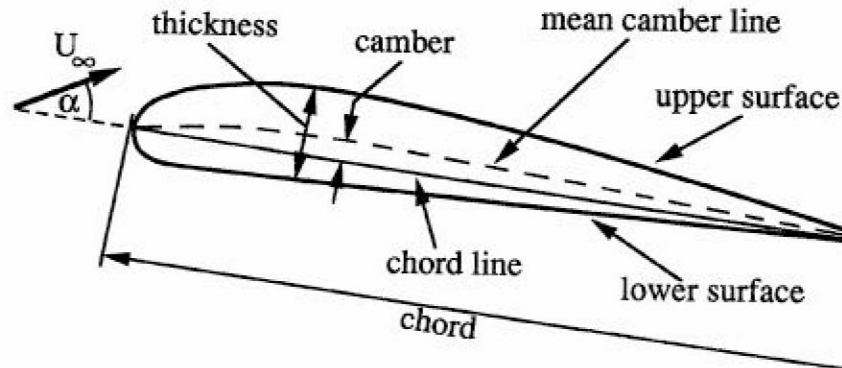
Typically, wings have been analyzed by splitting the problem into two parts: (1) analysis of a wing section (*airfoil*), and (2) modification of the airfoil properties to account for the effects of the complete, finite wing. An airfoil is the cross section of a wing in a plane parallel to the freestream velocity and perpendicular to the wing, as shown in Fig. 172.5.

**Figure 172.5** Diagram illustrating the relationship between an aircraft wing and an airfoil.



Airfoils are usually described using the nomenclature indicated in Fig. 172.6. The longest dimension of the airfoil section is termed the *chord*, and the line connecting the *leading edge* to the *trailing edge* is the *chord line*. Airfoil *thickness* is the maximum distance between the upper and lower surfaces along a line perpendicular to the chord line. Thickness is usually expressed as a percentage of the chord (e.g., a 12% thick airfoil has a maximum thickness of 12% of the chord). The *mean camber line* is the locus of points halfway between the upper and lower surfaces. The variation of the thickness along the mean camber line is termed the *thickness distribution*. *Camber* of the airfoil is defined as the maximum distance between the mean camber line and the chord line in a direction perpendicular to the chord line. When a fluid is moving past the airfoil, the angle of attack is defined as the angle between the freestream velocity and the chord line, denoted as  $\alpha$  in Fig. 172.6.

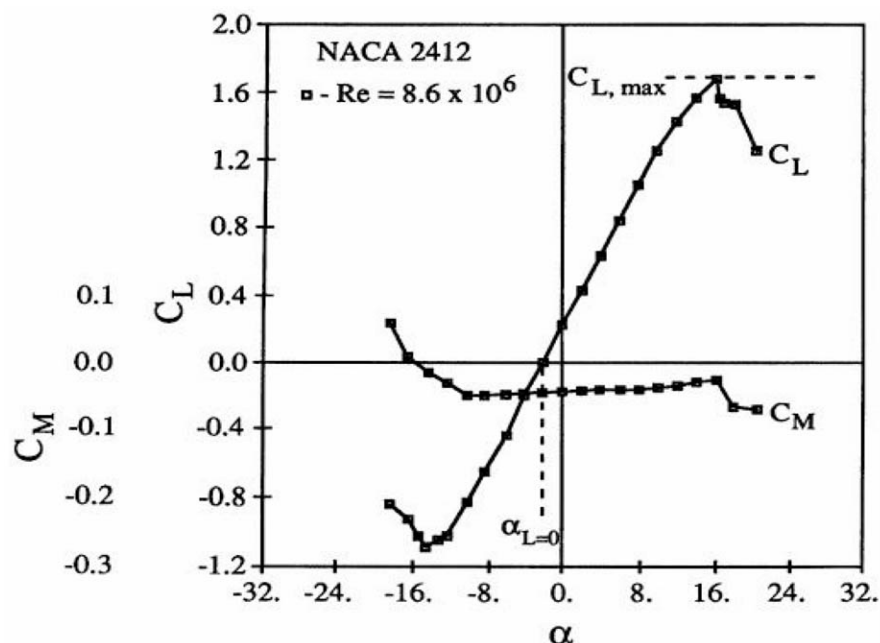
**Figure 172.6** Airfoil terminology.



Airfoil designs that cover a wide range of thickness, thickness distributions, and camber have been designed and tested by the National Advisory Committee for Aeronautics (NACA) (the predecessor of the current NASA). These include the *four-digit* series (e.g., the NACA 2412), the *five-digit* series (e.g., the NACA 23012), and the *six-digit* series (e.g., the NACA 65-218). The numbers indicate different characteristics of the airfoils. A complete description of the numbering system can be found in Abbott and von Doenhoff [1959]. These and other airfoil designs and their performance characteristics can be found in the books in the Further Information section.

In selecting an airfoil for a particular application, the variation of lift, drag, and moment with angle of attack is necessary. A typical variation of lift coefficient with angle of attack is shown in Fig. 172.7. Recall from the earlier discussion that this data represents the performance of a two-dimensional airfoil and is sometimes referred to as *infinite wing data*. The value of  $\alpha$  when the lift is zero is termed  $\alpha_{L=0}$ , the zero-lift angle of attack. As the angle of attack is increased from  $\alpha_{L=0}$ , the lift increases linearly over a wide range of  $\alpha$ . The slope of this line is the *lift slope*. As  $\alpha$  becomes large, the pressure gradients on the upper surface of the airfoil also become large, and separation of the boundary layer occurs. Separation causes a large reduction in lift and the airfoil is said to be stalled. The maximum lift generated by the airfoil is indicated by  $C_{L,max}$ . Before the decrease in lift occurs, the curve becomes nonlinear because of viscous effects.

**Figure 172.7** Lift and moment coefficients as a function of angle of attack for the NACA 2412 airfoil.



In the linear region of the curve where viscous effects have a small impact on the lift, *thin airfoil theory* can be used to predict the airfoil behavior. If the airfoil thickness is small in comparison to its chord, the flowfield about the airfoil can be represented by distributing vortices along the mean camber line. An infinite number of solutions are possible for the distribution of the strengths of the vortices as a function of chord. All but one of these solutions is eliminated by applying the *Kutta condition*. The Kutta condition essentially requires that the flow be physically realistic in that the fluid cannot flow around a sharp trailing edge. More specifically, it states that the following be true: (a) the value of the circulation,  $\Gamma$ , is such that the flow leaves the trailing edge smoothly; (b) if the included angle at the trailing edge is finite, then the trailing edge is a stagnation point; and (c) if the trailing edge is cusped (zero included angle), then the velocities from the upper and lower surfaces are equal in magnitude and direction at the trailing edge. Thin airfoil theory then yields a lift slope of  $2\pi$ , so in the case of a symmetric airfoil  $C_L = 2\pi\alpha$ . For relatively thin airfoils, less than about 12%, this prediction is accurate to within several percent of experimentally determined values.

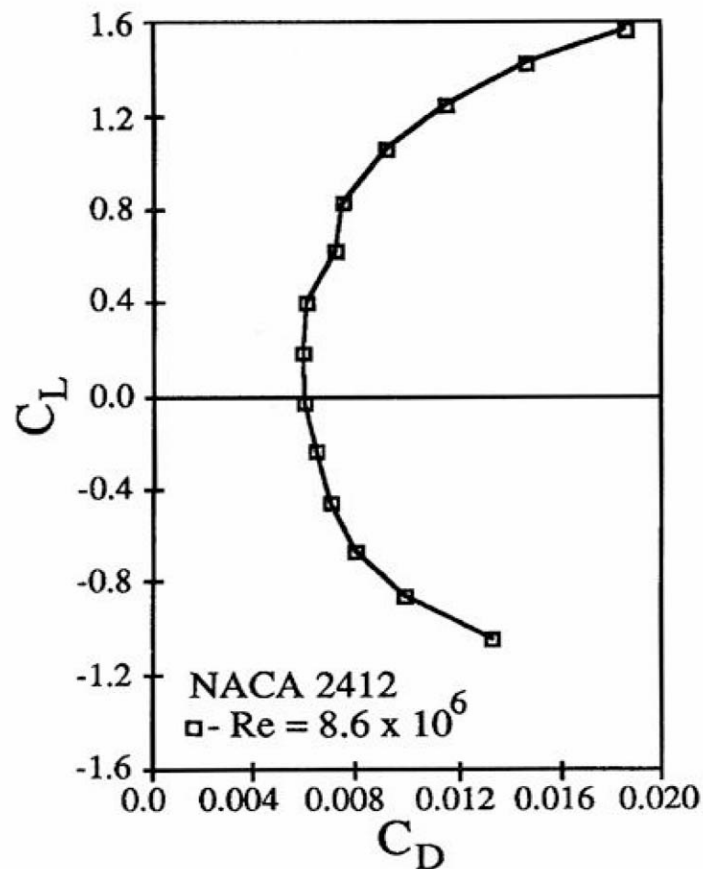
Camber controls the amount of lift generated at zero angle of attack. A symmetric airfoil (zero camber) generates no lift at  $\alpha = 0$ . An airfoil with positive camber generates lift at  $\alpha = 0$ , and zero lift occurs at a negative angle of attack, as shown in Fig. 172.7.

Figure 172.7 also shows the variation of moment coefficient with angle of attack for a typical cambered airfoil. The moment is usually taken about a point one quarter of the way along the chord line from the leading edge. This point is known as the *quarter-chord point*. Note that the value of the moment is negative, indicating that the moment acting on the airfoil tends to rotate it in the counterclockwise direction for the airfoil depicted in Fig. 172.6. Camber also affects the moment—a symmetric airfoil produces no moment and an increase in the camber causes an increase in the magnitude of moment.

In Fig. 172.7, the moment coefficient is approximately constant over a wide range of angle of attack. The quarter-chord point was chosen as the point about which to determine moments for specifically this reason—it is the *center of pressure* for an airfoil. The center of pressure is defined as the point about which the moment is constant, independent of angle of attack. Another useful point in aerodynamics is the *aerodynamic center*, which is the point about which the moment is zero. For a symmetric airfoil, this point happens to be also at the quarter-chord point. In general, the aerodynamic center may or may not lie on the body itself. As viscous effects become important, the moment varies from its ideal behavior, as did the lift behavior. Near stall, the magnitude of the moment increases because of boundary layer separation altering the static pressure distribution.

Drag data is usually presented as a function of lift coefficient, instead of angle of attack, in what is termed a *drag polar*. An example of a drag polar is shown in Fig. 172.8 for a cambered airfoil. This drag coefficient includes both the viscous and pressure drags. The penalty in drag for generating high lift coefficients is clear from Fig. 172.8. In addition, when the airfoil approaches stall, and the upper surface boundary layer separates, the drag increases dramatically. This increase is caused by the large pressure drag associated with separation, as indicated in section 172.2. In many applications, the ratio of lift to drag,  $L/D$ , is an important performance parameter. The  $L/D$ s of different airfoils can easily be compared using a drag polar simply by looking at the slope of a line drawn between a point on the polar and the origin of the plot.

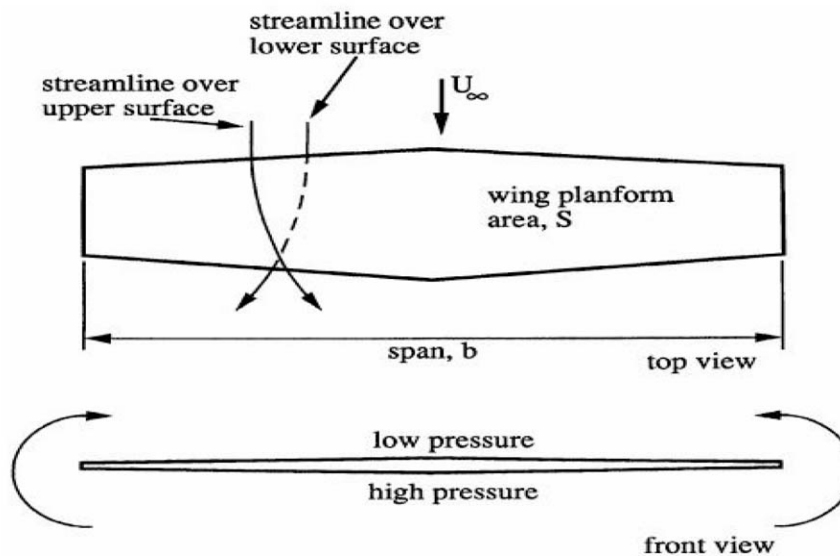
**Figure 172.8** Drag polar for the NACA 2412 airfoil.



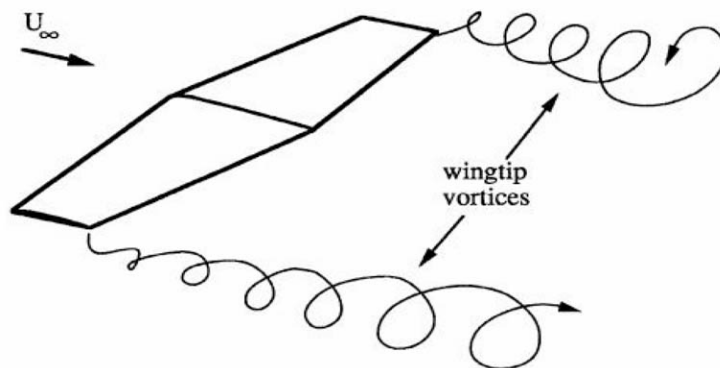
## 172.4 Finite Wing Effects

Two-dimensional, or infinite wing, performance was discussed in the previous section. Of course, all real aircraft have finite wings. In this section, the modification of infinite wing performance due to the finite nature of real wings is presented. The primary difference between the two cases is that, in the finite wing, the flow can be three-dimensional (i.e., flow can occur into and out of the paper in [Fig. 172.4](#) as well as in the plane of the paper). If a wing is generating lift, then the average pressure on the lower surface must be higher than that on the upper surface. Near the wing tips, this pressure difference causes flow from the lower surface to the upper surface, as shown in [Fig. 172.9](#). This leakage around the wing tips causes the flow to rotate about a streamwise axis as it leaves the wing tip, forming *wing tip vortices*, as shown in [Fig. 172.10](#). Inboard of the wing tips, the flow has a spanwise component due to the leakage around the wing tips. As shown in [Fig. 172.9](#), the flow over the lower surface tends to move out toward the wing tip, and the flow over the upper surface tends to move inboard. Clearly, the flow is not two-dimensional, as was assumed in the previous section.

**Figure 172.9** Finite wing effects.



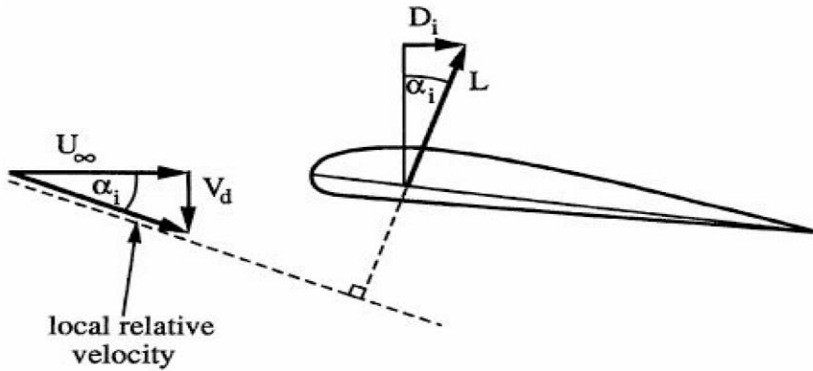
**Figure 172.10** Wing tip vortices formed by a finite wing.



Wing tip vortices induce flow about their axes. In the region of the wing, it can be seen from [Fig. 172.9](#) that the resulting induced flow will be downward. This induced downward flow is termed *downwash* and locally rotates the incoming velocity vector, as shown in [Fig. 172.11](#). Locally, the angle of attack is reduced by  $\alpha_i$ . Thus, the effective angle of attack is  $\alpha_{\text{eff}} = \alpha - \alpha_i$ . Lift is generated perpendicular to the local freestream velocity, and as it has been rotated by  $\alpha_i$ , then the lift vector is also rotated by  $\alpha_i$ . The lift vector now has a component parallel to the freestream that is far from the wing,  $D_i$ —the *induced drag*. In summary, the downwash induced in the region of the wing by the wing tip vortices reduces the effective angle of attack of the airfoil section. This tilts the lift vector downstream, resulting in a component of lift acting in the drag direction. Typical airfoil sections might have a  $L/D$  ratio of about 100. Thus, only a small rotation of the large lift vector can cause a significant increase in drag, which appears as pressure drag.



**Figure 172.11** Diagram illustrating the tilting of the lift vector due to the downwash.

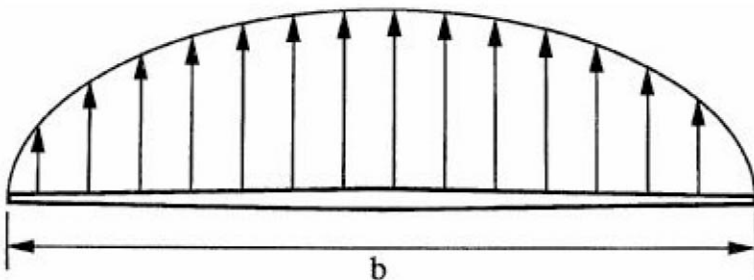


The induced drag of a wing depends on the downwash distribution along its span. It can be shown that the downwash distribution depends on the lift distribution across the wing span. Lift per unit span can vary for several reasons: (1) changes in chord; (2) changes in angle of attack; and (3) variations in airfoil shape. An elliptical distribution of lift across the span, as shown in [Fig. 172.12](#), provides a minimum induced drag. Lifting-line theory can be used to predict the downwash for an elliptical lift distribution and a detailed discussion of this technique can be found in Kuethe and Chow [1986]. In this case, the drag coefficient for the induced drag is

$$C_{D,i} = \frac{C_L^2}{\pi AR} \quad (172.12)$$

where  $AR$  is the *aspect ratio* of the wing defined as  $AR \equiv b^2/S$ . It can be seen from Eq. (172.12) that, as the aspect ratio increases (i.e., the wing becomes longer and more slender), the induced drag is reduced. This equation also indicates the dependence of induced drag on lift. As the lift increases, the strength of the tip vortices increases, and therefore the downwash increases, which causes the lift vector to tilt more in the downstream direction.

**Figure 172.12** Elliptical lift distribution over a finite wing.



In practice, it is difficult to obtain an elliptical lift distribution. Thus, the induced drag is higher than that for an elliptical distribution. In general,

$$C_{D,i} = \frac{C_L^2}{\pi e AR} \quad (172.13)$$

where  $e$  is the *span efficiency factor*. For elliptical wings,  $e = 1$ , but for typical subsonic aircraft,  $0.85 < e < 0.95$ . Thus, the total drag for a finite wing is the sum of the induced drag and the infinite wing drag:

$$C_{D,\text{tot}} = C_{D,2-D} + \frac{C_L^2}{\pi e AR} \quad (172.14)$$

In addition to affecting the drag of two-dimensional airfoil data, the finite nature of a wing also reduces the lift slope. The downwash reduces the effective angle of attack, and therefore reduces the lift generated. For a general wing, the effective angle of attack is

$$\alpha_{\text{eff}} = \alpha - \frac{57.3 C_L}{\pi e' AR} \quad (172.15)$$

where  $e'$  is another span effectiveness factor, which, in practice, is approximately equal to  $e$ . The reduced lift for a given angle of attack indicates a reduction in the lift slope,  $a$ , which becomes

$$a = \frac{a_0}{1 + 57.3 a_0 / (\pi e' AR)} \quad (172.16)$$

where  $a_0$  is the two-dimensional lift slope and  $a$  is the lift slope including finite wing effects.

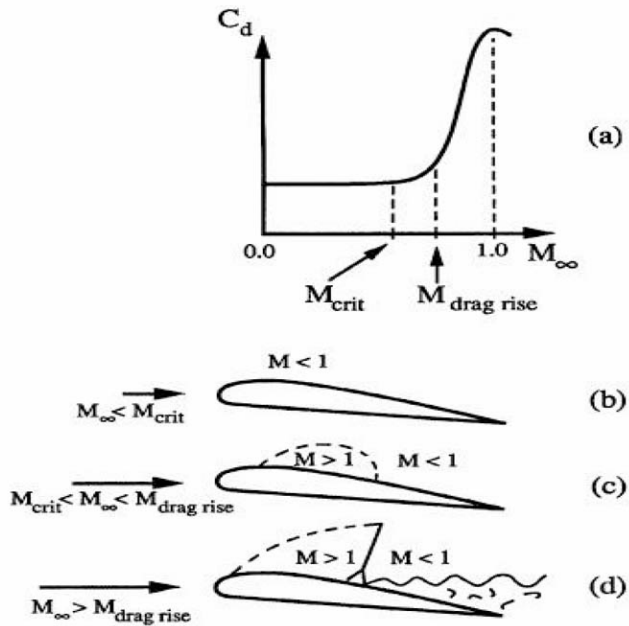
## 172.5 Effects of Compressibility

It has been assumed to this point that the flow speed has been low enough that the effects of compressibility of the fluid have been negligible (Section VI describes compressibility in more detail and should be reviewed at this time if the reader is unfamiliar with this topic). However, as the freestream speed and thus the freestream Mach number ( $M_\infty$ ) are increased, compressibility effects are seen. Only a brief discussion of some of these effects is presented here.

If the local Mach number increases above 1.0, there exists the possibility for the formation of shock waves. This effect is one of the major differences between compressible and incompressible flows and is depicted in Fig. 172.13. Over the upper surface of an airfoil, the flow moves faster than the freestream. Thus, the Mach number in a region above the upper surface can be supersonic even if the freestream Mach number is less than one [see Fig. 172.13(c)]. The freestream Mach number at which the flow over the upper surface reaches Mach 1.0 is termed the *critical Mach number*,  $M_{\text{crit}}$ . As the freestream Mach number increases, a shock wave forms on the upper surface, causing boundary layer separation because of the strong adverse pressure gradient [see Fig. 172.13(d)]. In this case, the drag of the airfoil increases dramatically, as shown in Fig. 172.13(a). The Mach number where the drag begins to rise is the *drag divergence Mach number*,  $M_{\text{drag rise}}$ .

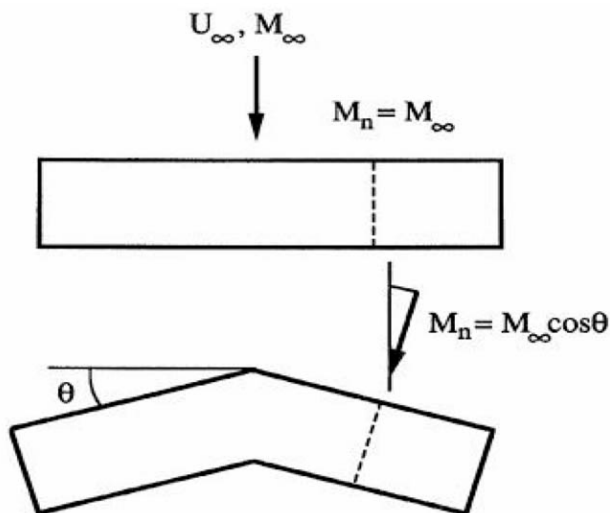


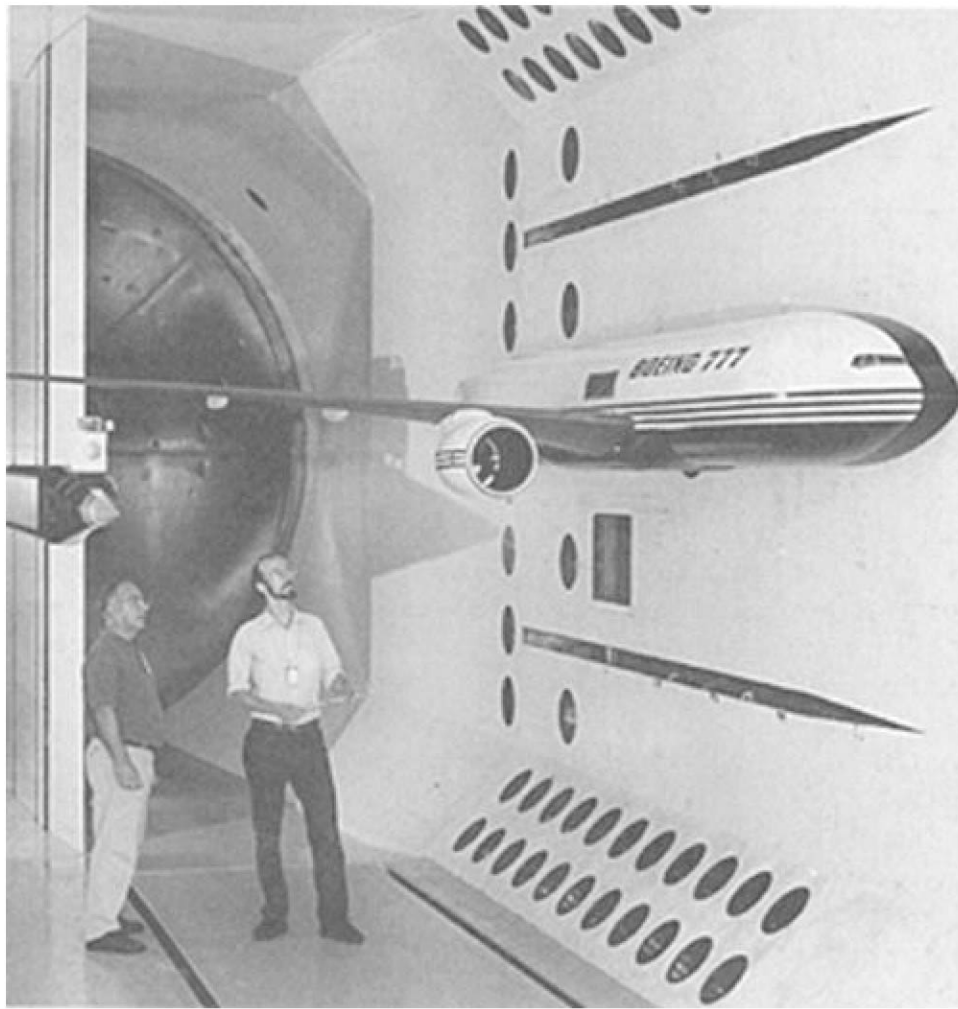
**Figure 172.13** Drag as a function of Mach number (a), and airfoil flowfields associated with the three Mach number ranges (b–d).



High-speed subsonic aircraft often have swept wings. The purpose of wing sweep is to increase  $M_{drag\ rise}$  so the aircraft can fly faster. By sweeping the wing back, the airfoil "sees" only the component of the freestream Mach number normal to the leading edge, which is lower than the total freestream Mach number, as shown in Fig. 172.14. Thus, the critical Mach number with sweep should be  $M_{cr, sweep} = M_{cr}/\cos \theta$ . In practice, because of the complex flowfield induced by wing sweep, the benefit of sweep is less than this. A more complete discussion of compressibility effects in aerodynamics can be found in Anderson [1984].

**Figure 172.14** The effect of wing sweep on the normal Mach number.





### WIND TUNNEL TESTING

The cornerstone of aerodynamics is wind tunnel testing. Here a model of the state-of-the-art Boeing 777 undergoes rigorous testing prior to the first flight of the actual Boeing 777 on June 12, 1994. This unique testing facility, the Transonic Dynamics Tunnel (TDT), is a national facility dedicated to aeroelastic research and engineering of high-speed aircraft and rotorcraft.

The TDT, a continuous-flow, closed-circuit, variable-density pressure tunnel, was completed in 1959. It is the world's first aeroelastic testing tunnel. It can test cable-mounted, sidewall-mounted, sting (strut)-mounted, or floor-mounted aircraft models. The test medium is Freon 12 or air. It operates in a speed range of up to Mach 1.2 (1.2 times the speed of sound) at pressures up to 350 lb/ft<sup>2</sup>. The TDT was designed to use Freon rather than air in order to achieve the desired tunnel test speed of Mach 1.2 at required pressure levels.

The principal focus of wind tunnel aeroelasticity studies today is "flutter." As an aircraft flies through the air, its wings and tail vibrate at certain frequencies, a harmless combination of structural responses to aerodynamic pressures. However, almost every aircraft is susceptible at certain speeds to a potentially harmful interplay of the wing vibrations and aerodynamic forces called flutter. In the most severe cases, unchecked flutter can damage or destroy a wing. Wind tunnel tests can identify a potential flutter problem, and changes can be made to the aircraft to make it safe to fly.

Future aircraft model tests in the TDT will ensure that new designs of high-speed aircraft and space vehicles are flutter free. (Photo courtesy of NASA, Langley.)

## Defining Terms

**Drag:** The net aerodynamic force on a body in a direction parallel to the freestream.

**Lift:** The net aerodynamic force on a body in a direction perpendicular to the freestream direction and to the wing platform.

**Vortex:** A flowfield in which fluid elements rotate about a common axis. In a vortex where viscous effects are not important, the velocity at which a fluid element circles the axis is inversely proportional to its distance from the axis.

**Wake:** The region of a flowfield downstream of a body where the momentum is less than that of the freestream.

## References

- Abbott, I. H. and von Doenhoff, A. E. 1959. *Theory of Wing Sections*. McGraw-Hill, New York.
- Anderson, J. D. 1978. *Introduction to Flight*. McGraw-Hill, New York.
- Anderson, J. D. 1984. *Fundamentals of Aerodynamics*. McGraw-Hill, New York.
- Kuethe, A. M. and Chow, C. Y. 1986. *Foundations of Aerodynamics*, 4th ed. John Wiley & Sons, New York.
- Schlichting, H. 1968. *Boundary Layer Theory*, 6th ed. McGraw-Hill, New York.
- Selig, M. S., Donovan, J. F., and Fraser, D. B. 1989. *Airfoils at Low Speeds*. Stokely, Virginia Beach, VA.

## Further Information

A much more complete discussion of aerodynamics can be found in *Fundamentals of Aerodynamics* by Anderson, and in *Foundations of Aerodynamics* by Kuethe and Chow. A good introductory text is *Introduction to Flight* by Anderson.

Detailed lift, drag, and moment data for a wide variety of aerodynamic shapes over a wide speed range can be found in *Fluid Dynamic Drag* by Hoerner. A good source of low-Reynolds number airfoil data is *Airfoils at Low Speeds* by Selig, Donovan, and Fraser.

Boundary layers were only mentioned briefly in this section, but for further information, *Boundary Layer Theory* by Schlichting is a good source.

Hess, R. A. "Stability and Turbulence"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Stability and Turbulence

---

### 173.1 Descriptions of Atmospheric Movement

Background • Simplified Descriptions

### 173.2 Turbulence and Aircraft Dynamics

Fourier Integral Representation • Aerodynamic Force/Moment Prediction

### 173.3 An Example

### 173.4 Other Applications

#### **Ronald A. Hess**

*University of California*

The aerodynamic forces which allow the sustained flight of heavier-than-air vehicles are created by the relative motion of the aircraft and the fluid mass through which it moves. In the case of a nonquiescent atmosphere, the motion of the fluid mass itself contributes to this relative motion. Since the atmosphere is rarely quiescent, the subject of aircraft stability and control in a moving atmosphere is of considerable importance to aeronautical engineers from a standpoint of performance, passenger comfort, and safety.

## 173.1 Descriptions of Atmospheric Movement

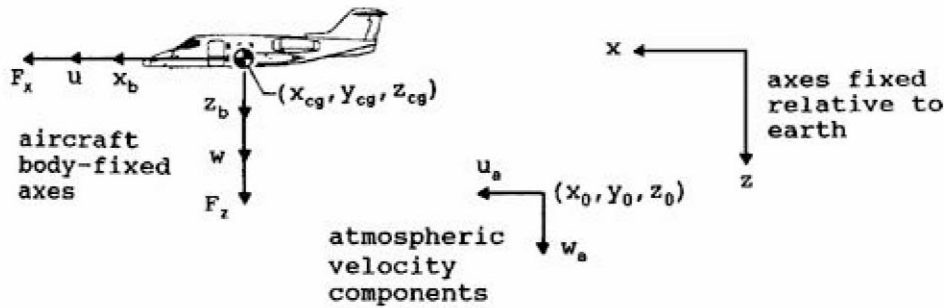
---

### **Background**

The fundamental approach which physicists and engineers adopt in describing the motion of a continuum, such as the atmosphere, is a field description. The velocity of the atmosphere is considered to be a continuous function of location and time, referred to as a velocity field.

Consider Fig. 173.1, showing an orthogonal axis system fixed relative to the earth. At some points  $x_0$ ,  $y_0$ , and  $z_0$ , and at some time instant  $t_0$ , the three components of the velocity of the atmosphere are  $u_a(x_0, y_0, z_0, t)$ ,  $v_a(x_0, y_0, z_0, t)$ , and  $w_a(x_0, y_0, z_0, t)$ . The preceding description allows a convenient means of categorizing atmospheric motion.

**Figure 173.1** Axes systems for description of atmospheric movement.



### Mean Winds

If, for example,  $u_a = U$ ,  $v_a = V$ , and  $w_a = W$ , where  $U$ ,  $V$ , and  $W$  represent constants, one has defined a **mean wind**. Obviously, this is an idealized description, but nonetheless a useful one. For example, airliner flight in the jetstream can often be adequately described by such a mean wind condition.

### Wind shears

If, for example,  $u_a = u_a(x)$ ,  $v_a = V$ , and  $w_a = W$ , one has defined a simplified **wind shear**—a velocity field in which the velocity varies with position, but not with time. Here, the changes in  $x$  ( $\Delta x$ ) over which significant changes in  $u_a$  occur are assumed to be such that  $\Delta x \gg l_f$ , where  $l_f$  represents the aircraft fuselage length.

### Turbulence

If, for example,  $u_a = U$ ,  $v_a = V$ , and  $w_a = w_a(x, t)$ , and if the changes  $\Delta x$  over which significant changes in  $w_a$  occur are such that  $\Delta x$  is the same order of magnitude as  $l_f$ , one has defined atmospheric **turbulence** or **gusts**. If, as in Fig. 173.1, the direction of the  $w_a$  axis has been chosen to be vertical, the field just described is referred to as a vertical turbulence field or a one-dimensional *upwash* field [Etkin, 1972].

The intensity of turbulence is typically quantified by the **root mean square (RMS) value** of each of the components. Thus, at some point  $x_0$  in the one-dimensional upwash field,

$$\sigma_{w_a} \triangleq \text{RMS value of } w_a(x_0, t) = \sqrt{\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T w_a^2(x_0, t) dt} \quad (173.1)$$

### Simplified Descriptions

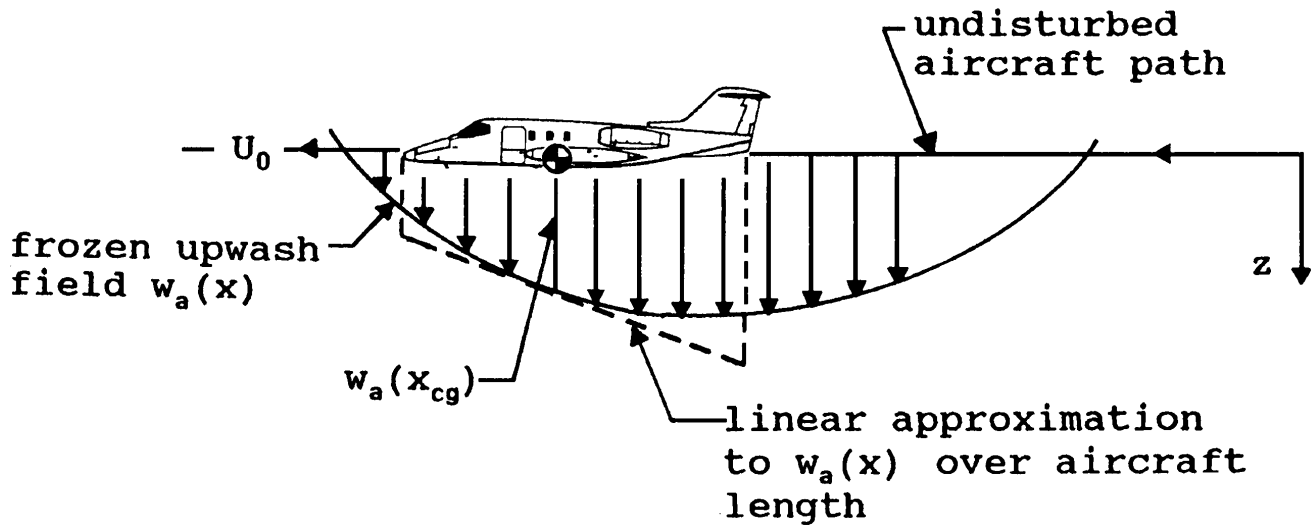
Concentrating upon turbulence, it is often assumed in aircraft stability and control analyses that the statistical description of the turbulence field (e.g., the RMS velocity at a point) is not a function of time, the origin, or the orientation of the axis system used to describe the field. If, as is typically the case, the aircraft is moving through the field with a velocity that is significantly larger in magnitude than the RMS value of the turbulence velocity at any point along its path, then the

spatial and temporal variation of the turbulence field may be replaced by the spatial variation alone. Thus,  $w_a(x, t) = w_a(x)$ . This is referred to as a **frozen turbulence field hypothesis** or **Taylor's hypothesis** [Houbolt, 1973].

## 173.2 Turbulence and Aircraft Dynamics

Figure 173.2 is a representation of an aircraft flying through the one-dimensional frozen upwash field described in the preceding section. The velocity profile of the turbulence field is represented in the figure, with the vertical arrows indicating the magnitude and direction of the atmospheric velocity at each point on the idealized linear flight path. The problem now becomes one of determining the aerodynamic forces and moments that are created on the aircraft by the indicated turbulence velocities. To do this, a mathematical description of the frozen turbulence field,  $w_a(x)$ , is needed.

**Figure 173.2** Aircraft in a one-dimensional, frozen upwash field.



### Fourier Integral Representation

It is often assumed that  $w_a(x)$  is a sample function from an ergodic random process [Bendat and Piersol, 1966]. An important measure associated with such sample functions is the autocorrelation function, defined as

$$\phi_{w_a w_a}(\xi) = \lim_{X \rightarrow \infty} \frac{1}{2X} \int_{-X}^X w_a(x) w_a(x - \xi) dx \quad (173.2)$$

The autocorrelation function has a Fourier Integral or Fourier Transform representation:

$$\begin{aligned}\phi_{w_a w_a}(\xi) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi_{w_a w_a}(\Omega) e^{j\xi\Omega} d\Omega \\ \Phi_{w_a w_a}(\Omega) &= \int_{-\infty}^{\infty} \phi_{w_a w_a}(\xi) e^{-j\xi\Omega} d\xi\end{aligned}\quad (173.3)$$

Here,  $\xi$  represents spatial frequency with units of rad/ft or rad/m.  $\Phi_{w_a w_a}(\Omega)$  is referred to as the **power spectral density** of  $w_a(x)$ . Power spectral density representation constitutes one of the primary ways in which the characteristics of atmospheric turbulence have been measured and reported [Houbolt, 1973].

Two power spectral forms widely used in turbulence analyses are the von Karman and Dryden spectra [Etkin, 1972]. For example, the von Karman spectrum is given by:

$$\Phi_{w_a w_a}(\Omega) = 2\sigma_{w_a}^2 L \frac{1 + \frac{8}{3}(1.339L\Omega)^2}{[1 + (1.339L\Omega)^2]^{\frac{11}{6}}}\quad (173.4)$$

where  $L$  is referred to as the **turbulence scale length**. Depending upon altitude and turbulence intensity, scale lengths from 200 to 5000 ft have been measured [Houbolt, 1973]. Often in flight control system analysis, a spectrum simpler in form than those of Eq. (173.4) is used. For example,

$$\Phi_{w_a w_a}(\Omega) = 2\sigma_{w_a}^2 L_{w_a} \frac{1}{1 + (L_{w_a}\Omega)^2}\quad (173.5)$$

If one limits the spatial duration of  $w_a(x)$  to a large but finite value,  $X$ , then  $w_a(x)$  is itself amenable to representation by a Fourier Integral or Fourier Transform as

$$w_a(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} W_a(j\Omega) e^{j\Omega x} d\Omega\quad (173.6a)$$

$$W_a(j\Omega) = \int_{-\infty}^{\infty} w_a(x) e^{-j\Omega x} dx\quad (173.6b)$$

Equation (173.6a) can be used to show that  $w_a(x)$  may be envisioned as the sum of an infinite number of sinusoids, each of infinitesimal amplitude, with the frequency of each constituent sinusoid differing only infinitesimally from sinusoid to sinusoid [McRuer *et al.*, 1973].

## Aerodynamic Force/Moment Prediction

The problem of determining the aerodynamic forces and moments exerted on an aircraft as it moves through an upwash field can be simplified as follows. Compare the magnitudes of  $l_f$  and the spatial wavelengths ( $2\pi/\xi$ ) of the constituent infinitesimal sinusoids of  $w_a(x)$ . Consider the



case where  $w_a(x)$  has the majority of its power below a frequency  $\omega_0$ . If  $l_f \leq 8(2\pi/\omega_0)$ , the variation in  $w_a(x)$  over the length of the aircraft can be considered to be linear. This is shown in Fig. 173.2. The relationship  $l_f \leq 8(2\pi/\omega_0)$  is derived from the simple fact that a straight line can be considered an adequate approximation to a sinusoid if the length of the line is less than one-eighth of the period (wavelength) of the sinusoid.

Figure 173.2 suggests that the relative motion of the aircraft and upwash field (and the aerodynamic forces and moments being produced) is, at this instant, equivalent to that which would be in evidence if the aircraft were moving through quiescent air but with an instantaneous vertical velocity component

$$w(t) = -w_a(x_{cg}) \quad (173.7)$$

and an instantaneous angular pitching velocity (pitch rate) of

$$q(t) = \frac{dw_a(x_{cg})}{dx} = \frac{1}{U_0} \frac{dw_a(x_{cg})}{dt} \bigg|_{x_{cg}=U_0 t} \quad (173.8)$$

Here  $x_{cg}$  refers to the  $x$  coordinate of the aircraft center of gravity in the  $xyz$  axis system used to describe the frozen turbulence field.

The manner in which turbulence effects are included in the aircraft equations of motion can now be described. The force equations can be written in the familiar form:

$$\mathbf{F} = m \frac{d\mathbf{v}}{dt} \quad (173.9)$$

where  $\mathbf{F}$  represents the vector sum of the aerodynamic and propulsive forces acting on the aircraft of mass  $m$ , and  $\mathbf{v}$  represents the velocity vector of the aircraft's center of gravity.

For aircraft applications, Eq. (173.9) is typically linearized about a condition of steady, wings-level flight [Etkin *et al.*, 1972]. The components of  $\mathbf{F}$  in this equation are expressed as linear functions of the aircraft's linear and angular velocity perturbations and their derivatives. Thus, for example, referring to Fig. 173.1, the  $x_b$  and  $z_b$  body axis component of  $\mathbf{F}$  can often be simplified to

$$F_x = m[-g(\cos \theta_0)\Delta\theta + X_u\Delta u + X_w\Delta w + X_\delta\Delta\delta] \quad (173.10a)$$

$$F_z = m[-g(\sin \theta_0)\Delta\theta + Z_u\Delta u + Z_{\dot{w}}\Delta\dot{w} + Z_w\Delta w + Z_q\Delta q + Z_\delta\Delta\delta] \quad (173.10b)$$

where  $\Delta\theta$ ,  $\Delta u$ ,  $\Delta w$ , and  $\Delta\delta$  represent, respectively, perturbations in the pitch attitude, the components of  $\mathbf{v}$  in the aircraft's  $x_b$  and  $z_b$  body axes, and control/propulsive inputs.  $X$  and  $Z$  represent the components of the aerodynamic and propulsive forces in the  $x_b$  and  $z_b$  body axis directions. The quantities  $X_u$ ,  $X_w$ , etc., are referred to as mass normalized stability derivatives [McRuer *et al.*, 1973]. Equation (173.10) can now be modified to include the effects of the

one-dimensional upwash field by simply replacing  $\Delta w$  with  $(\Delta w - w_a)$  and  $\Delta q$  with  $[\Delta q + (\dot{w}_a/U_0)]$ . The result will be the addition of a disturbance term to Eqs. (173.10a) and (173.10b).

### 173.3 An Example

It is useful at this juncture to consider a simple flight control example utilizing the modeling procedure outlined in the previous paragraphs. Consider a small business jet flying at an equilibrium velocity of 677 ft/s (Mach No. = 0.7) at 40 000 ft and encountering a one-dimensional upwash field with a power spectral density described by Eq. (173.5). The turbulence intensity is  $\sigma_{w_a} = 10$  ft/s, and the scale length is  $L = 2000$  ft. The response variable of interest in this analysis is the normal acceleration at the aircraft's center of gravity,  $a_{z_{cg}}(t)$ . With the frozen turbulence assumption, the RMS value of  $a_{z_{cg}}(t)$  is given by

$$\sigma_{a_{z_{cg}}} = \sqrt{\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T a_{z_{cg}}^2(t) dt} = \sigma_{a_{z_{cg}}} = \sqrt{\frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi_{a_z a_z}(\omega) d\omega} \quad (173.11)$$

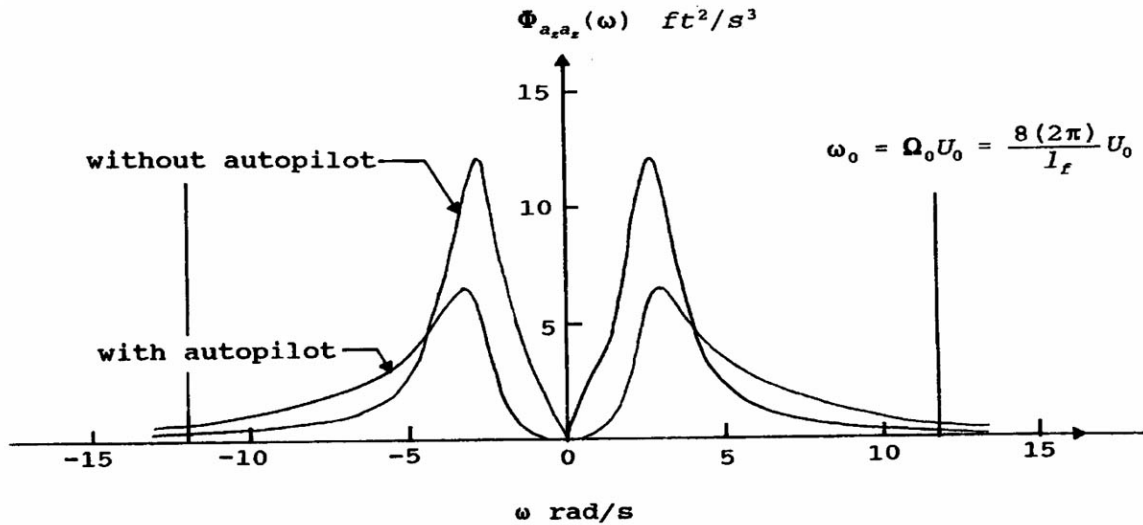
where

$$\Phi_{a_z a_z}(\omega) = \left| \frac{a_{z_{cg}}}{w_a}(j\omega) \right|^2 \Phi_{w_a w_a}(\omega) = \left| \frac{a_{z_{cg}}}{w_a}(j\omega) \right|^2 \left[ \frac{1}{U_0} \Phi_{w_a w_a}(\Omega) \right]_{\Omega = \frac{\omega}{U_0}} \quad (173.12)$$

and where  $a_{z_{cg}}/w_a(s)$  is obtained from the vehicle equations of motion, modified with turbulence terms as just described.

The aircraft possesses an autopilot designed to maintain a desired or equilibrium value of normal acceleration. Figure 173.3 shows  $\Phi_{a_z a_z}(\omega)$  for the aircraft, with and without the autopilot. The temporal frequency corresponding to a spatial wavelength eight times  $l_f$  is also shown. The RMS value of  $a_{z_{cg}}(t)$  for the aircraft with control system is approximately 2.8 ft/s<sup>2</sup>.

**Figure 173.3** Normal acceleration power spectral densities for aircraft in example problem.



## 173.4 Other Applications

---

The modeling technique which has been outlined is applicable to more complex turbulence fields than the simple upwash field discussed here. Extension to two-dimensional fields is, of course, possible. In addition, the effects of atmospheric movement other than random turbulence, such as flight through microbursts, is obviously possible [Psiaki and Stengel, 1985]. This type of low-altitude meteorological phenomenon produces large wind shears and can be particularly dangerous to an aircraft in takeoff and landing [Wingrove and Bach, 1989].

### Defining Terms

**Frozen turbulence field (Taylor's hypothesis):** A hypothesis that considers the time and spatial variation of the atmospheric velocities of a turbulence field as a spatial variation only.

**Mean wind:** A constant atmospheric velocity, independent of location and time.

**Power spectral density:** A means of describing the distribution, with frequency, of the power in a signal.

**Root mean square (RMS) value:** The square root of the integral of a squared function, divided by the integration interval.

**Turbulence (gusts):** The rapid, random variations of atmospheric velocities at a point in a space.

**Turbulence scale length:** A parameter appearing in the power spectral density representation of a frozen turbulence field. It is related reciprocally to the break frequency of the power spectral density (i.e., the larger the characteristic length, the lower the spatial or temporal break frequency).

**Wind shear:** The variation in atmospheric velocity as a function of location, often associated with microbursts.

### References

- Bendat, J. S. and Piersol, A. G. 1966. *Measurement and Analysis of Random Data*, 1st ed. John Wiley & Sons, New York.
- Etkin, B. 1972. *Dynamics of Atmospheric Flight*, 1st ed. John Wiley & Sons, New York.
- Houbolt, J. C. 1973. Atmospheric Turbulence. *AIAA J.* 11(4):421–437.
- McRuer, D. T., Ashkenas, I., and Graham, D. 1973. *Aircraft Dynamics and Automatic Control*, 1st ed. Princeton Univ. Press, Princeton, NJ.
- Psiaki, M. L. and Stengel, R. G. 1985. Analysis of Aircraft Control Strategies for Microburst Encounter. *J. Guid. Cont. Dyn.* 8(5):553–559.
- Wingrove, R. C. and Bach, R. E. 1989. Severe Winds in the Dallas/Ft. Worth Microburst Measured from Two Aircraft. *J. Aircraft* 26(3):221–224.

### Further Information

A thorough overview of the topic of turbulence and flight is presented in the 1981 Wright Brothers Lectureship in Aeronautics—Etkin, B. 1981. Turbulent Wind and Its Effect on Flight. *J. Aircraft*. 18(5):327–345.

Agarwal, R. K. "Computational Fluid Dynamics"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Computational Fluid Dynamics

---

174.1 Geometry Modeling and Grid Generation

174.2 Flow Simulation Algorithms

174.3 Turbulence Modeling

174.4 Flow Simulation Examples

174.5 Future Directions and Challenges

**Ramesh K. Agarwal**

*Wichita State University*

In the past two decades, enormous progress has been made in computational methods for the study of fluid flow phenomena in a broad variety of scientific and engineering disciplines. Examples include aircraft, ship, and automobile design; oceanography; meteorology; and astrophysics. The focus of this chapter is, however, on computational aerodynamics, which is rapidly emerging as a crucial enabling technology for the analysis and design of flight vehicles. Computational simulations, coupled with targeted experimental testing, have proven to be a cost-effective way of developing flight vehicles. Such combined evaluations are performed to ensure that the final product will efficiently meet target performance characteristics, using few design cycles. On the other hand, for some flight regimes, such as hypersonic flows at high altitude, experimental testing may not be feasible, and in such cases even greater reliance is placed on computational simulations.

In aeronautical applications, the computational analysis of the aerodynamic performance of a flight vehicle—be it an aircraft, a helicopter, or a launch vehicle—is a multistep process. First, a geometric description of the configuration is obtained in discretized form. Second, a grid is generated around the object to provide a finite set of points over which the flow field solution is calculated. The flow is then solved, and enormous quantities of data for pressure, temperature, and velocity variables are processed to obtain the aerodynamic quantities of interest—namely, the lift, drag, moment coefficients, and other parameters required to assess the flight vehicle performance.

In this respect, recent developments in computational fluid dynamics primarily fall into two broad categories: (a) computational prediction of the flow field about increasingly complex configurations, and (b) the simulation of flow physics leading to a better understanding of such flow phenomena as transition and turbulence. This chapter covers only the first category: the computational tools employed in aerodynamic analysis and their application in analyzing the aerodynamic performance of complex configurations.

### 174.1 Geometry Modeling and Grid Generation

---

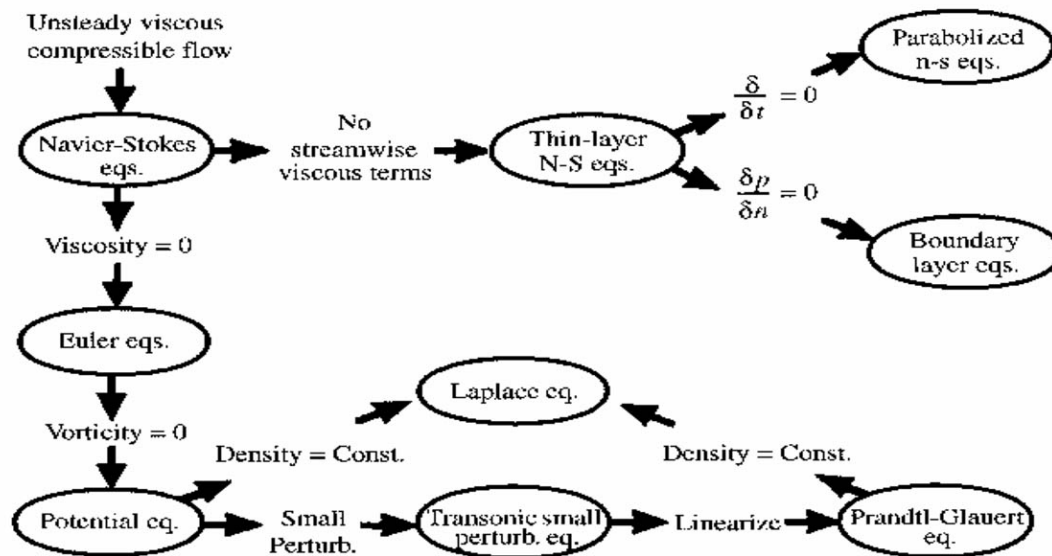
In the aerospace industry, the geometric description of the configuration in discretized form is routinely created using computer-aided-design (CAD) systems. However, the generation of a suitable three-dimensional grid about complete aircraft configurations remains a formidable task. In this respect, there have been significant developments in the past decade, with both the algebraic and partial differential equations–based techniques fairly well developed for the generation of

structured global grids about complex three-dimensional shapes. The recent thrust has been toward zonal grids and unstructured grid technologies since they are thought to be more effective for treating flows over complex three-dimensional objects. In addition, there have been major advances in adaptive grid and grid embedding techniques, for optimal distribution of grid points around an object in order to resolve sharp gradients and discontinuities, at the least cost possible. A review article by Steinbrenner and Anderson [1990] appropriately describes these developments.

## 174.2 Flow Simulation Algorithms

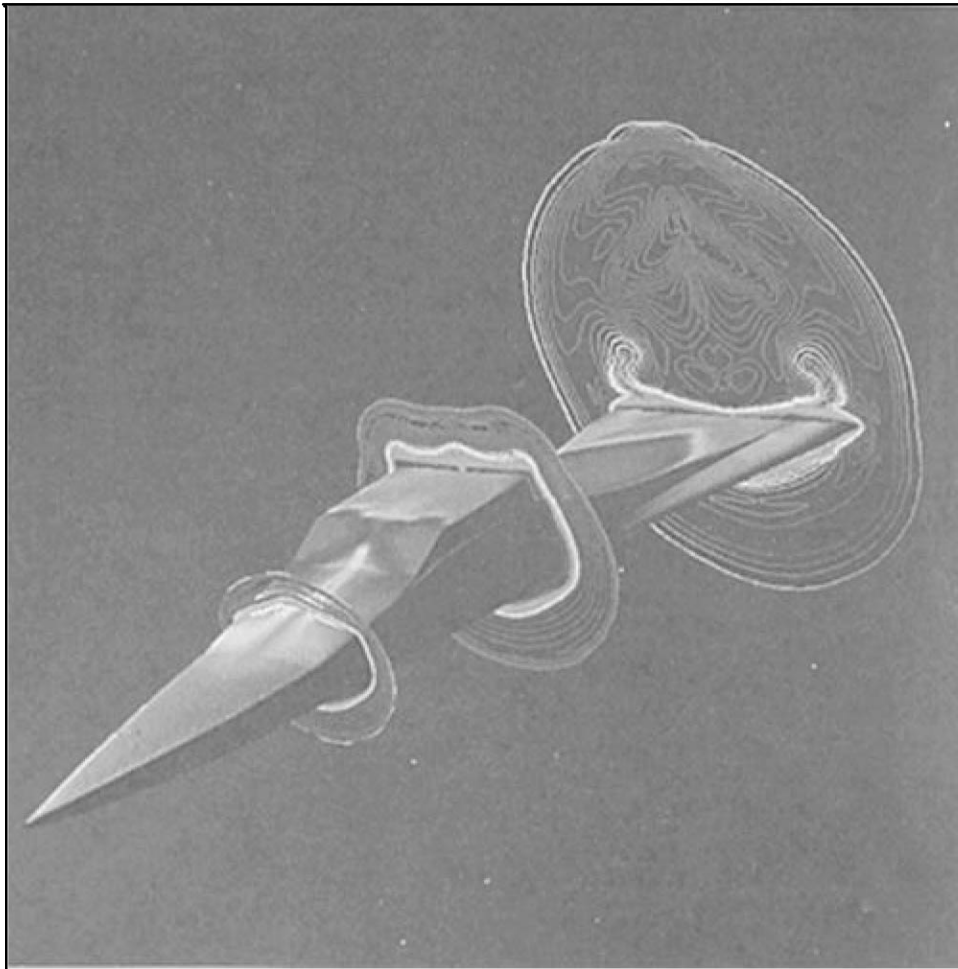
For flow field solutions, mathematical models vary in complexity from the simple Laplace equation to the unsteady compressible Navier-Stokes equations. Figure 174.1 describes the equations of fluid dynamics for mathematical models of varying complexity.

**Figure 174.1** Equations of fluid dynamics for mathematical models of varying complexity.



During the late 1960s, panel methods were developed for calculating inviscid subsonic flow about complex aerodynamic shapes by solving the Laplace equation. Also, boundary-layer solution techniques were developed to allow for viscous effects, primarily for attached flows. During the early 1970s, a major breakthrough was achieved by Murman and Cole [1971], who solved the mixed elliptic-hyperbolic transonic small-disturbance equation. Later, Jameson [1974] extended the Murman-Cole technique to the solution of the transonic full-potential equation. Such codes are routinely used in transport aircraft design and have proven to be an effective tool, when combined with boundary-layer corrections, for predicting the aerodynamic performance of a variety of flight vehicles in the transonic range.

Most of the recent progress in the development of numerical algorithms has been for the solution of the Euler and Navier-Stokes equations, especially the development of efficient and accurate shock-capturing algorithms. Central-difference algorithms with artificial viscosity (dissipation), characteristics-based upwind schemes, and total-variation-diminishing schemes have been developed for hyperbolic conservation laws. Strong discontinuities (shocks) are captured by these schemes with little oscillation and minimal dispersive and dissipative errors. It is beyond the scope of this chapter to provide details of these techniques, and the reader is referred, for example, to a survey paper by Jameson [1990].



### SCRAMJET EXHAUST

SCRAMJET engine exhaust is modeled in this supercomputer-generated image of an aerospace vehicle as part of the National Aero-Space Plane (NASP) program. Contour lines which ring the airplane indicate air density away from the surface at selected points. Shadings on the aircraft indicate areas of highest air density.

Researchers at NASA's Langley Research Center, Hampton, VA, are studying engine nozzle performance and exhaust effects on tail control surfaces. The image simulates wind tunnel test conditions at Mach 10 (approx. 6500 mph).

NASP was a joint program between the Department of Defense, NASA, and U.S. industry that made great strides in demonstrating technology for an aerospace vehicle that would take off horizontally like an airplane, accelerate to 17000-plus miles per hour and reach low earth orbit (generally less than 300 miles altitude). Once in orbit, such a vehicle could deploy payloads, pick up or repair satellites, or visit space station *Freedom* before returning to earth for a runway landing. (Photo courtesy of National Aeronautics and Space Administration.)



## 174.3 Turbulence Modeling

---

The currently intractable problem of accurately and practically modeling turbulence effects must be addressed before computational aerodynamics becomes a reliable routine prediction tool for flow situations where viscous effects are dominant. Examples of such situations include separated flows over control surfaces at high angles of attack and aerodynamic heating on high-speed vehicles. The computation of the effect of turbulence from first principles remains an elusive goal at the present time, even for simple two-dimensional shapes such as airfoils, and using maximum available computing power.

In current aerodynamic analyses, turbulence effects are accounted for by phenomenological models. The time-averaged form of the Navier-Stokes equations (known as the Reynolds-averaged Navier-Stokes equations) are solved. Time-averaging introduces the turbulent Reynolds stresses in the Navier-Stokes equations, and these are calculated by multiplying an eddy-viscosity coefficient with the strain-rate tensor. The phenomenological description of eddy viscosity is known as "turbulence modeling." At present, there is no universal turbulence model which works for all flow situations. Generally, turbulence models are developed by validating them against experimental data for simple flow situations, and are then used for calculation of complex flow fields. This approach introduces an element of uncertainty into the prediction of complex flows.

A variety of turbulence models of varying complexity have been developed over the last half-century. The underlying theoretical framework behind these models and their range of applicability are not given here, but an interested reader can find a comprehensive discussion in a review paper by Rubesin [1987].

## 174.4 Flow Simulation Examples

---

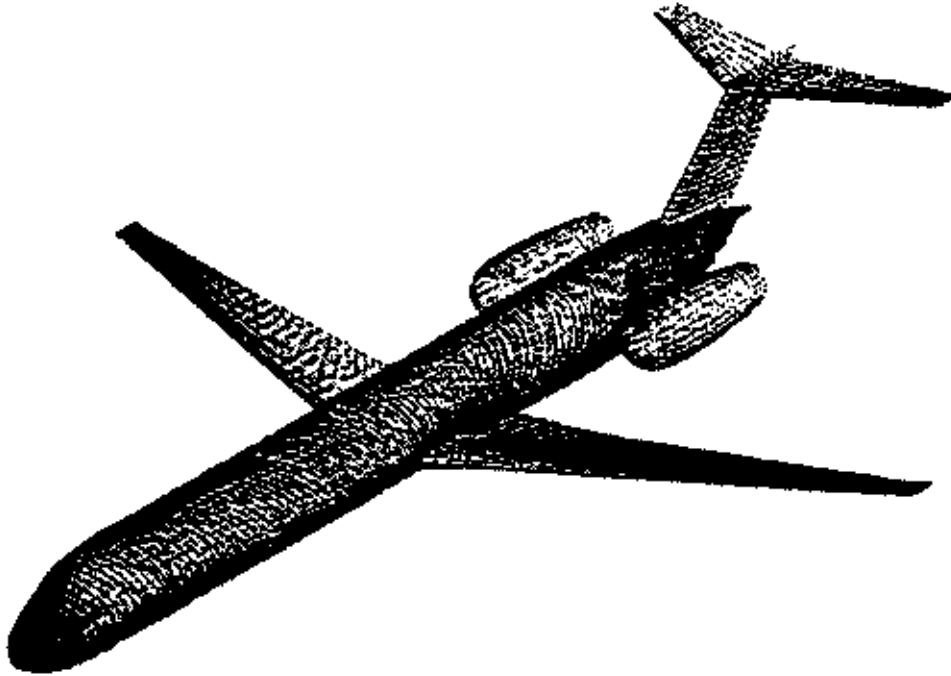
The selected flow simulation examples given here basically draw upon the author's own work and that of co-workers at the McDonnell Douglas Research Laboratories. The purpose of these examples is to illustrate the current state in computational aerodynamics for the prediction of aircraft, launch vehicle, and helicopter flow fields, from a manufacturer's point of view.

Figure 174.2 shows the surface grid on a complete twin-jet transport aircraft. Figure 174.3 shows the pressure distribution on the aircraft at a cruise Mach number of 0.76 and angle of attack of  $2^\circ$  obtained with the Euler code described by Deese and Agarwal [1988]. Figure 174.4 shows the comparison of the computed pressure distribution, at various spanwise locations of the wing, with experimental data. The inviscid Euler solution tends to predict shocks which are too strong and too far downstream on the wing upper surface. Addition of the viscous terms moves the shock location and strength toward better agreement with the experimental data, except in the region near the suction peak, where better grid resolution is needed to capture the high flow field gradients. Figure 174.5 shows the surface grid on a complete fighter aircraft with faired inlets. Figure 174.6 shows the pressure distribution on the aircraft at a Mach number of 0.90 and an angle of attack of  $4.84^\circ$ , obtained with an Euler code described in Deese and Agarwal [1988]. Figure 174.7 shows the comparison of the computed and experimental pressure distributions at various spanwise locations of the fighter wing. The Euler solution tends to predict a suction peak higher than that observed experimentally. Better resolution of the relatively small-radius fighter wing should improve the

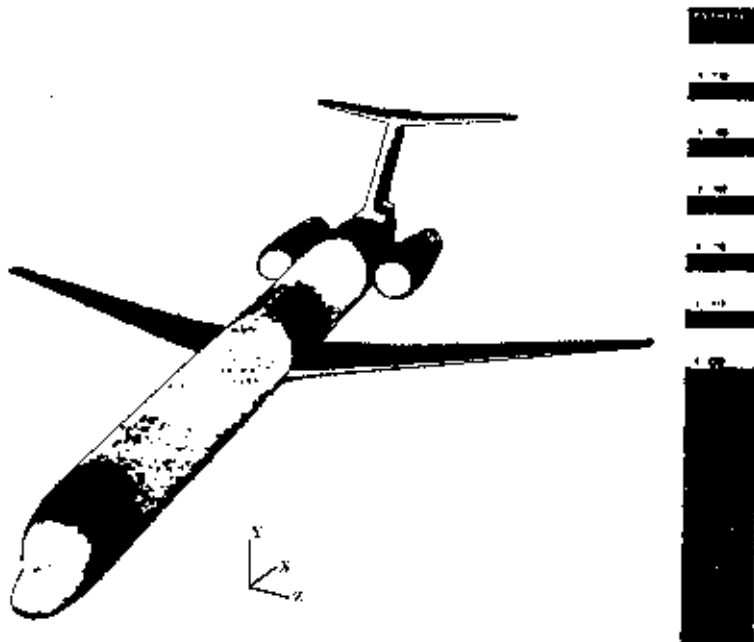


agreement with data. The shocks on the wing upper surface are predicted to be stronger and slightly downstream of the measured shocks, as is typical of inviscid solutions.

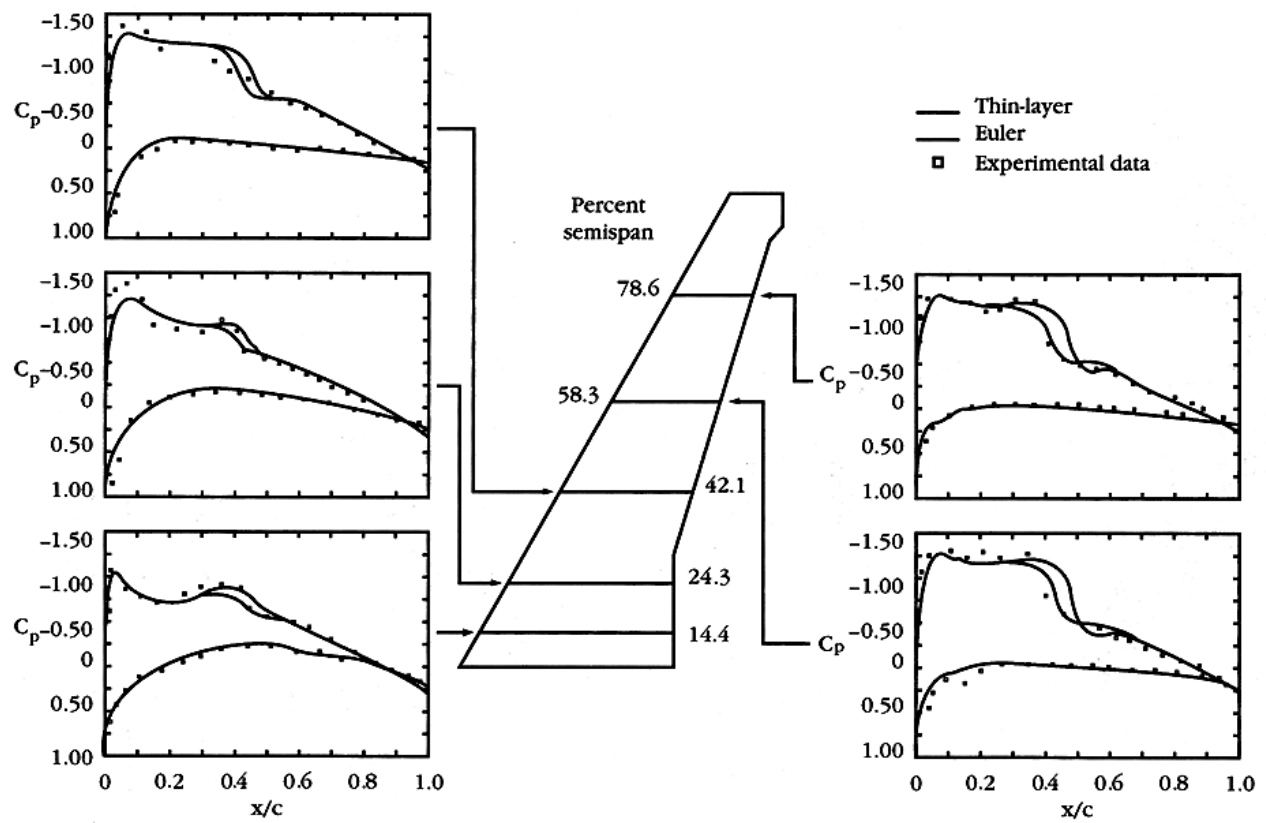
**Figure 174.2** Surface grid over a complete twin-jet aircraft.



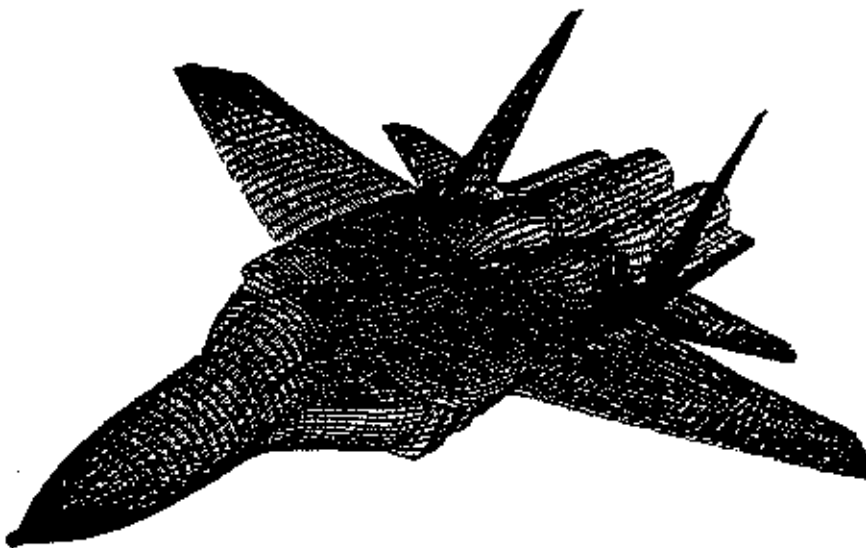
**Figure 174.3** Euler solution predictions of surface pressure on a generic twin-jet transport aircraft;  $M_\infty = 0.76$ ,  $\alpha = 2.0^\circ$ .



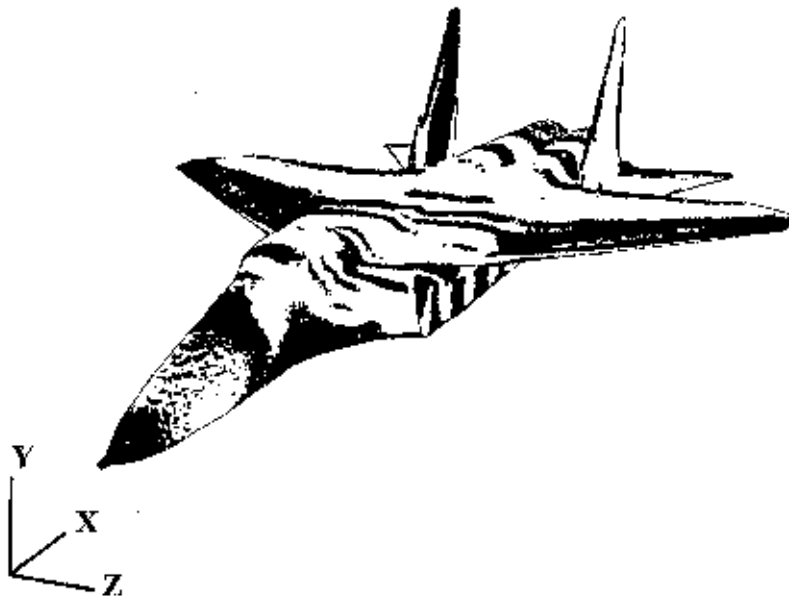
**Figure 174.4** Comparison of Euler and thin-layer Navier-Stokes pressure distribution with experimental data for typical transport wing-body at  $M_\infty = 0.76$ ,  $Re = 6.39 \times 10^6$ ,  $\alpha = 2.0^\circ$ ,  $160 \times 34 \times 42$  mesh.



**Figure 174.5** Surface grid on a generic fighter with faired inlets.



**Figure 174.6** Surface pressure distribution on the generic fighter, Mach 0.9,  $\alpha = 4.84^\circ$ .



**Figure 174.7** Surface pressure distribution on the wing of the generic fighter;  $M = 0.9$ ,  $\alpha = 4.84^\circ$ .

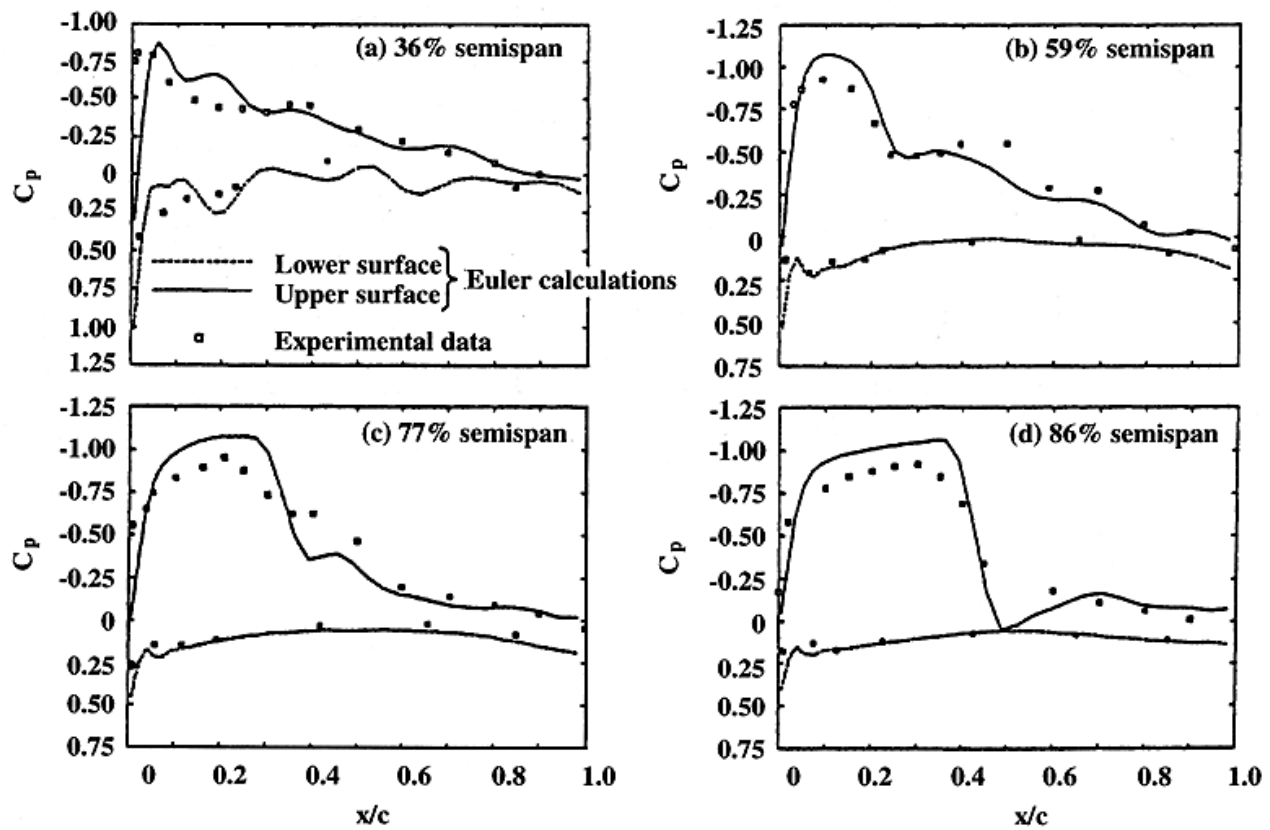
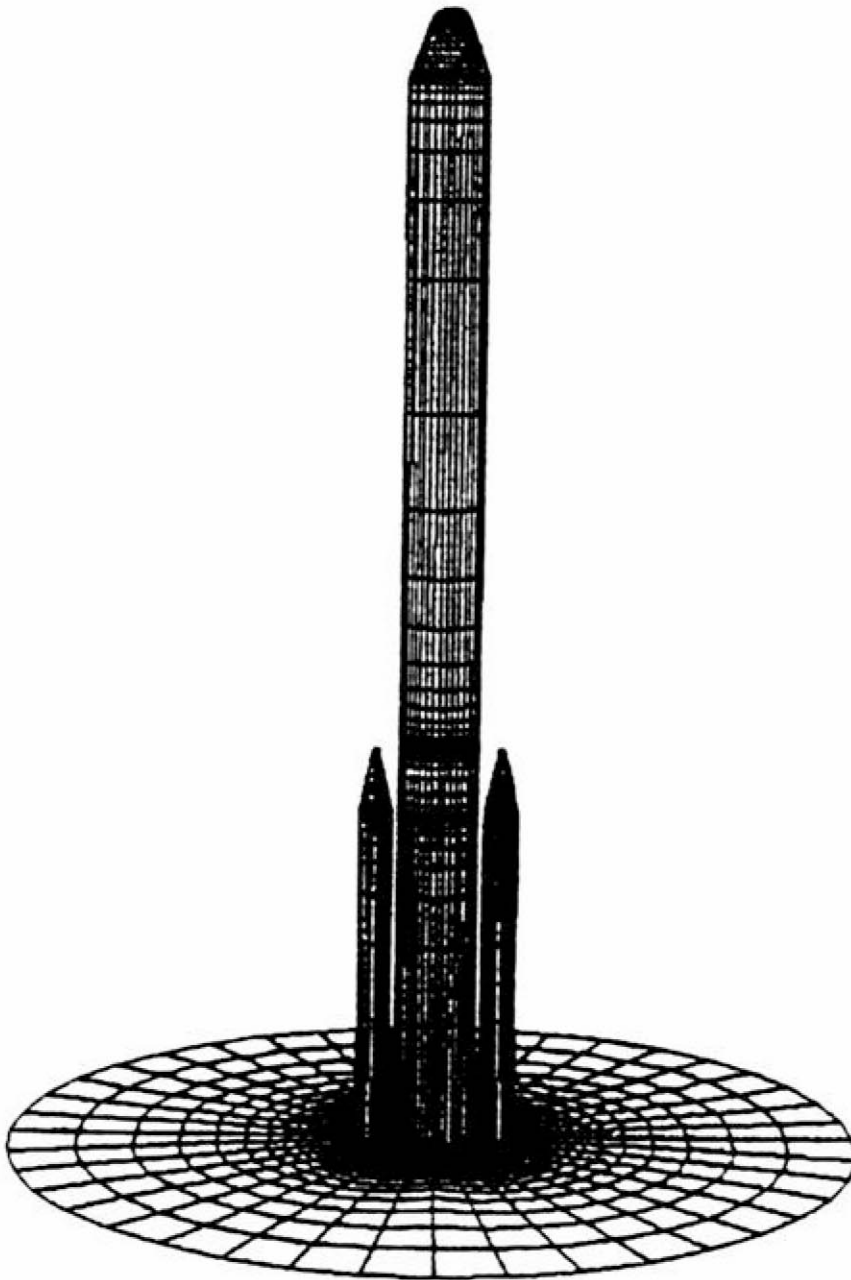


Figure 174.8 shows the surface grid on a launch vehicle with two boosters. Figure 174.9 shows the comparison of the computed pressure distribution on the main vehicle of a two-booster configuration with experimental data. The predictions are in good agreement with the data, except for the region just upstream of the booster nose. Inclusion of viscous effects should improve the agreement between the calculations and the experimental data.

**Figure 174.8** Surface grid over a launch vehicle with two boosters.



**Figure 174.9** Surface pressure distribution on the core of a launch vehicle with two boosters;  $M_\infty = 1.05$ ,  $\alpha = 0.0$ .

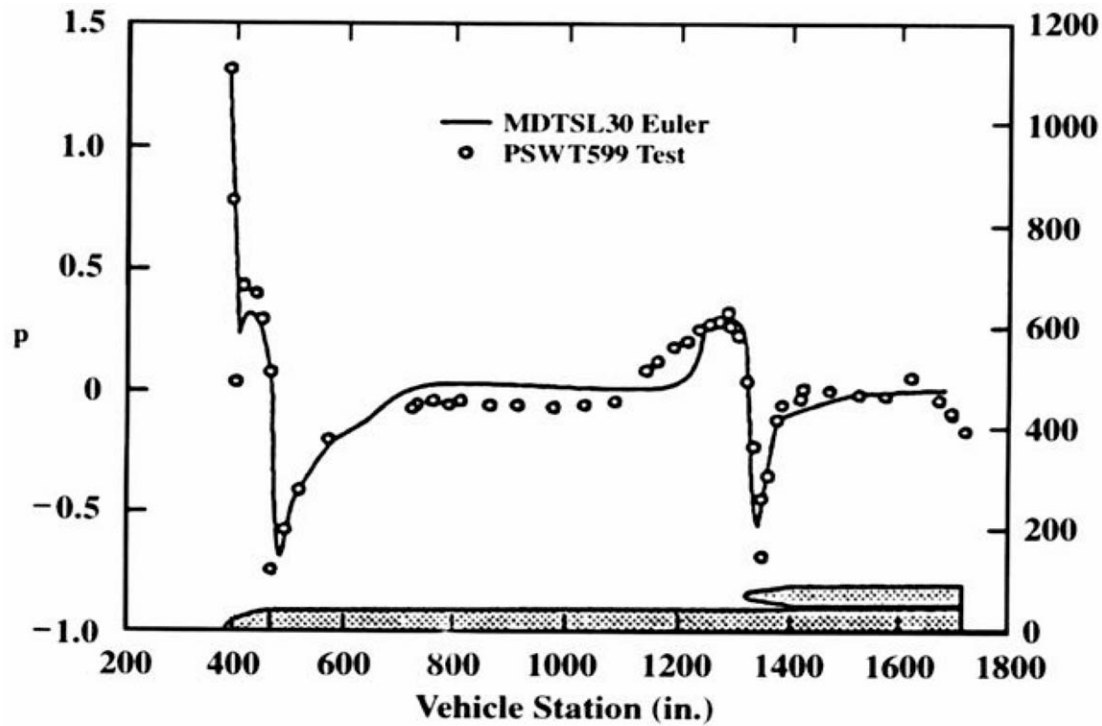
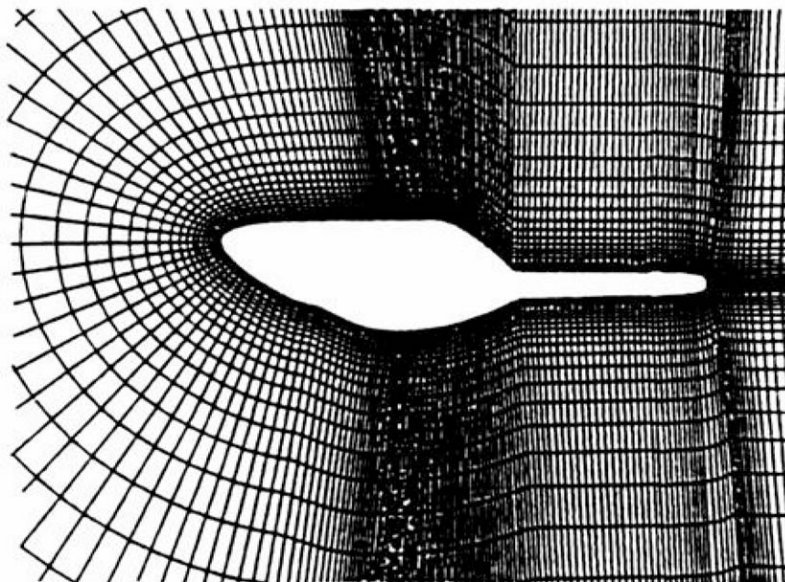
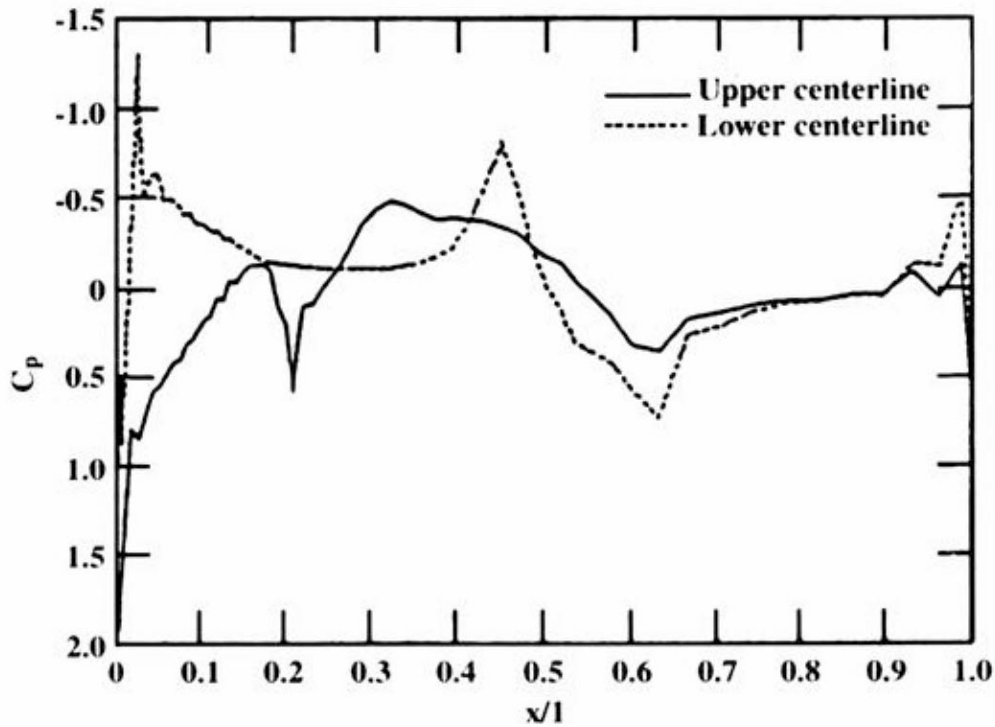


Figure 174.10 shows the grid distribution in the symmetry plane for a generic helicopter fuselage. Figure 174.11 shows the pressure distribution on an isolated helicopter fuselage obtained with the Navier-Stokes code described by Deese and Agarwal [1988]. Figure 174.12 shows the pressure distribution on an ONERA rotor blade in forward flight, obtained with the rotary-wing Euler code described by Agarwal and Deese [1987]. Figure 174.13 compares computed and experimental pressure distributions on the ONERA rotor in forward flight.

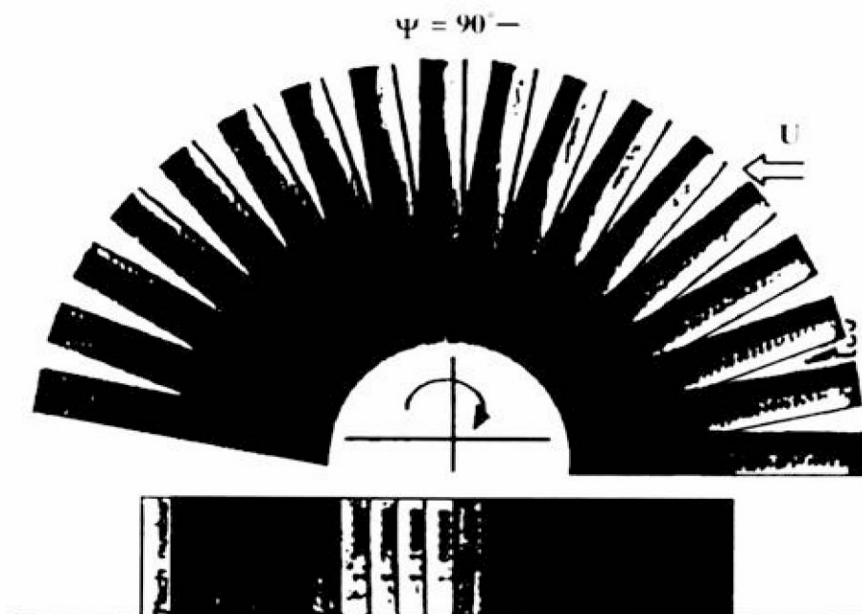
**Figure 174.10** Grid distribution in the symmetry plane for a generic helicopter fuselage.



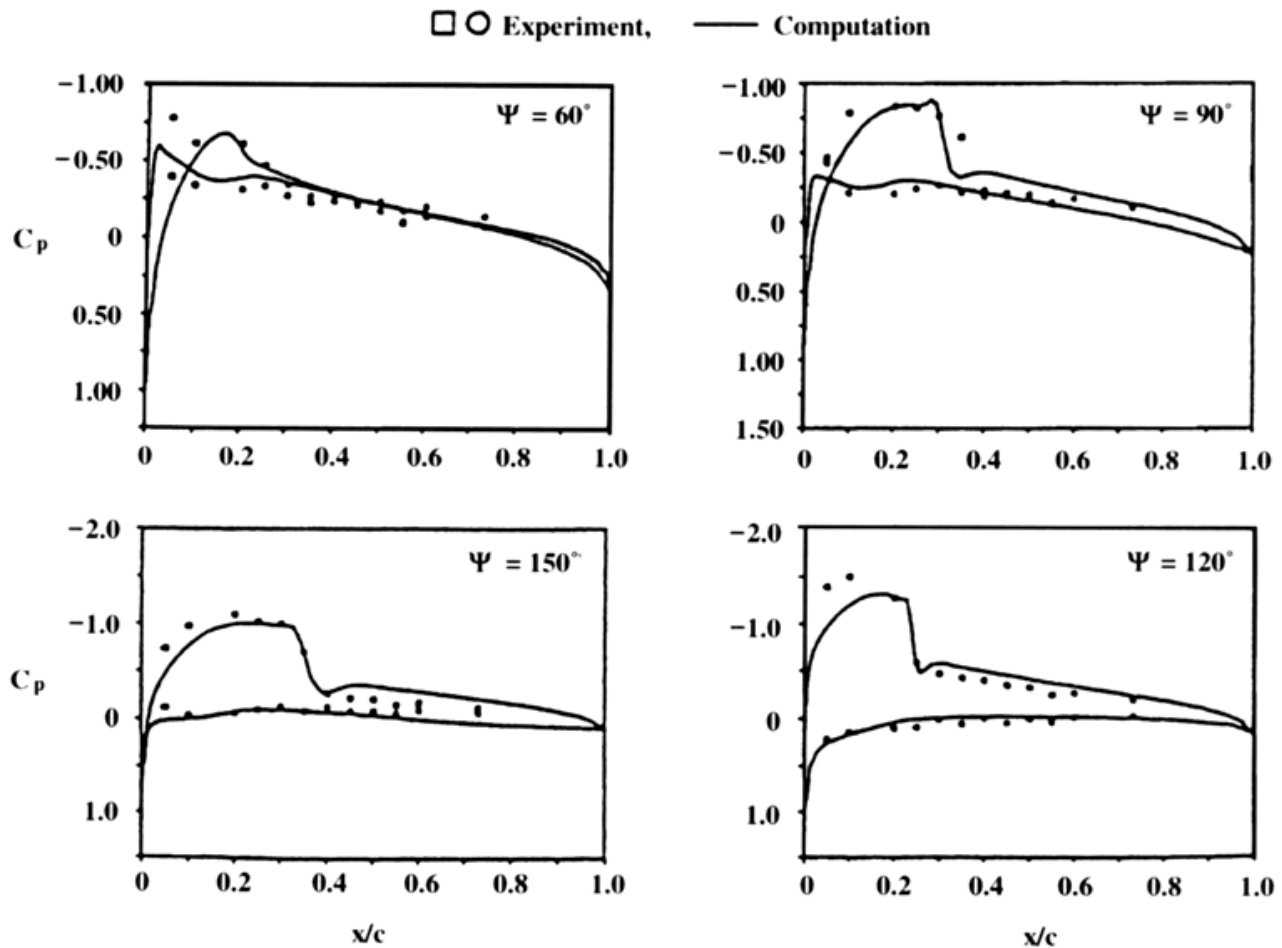
**Figure 174.11** Pressure distributions computed by the Navier-Stokes code on the symmetry plane of a generic fuselage configuration;  $M_\infty = 0.4$ ,  $\alpha = -5^\circ$ .



**Figure 174.12** Euler solutions for flow about an ONERA three-bladed rotor in forward flight;  $M_t = 0.629$ ,  $\mu = 0.388$ .



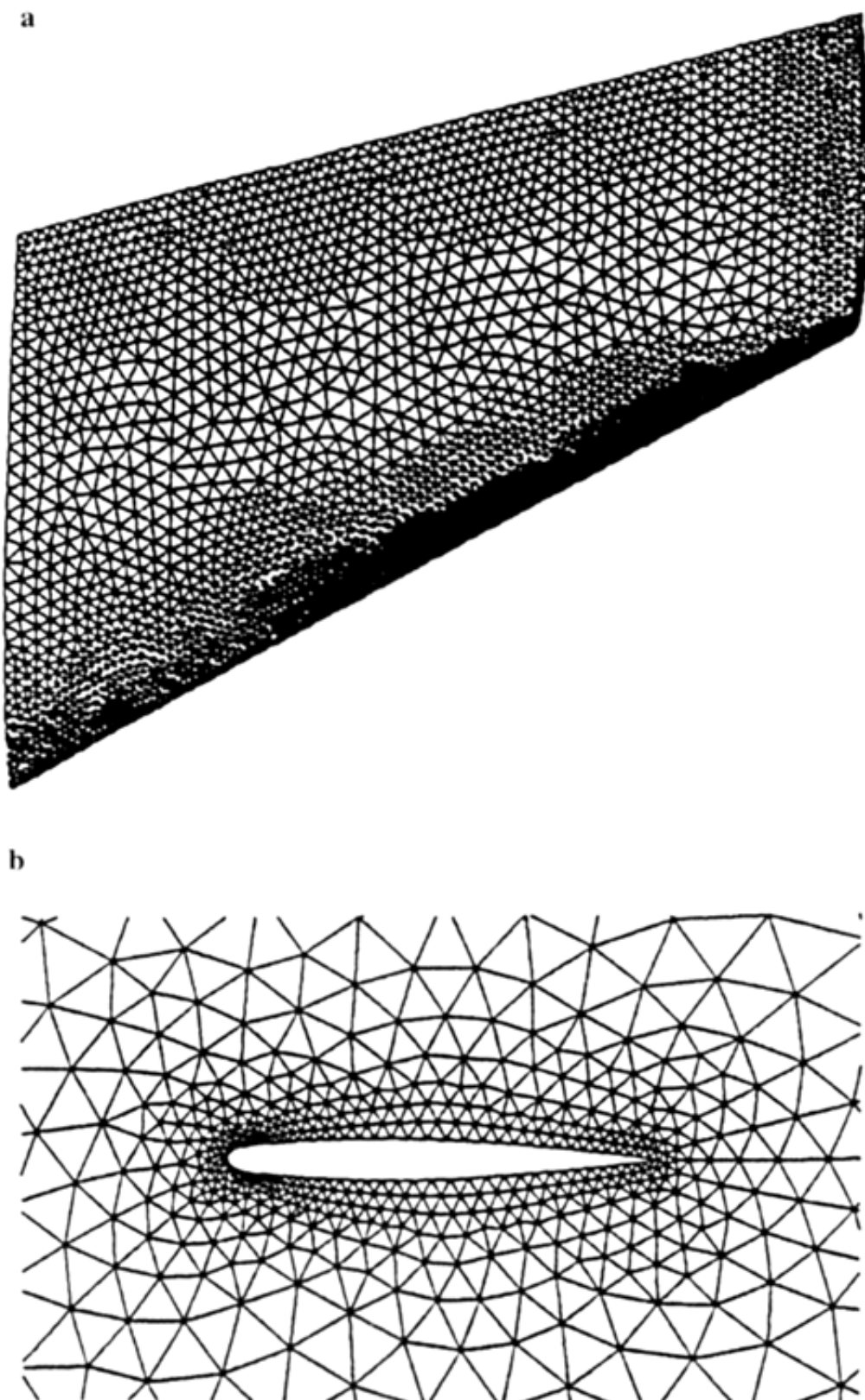
**Figure 174.13** Comparison of predicted and experimental rotor surface pressure distributions for the ONERA three-bladed rotor in forward flight;  $M_t = 0.629$ ,  $\mu = 0.388$ , 90% span location.



As a final example, [Fig. 174.14](#) shows the unstructured grid about an ONERA M6 wing. [Figure 174.15](#) shows the computed pressure distribution on the wing at various spanwise locations and its comparison with experimental data. These calculations were performed with the Navier-Stokes code described by Marcum and Agarwal [1990].

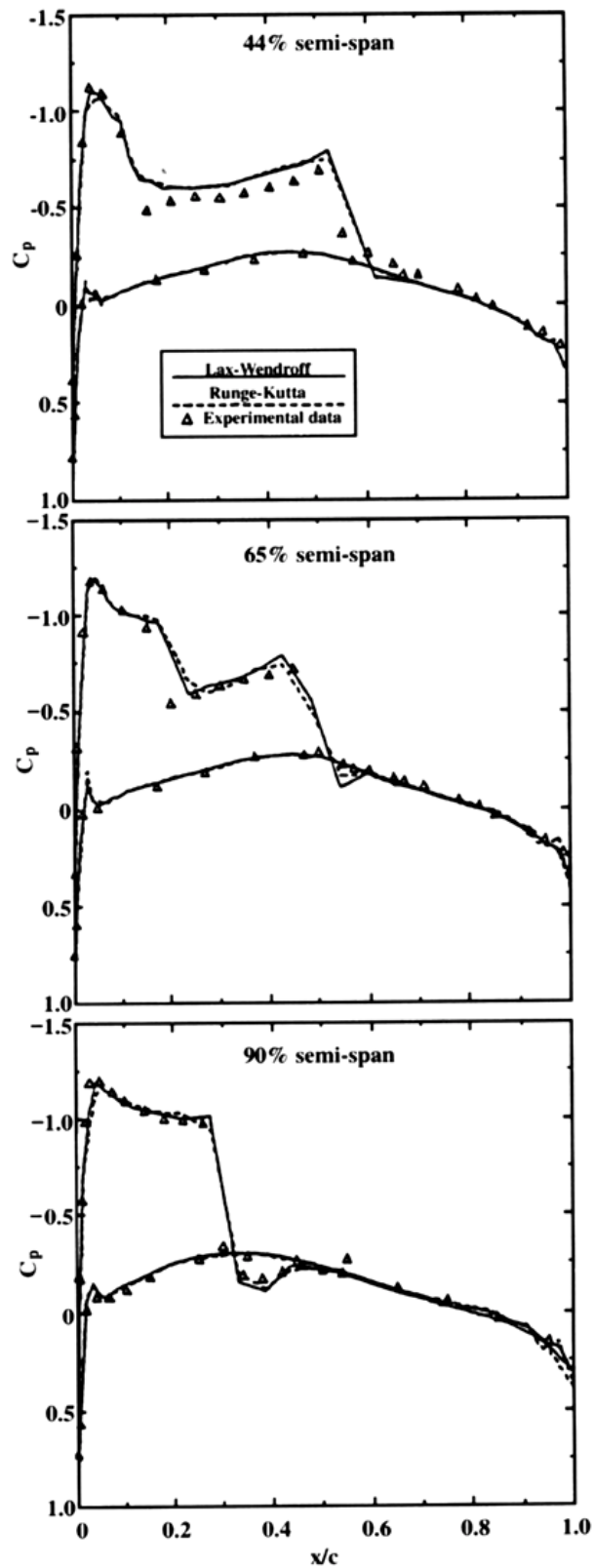


**Figure 174.14** Surface grids for ONERA M6 wing grid with 231507 elements and 42410 nodes. (a) Wing surface with 15279 faces and 7680 nodes. (b) Symmetry plane with 1525 faces and 813 nodes.





**Figure 174.15**  $C_p$  distribution for ONERA M6 wing at  $M_\infty = 0.84$ ,  $\alpha = 3.06^\circ$ , as predicted by the unstructured grid code.



Many examples of a similar kind exist in the aerospace literature, showing the power of computational aerodynamics in simulating flows about complex configurations.

## 174.5 Future Directions and Challenges

---

As we step into the second half of the 1990s, we can look back on the progress in computational aerodynamics during the 1970s, 1980s, and early 1990s with great satisfaction. At present, we can generate three-dimensional grids about complete vehicle configurations and employ efficient and accurate numerical algorithms for the solution of the Euler and Navier-Stokes equations. Inviscid flow can be computed quite accurately over the entire Mach number range.

However, confidence in the predictive capability of computational codes for viscous-dominated flows remains low because of turbulence model limitations and insufficient computing power. Furthermore, although significant advances have taken place in the analysis of flow fields, their impact on the development of direct design methods has been limited. The development of inverse design techniques has lagged far behind that of analysis tools. In the future, a strong emphasis is needed on the development of such design methods.

The next computational challenge will be in developing methods and techniques for multidisciplinary design optimization, that is, to include several disciplines simultaneously in the design optimization process, rather than in a sequential manner. A configuration could be optimized taking into account aerodynamic, structural, and signature constraints simultaneously.

Finally, two of the emerging computer-related technologies, artificial intelligence and massively parallel technology, are likely to have major impacts on flight vehicle design. Expert system technology will have a substantial impact on automating the design process, thus reducing the cost and time of a design cycle. Massively parallel technology will make real-time interactive analysis possible, opening up great opportunities for real-time design modifications and improvements in product quality, at a substantially reduced cost.

## References

- Agarwal, R. K. and Deese, J. E. 1987. Euler calculations for the flow field of a helicopter rotor in hover. *J. Aircr.* 24:231–238.
- Deese, J. E. and Agarwal, R. K. 1988. Navier-Stokes calculations of transonic flow about wing/body configurations. *J. Aircr.* 25:1106–1112.
- Jameson, A. 1974. Interactive solution of transonic flows over airfoils and wings, including flows at Mach 1. *Commun. Pure Appl. Math.* 27:283–309.
- Jameson, A. 1990. Full potential, Euler and Navier-Stokes schemes. In *Applied Computational Aerodynamics: Progress in Astronautics and Aeronautics*, ed. A. R. Seebass, p. 39–88. American Institute of Aeronautics and Astronautics, Washington, DC.
- Marcum, D. L. and Agarwal, R. K. 1990. Finite element Navier-Stokes solver for unstructured grids. *AIAA J.* 30(3):648–654.
- Murman, E. M. and Cole, J. D. 1971. Calculation of plane steady transonic flows. *AIAA J.* 9:114–121.

- Rubesin, M. W. 1987. *Turbulence Modeling, Supercomputing in Aerospace*. NASA CP-254.
- Steinbrenner, J. P. and Anderson, D. A. 1990. Grid generation methodology. In *Applied Computational Aerodynamics: Progress in Astronautics and Aeronautics*, ed. A. R. Seebass, p. 91–130. American Institute of Aeronautics and Astronautics, Washington, DC.

June, R. R. "Aerospace Materials"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

175.1 System Requirements and Materials Selection

175.2 Design Considerations

175.3 Nonstructural Materials

175.4 The Future

**Reid R. June**

*The Boeing Company*

Aerospace structural materials include aluminum, titanium, steel- and nickel-based alloys, and fiber reinforced plastics. Nonstructural materials include paint, rubber, transparencies, sealants, lubricants, wiring and electronics, and systems fluids. All aerospace materials must satisfy stringent performance requirements at minimum weight and cost.

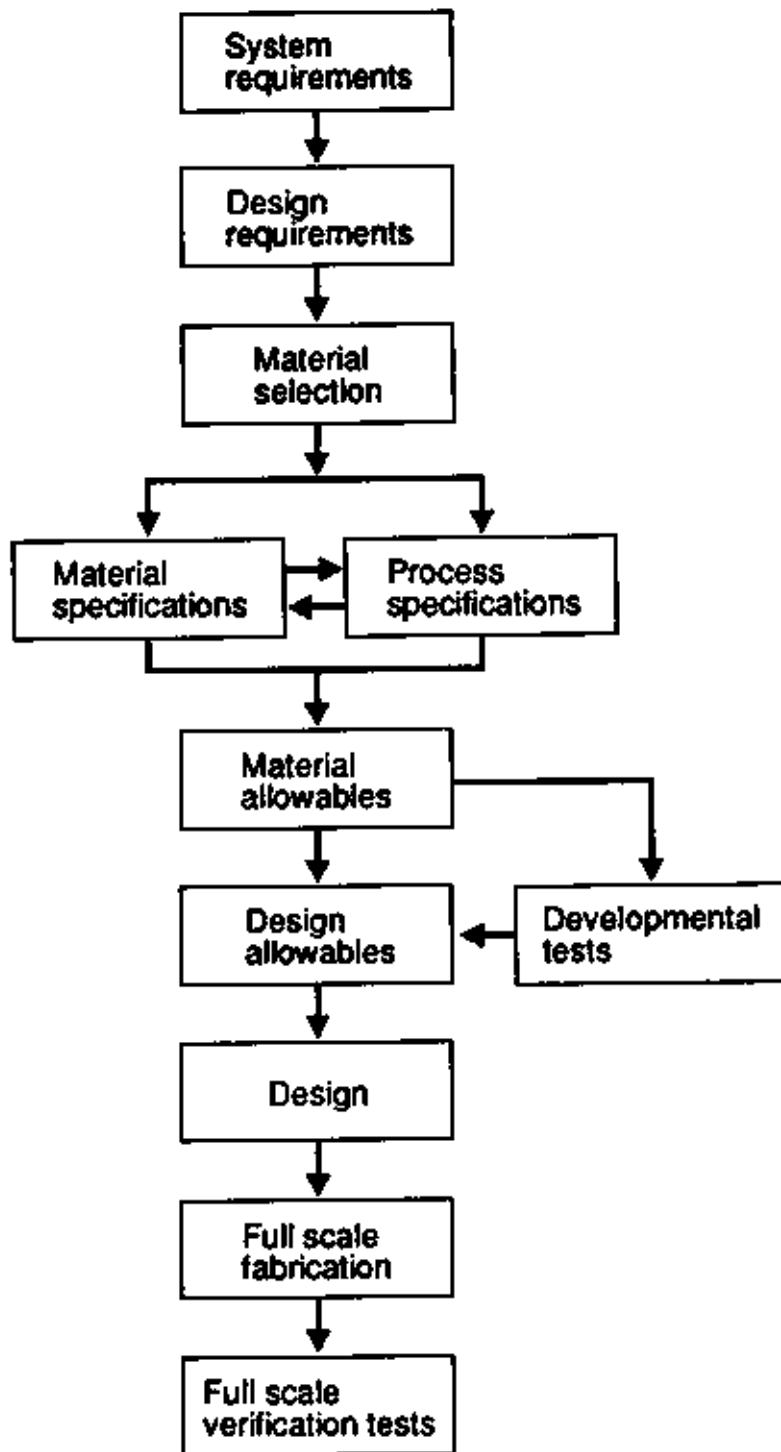
## **175.1 System Requirements and Materials Selection**

---

Aerospace materials selection begins with a thorough understanding of the design requirements for the system to which the materials are to be applied. Design requirements include loads, criteria, and operating environment. Crichlow [1969] presents an excellent approach for integrating structural requirements and material selection.

Figure 175.1 shows that system and design requirements drive materials selection. It is essential that all critical design requirements be understood as early in the design process as possible to enable the materials job to be done correctly the first time, thereby avoiding costly redesign and rework. Product development teams, where all disciplines are concurrently involved in design evolution, and expanded use of computer-aided design and manufacturing have helped to achieve this objective.

**Figure 175.1** Flow of system requirements through verification testing.



When the designers understand the materials requirements, the selection process can proceed with off-the-shelf materials or the development of new materials, if required. Most new systems include a combination of proven materials and new materials and processes developed to (1) improve performance, producibility, or both, (2) to satisfy unique requirements of the system, or

(3) to solve problems with earlier materials, such as excessive maintenance costs. Customer input is essential in this regard. In all cases, comprehensive materials and process specifications are written, usually in conjunction with material suppliers, to ensure that maximum advantage is taken of existing experience and knowledge.

Next, needed **material allowables** data are developed. For structural materials, these data typically include strength, modulus of elasticity, effects of temperature and environment, fatigue, toughness, corrosion resistance, and any other special data required for design of the system. The data are obtained from coupon tests. Design allowables are derived from material allowables and provide allowance for scale-up effects and uncertainties in materials, analysis, loads, and other system requirements. Design, fabrication, and test of full-scale test articles verify that design requirements are satisfied. If testing shows that requirements are exceeded, the design will accommodate growth in system performance parameters. If premature failure occurs, root causes must be identified and corrected.

Some materials have proven to be unsatisfactory in particular service applications. This condition is often traceable to inadequate development and testing prior to production, to discovery of material behavior that became evident only after several years of exposure to the service environment, or to systems operational conditions that were not identified in the original requirements. These failures underscore the importance of having accurate and complete requirements and obtaining the critical information to ensure that they are satisfied prior to entering production.

Aerospace system performance requirements demand the highest standards of quality from all elements of the system, including materials and related processes. Process standards, including **control limits**, must be established and followed during development, production, and operation of the system. Material suppliers, system manufacturers, and system operators all have responsibilities for maintaining process control during the system life cycle.

## 175.2 Design Considerations

---

Design concept, design criteria, material characteristics and processing, and operator maintenance all play key roles in ensuring structural integrity over the service life of the system. Metallic materials, for example, must have crack growth characteristics that ensure that normal inspections will reveal any crack before it can grow to a size that will jeopardize structural integrity. Advanced composite structures are designed so that damage will be detectable before structural integrity is compromised.

Minimum weight is essential in all aerospace systems. Materials with low specific weights or densities are key to lightweight structures. For this reason, aluminum alloys with densities of about 0.1 lb/in.<sup>3</sup> are widely used, comprising, for example, 60 to 70% of the structural weight of modern commercial aircraft. Lighter still are **advanced composite materials** with densities of about 0.06 lb/in.<sup>3</sup>. These materials offer 20 to 25% weight savings compared to aluminum structures and have excellent corrosion and fatigue characteristics. Mechanical property ranges for some widely used aerospace metals and composites are shown in [Tables 175.1](#) and [175.2](#), respectively. It should be noted that composite materials are continuing to evolve, especially for high-temperature applications. The ranges shown in [Table 175.2](#) are expected to change significantly over time.

**Table 175.1** Range of Mechanical Properties for Typical Aerospace Metallic Materials

Material	Density (pci)	Range of mechanical properties				Maximum Service Temp. (deg F)	Selection Drivers	Applications
		Modulus of Elasticity (msi)	Ultimate Tensile Strength (ksi)	Compressive Yield Strength (ksi)	Fracture Toughness (ksi √in)			
Aluminum								
2000-series	0.100–0.103	10.2–10.7	41–70	23–50	20–40	200–400	Fatigue, toughness	Lower wing structure, fuselage skins
7000-series	0.101–0.102	10.1–10.3	54–88	43–85	25	200	Strength	Upper wing structure, fittings, and structure
6000-series	0.098	9.9	26–42	12–36	2	200	Corrosion resistance, weldability	Welded ducts
Steel								
4000-series	0.283	29.0	125–280	109–244	40–100	500–900	Strength	Landing gear, high-strength parts
9Ni-4Co-.3C	0.284	28.5	220	209	80	900	Combined strength and toughness	Fracture-critical parts and fittings
CRES								
300-series	0.286	26.0–29.0	70–175	23–83	2	700–1000	Corrosion resistance, weldability	Sheet metal, welded parts, castings
PH-series	0.279–0.283	28.3–28.5	125–200	95–200	75	600–800	Corrosion resistance, strength, and toughness	Fittings
Titanium								
Annealed or cold-worked CP, 6-4	0.160–0.163	15.0–16.0	80–135	70–130	≤ 70	500–750	Weldability, corrosion resistance, lightweight	Sheet metal parts, welded parts, hydraulic tubes, and structural parts
Heat-treated 6-4, 10-2-3	0.160–0.174	15.0–16.0	130–173	126–168	40	500–750	Strength, corrosion resistance, lightweight	Landing gear fittings, structural parts, fasteners, and standard parts
Nickel								
Annealed 625	0.305	29.8	120	53	2	1800	Temperature resistance, weldability	Welded ducts
Heat-treated 718	0.297	29.4	180	158	80	1800	Temperature resistance, strength, toughness	Fittings, fasteners

<sup>1</sup> Ranges shown are for available material forms, including alloy and heat treatment. Values are for reference only, and are not to be used for design.

<sup>2</sup> Materials not used where fracture toughness is a consideration.

**Table 175.2** Range of Mechanical Properties for Typical Aerospace Composite Materials

Material	Range of Mechanical Properties <sup>1</sup>					Maximum Service Temp. (deg F)	Selection Drivers	Applications
	Density (pci)	Modulus of Elasticity (msi)	Design Tensile Strength (ksi)	Hot-Wet Compressive Strength (ksi)				
	<sup>2</sup>	<sup>3</sup>	<sup>3</sup>	<sup>3</sup>				
Fiberglass; polymeric matrix	0.08	3–6	20–100	15–75	150–500	150–500	Low cost, corrosion resistance, toughness	Fairings, doors
Polymeric fiber; polymeric matrix	0.05	4–11	12–50	Low	150–500	150–500	Low density, energy absorption	Interiors, fairings, damage containment
Carbon fiber; polymeric matrix	0.06	6–30	25–100	20–80	150–500	150–500	Low density, strength, corrosion, fatigue, stiffness, low thermal expansion	Primary and secondary aerospace structures
Boron fiber; polymeric matrix	0.07	15–30	50–100	50–100	150–500	150–500	High compression strength, high stiffness	Compression dominated structure
Boron, carbon, or silicon carbide fiber; aluminum or titanium matrix	0.08–0.12	10–30	50–100	75–125	800–1000	800–1000	High temperature	Propulsion, space structures
Carbon fiber; carbon matrix	0.07	2	15–30	N/A	5000	5000	Very high temperature	Aircraft brakes, rocket nozzles, re-entry heat shields

<sup>1</sup> Reference only. Not to be used for design.

<sup>2</sup> Composite density depends on volume fraction and density of constituents.

<sup>3</sup> Composite design strengths and modulus depend on constituent property values, fiber volume fraction, and fiber orientation, tailored to directional load requirements, and must account for the notch-sensitivity of most composite materials.



Most material choices involve engineering judgment in seeking the optimum combination of material properties for particular applications and requirements. For example, designers must consider the trade between strength and toughness when selecting alloys and their heat treatments. For minimum weight, high specific strength is desired. For damage tolerance, high toughness is needed. In general, however, as strength increases, toughness decreases. The best engineering solution is a material that is strong enough and tough enough, and provides the minimum weight design at lowest cost. Depending on the operating environment and the system and design requirements, one property or another may be critical, and the material choice is made accordingly. [Table 175.3](#) shows structural material weight percentages for several current airplanes.

**Table 175.3** Structural Material Weight Percentage for Some Current Airplanes

Material	Airplane and Manufacturer			
	McDonnell Douglas AV-8B	McDonnell Douglas F/A-18	Boeing 747	Boeing 777
Aluminum	38	29	70	59
Steel	12	14	11	13
Titanium	8	15	4	6
Composites	26	22	3	9
Other	16	20	12	13

Testing plays an important role in design evolution. It is very important that the correct tests be done in the correct manner to avoid costly mistakes that can result when improper materials are chosen for an application. The scale of testing to interrogate materials and structural capabilities ranges from small, relatively inexpensive material coupon tests to large, relatively expensive tests of full-scale articles, with intermediate levels of element, subcomponent, and component tests. The goal of a successful materials test program is to obtain the required data in the least expensive way that will ensure that, while in service, the materials will perform in the way intended, and will neither experience premature failure nor require excessive maintenance. Testing of structures containing both composite and metallic materials presents special challenges because of differing critical design conditions and strain responses to loading.

The American Society for Testing and Materials (ASTM) has developed standard test specimen geometries and test techniques for many aerospace (as well as other) materials. It is desirable to use standard test methods wherever possible. Underlying principles of materials testing include assurances that test techniques do not influence test results, that test data are reproducible, and that the tests are in fact providing reliable information that can be used in making sound engineering design decisions.

## 175.3 Nonstructural Materials

As with structural materials, nonstructural aerospace materials must satisfy high standards for performance and reliability. All are qualified through analysis and test before commitment to production. For example, interior materials for commercial aircraft must satisfy federal requirements for low smoke and toxic emission in a fire. In addition, these materials must be durable, decorative, and cleanable. Windshields must meet standards for birdstrikes. Compliance is demonstrated by analysis and test. Wiring and electronic components must meet standards for reliability in the service environment. Paints must be environmentally friendly and strippable to enable vehicle refurbishment without damage to the basic structure. Sealants must have good work-life properties for application, and must perform in the service environment for the life of the system.

## 175.4 The Future

---

Aerospace systems of the future must be more affordable. Potential materials and process contributions to this thrust include lower costs for raw materials, fabrication, assembly, and maintenance. Innovative design concepts and materials and process improvements will help achieve the goal of more affordable systems.

There has been significant progress in the understanding of fundamental materials behavior, enabling improved prediction of materials performance in service. The understanding is not complete, however. Research must continue to ensure higher standards of excellence in aerospace materials.

Increased concern for the environment has had a significant influence on how materials are produced, processed, and disposed. Some materials are being replaced by more environmentally friendly materials that meet all performance requirements. The downward trend of toxic chemical releases into the environment is expected to continue.

**Acknowledgment:** Contributions to this chapter from Boeing colleagues Chuck Hammerberg, Ted Porter, Bert Bannink, Bill Mannick, and Dana Vana are gratefully acknowledged.

### Defining Terms

**Advanced composite material:** Stiff, strong, and lightweight material consisting of a matrix, such as epoxy, reinforced by strong, stiff filaments, such as carbon.

**Control limits:** Limits of critical process parameters for the process to be considered in control; the boundaries of acceptable performance.

**Material allowable:** A statistically derived material mechanical property value, based on coupon tests, that is used in developing design allowables. An A-basis material allowable is a value that is equaled or exceeded by 99% of the population of values with a confidence of 95%. A B-basis material allowable is a value that is equaled or exceeded by 90% of the population of values with a confidence of 95%.

### Reference

Crichlow, W. J. 1969. The materials-structures interface—A systems approach to airframe structural design, *Proceedings of the Tenth AIAA Structures, Structural Dynamics and Materials Conference*.

### Further Information

Federal Aviation Regulations (FAR) establish standards that commercial airplanes must satisfy.

MIL-STD-1530 defines the Air Force Structural Integrity Program (ASIP) to ensure structural integrity of USAF airplanes.

American Society for Testing and Materials (ASTM) is one of the largest nonprofit voluntary standards development organizations in the world.

MIL-HDBK 5 provides standardized design values and related design information for metallic materials and structural elements used in aerospace structures.

MIL-HDBK 17 provides property data on current and emerging polymeric matrix composite materials.

Three good textbooks on composites are: *Mechanics of Composite Materials* by R. M. Jones (1975); *Composites Design (4th Ed.)* by S. W. Tsai (1988); and *Mechanics of Composite Structures* by V. V. Vasiliev, (English Edition Editor R. M. Jones, 1993).

Monk, J. C. "Propulsion Systems"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## 176.1 Performance Characteristics

## 176.2 Liquid Rocket Engine Cycles

Pressure-fed • Expander • Gas Generator • Staged Combustion

## 176.3 Major Components

Main Injector • Thrust Chamber • Turbomachinery

## 176.4 System Preliminary Design Process

## 176.5 Conclusion

### Jan C. Monk

*National Aeronautics and Space Administration*

Rocket propulsion is an application of Newton's first, second, and third laws of motion. Newton's first law of motion states that a particle not subjected to external forces remains at rest or moves with constant velocity in a straight line. A rocket lifting off the launch pad goes from a state of rest to a state of motion. Newton's second law of motion states that force equals mass times acceleration. Force in the equation is the rocket thrust, where mass is the amount of rocket fuel being burned and converted into gas, which expands and then escapes from the rocket. As the gas exits the combustion chamber through a nozzle, it picks up speed. Newton's third law of motion states for every action, there is an equal and opposite reaction. With rockets, the action is the expelling of gas out of the engine; the reaction is the force or thrust of the rocket in the opposite direction.

## 176.1 Performance Characteristics

---

In the process of producing thrust, rocket engines generate more power per unit weight than any other engine. To enable a rocket to climb into low-earth orbit, it is necessary to achieve velocities in excess of 28 000 km per hour. Escape velocity is a speed of about 40 250 km per hour. To achieve these velocities, the rocket engine must burn a large amount of fuel and push the resulting gas out of the engine as rapidly as possible. Containing and controlling this power is the basic challenge in the development of these devices. For example, the power density produced by liquid hydrogen (LH<sub>2</sub>) turbomachinery utilized by the space shuttle main engine (SSME) is approximately 83 horsepower per pound of turbopump weight.

Rocket propulsion system design solutions are quite varied: thrust levels from ounces to millions of pounds force; liquid and solid **propellants**; and liquid systems with pressures that are maintained by turbopumps or pressurized tanks. Liquid system applications vary from small

pressure-fed, storable monopropellant thrusters for keeping satellites stationary, to large turbopump-fed, cryogenic bipropellant engines for boost propulsion. **Combustion chamber** pressures vary from a few pounds per square inch (psi) to several thousand psi. Generally, liquid propulsion systems consist of a propellant feed system, an **injector**, a combustion chamber, and a nozzle. The propellant feed system includes ducting and valves for controlling flows and, in the case of pump-fed systems, turbomachinery that draws propellants from lightweight propellant tanks and increases the pressure to the level necessary to support the desired combustion chamber pressure.

The ideal rocket propulsion equation is

$$\Delta V_{\text{ideal}} = g_0 I_{\text{sp}} \ln \frac{M_0}{M_1} \quad (176.1)$$

where  $\Delta V_{\text{ideal}}$  is the ideal delta velocity imparted on a vehicle,  $g_0$  is the gravitational constant,  $I_{\text{sp}}$  is the propulsion system's specific impulse,  $M_0$  is the initial mass of the vehicle, and  $M_1$  is the final or burnout mass of the vehicle. This equation provides two important performance parameters: specific impulse, which is a measure of propulsion system efficiency expressed in seconds, and vehicle burnout mass, which includes all structures (tankage, thrust structure, etc.), residual propellants, engine systems, feed systems, pressurization systems, auxiliary systems, electronic systems, upper stages, payload supporting structures, and the payload itself.

One of the more important internal rocket engine parameters is characteristic exhaust velocity, commonly referred to as C-star ( $C^*$ ), which relates combustion chamber pressure, chamber throat area, and propellant flow rate. Theoretical characteristic exhaust velocity  $C^*$  is computed as follows:

$$C^* = \frac{P_{ns} A_t g_0}{\dot{w}_{tc}} \quad (176.2)$$

where  $P_{ns}$  is **nozzle** stagnation pressure in psi,  $A_t$  is throat area in square inches, and  $\dot{w}_{tc}$  is chamber propellant mass flow rate in pounds-mass per second. A number of losses will reduce the actual  $C^*$  realized. These losses are generally a function of injector design and are related to mixture ratio maldistribution, mixing, etc. The actual  $C^*$  realized by a given design is,

$$C_{\text{act}}^* = \eta_{c^*} C^* \quad (176.3)$$

where  $\eta_{c^*}$  is  $C^*$  efficiency, typically between 0.80 and 0.99.

Another useful parameter is thrust coefficient, which relates thrust  $F$ , chamber pressure, and throat area as follows:

$$F = C_F P_{ns} A_t \quad (176.4)$$

where  $C_F$  is the thrust coefficient,  $P_{ns}$  is nozzle stagnation pressure, and  $A_t$  is throat area. Once again, additional parameters must be added to reflect actual values. This yields the following:

$$F = \eta_{CF} C_F P_{ns} A_t - P_a A_e \quad (176.5)$$

where  $\eta_{CF}$  is thrust coefficient efficiency, typically between 0.90 and 0.97,  $P_a$  is local atmospheric pressure in psi, and  $A_e$  is exit area in square inches. This equation yields thrust at any point between **sea level** and **vacuum** conditions.

Specific impulse  $I_{sp}$  is an overall efficiency term and is defined as

$$I_{sp} = \frac{F}{\dot{w}_t} \quad (176.6)$$

where  $F$  is thrust level in pounds-force and  $\dot{w}_t$  is the total mass flow rate in pounds-mass per second. Specific impulse can be computed for the engine or thrust chamber by utilizing either engine thrust and flow rate or thrust chamber thrust and flow rate, as appropriate. Specific impulse can also be computed if  $C^*$  and the thrust coefficient are known. This relationship is expressed as

$$I_{sp} = \frac{C^* C_F}{g_0} \quad (176.7)$$

Again, one must maintain consistency between theoretical values and actual values.

Thrust and specific impulse are commonly calculated at either sea level or vacuum conditions for reference or comparative purposes. Later discussions will refer to sea level thrust ( $F_{sl}$ ), vacuum thrust ( $F_{vac}$ ), sea level specific impulse ( $I_{sp_{sl}}$ ), and vacuum specific impulse ( $I_{sp_{vac}}$ ).

Mixture ratio is the ratio between the oxidizer and fuel flow rates, and is expressed in equation form as

$$MR = \frac{\dot{w}_o}{\dot{w}_F} \quad (176.8)$$

where  $\dot{w}_o$  is oxidizer flow rate in pounds per second and  $\dot{w}_F$  is fuel flow rate in pounds per second. Mixture ratio can be computed for the engine or thrust chamber by utilizing either engine flow rates or thrust chamber flow rates, as appropriate.

Expansion ratio  $\varepsilon$  is a ratio of the thrust chamber nozzle exit area,  $A_e$ , and the thrust chamber throat area,  $A_t$ :

$$\varepsilon = \frac{A_e}{A_t} \quad (176.9)$$

A more complete definition of these and other rocket engine equations, including solid propellant systems, can be found in *Rocket Propulsion Elements* [Sutton, 1992].

## 176.2 Liquid Rocket Engine Cycles

A number of power cycles are available for liquid propellant systems. These include pressure-fed,

expander, gas generator, and staged combustion cycles. Each cycle has advantages and disadvantages; the one selected for a given application is determined after a series of system trade studies. A description of a number of engine systems is given in [Table 176.1](#). A brief description of each of these power cycles follows.

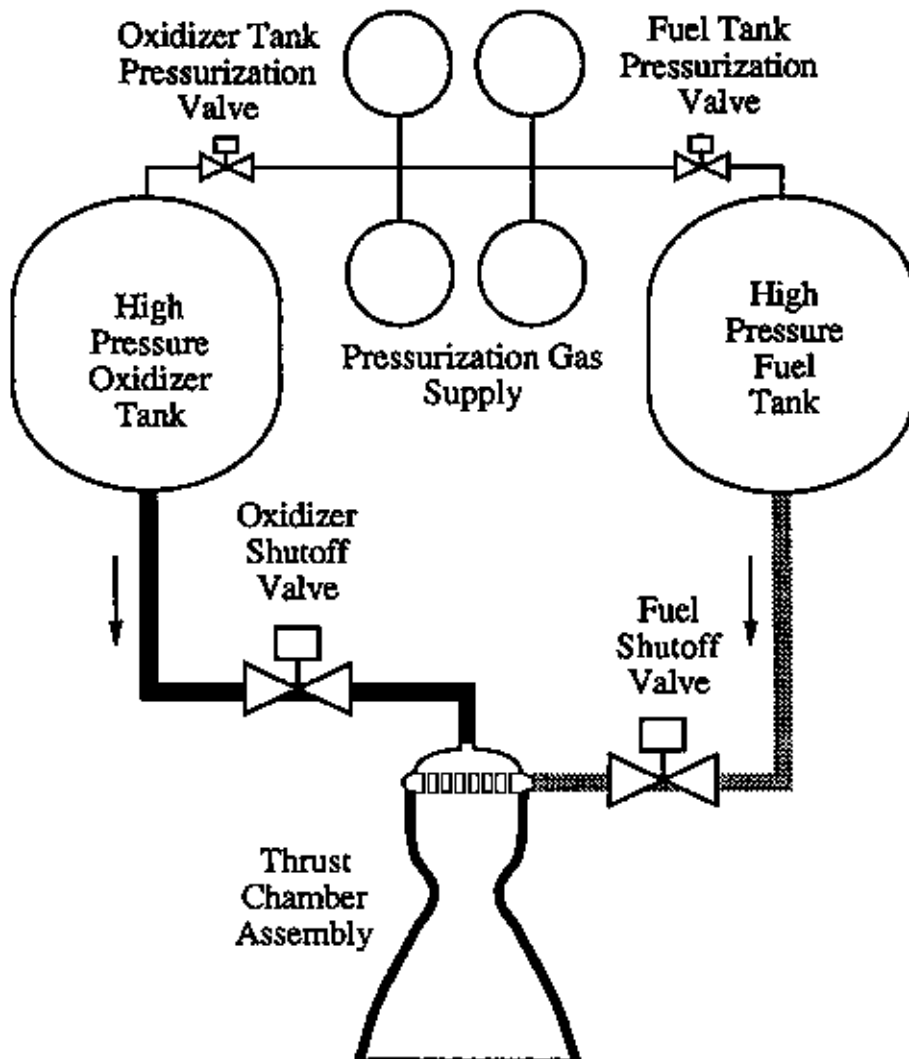
**Table 176.1** Liquid Rocket Engine Characteristics

Engine	SSME	RL-10-A4	LR87	LR91	RS-27A	F-1	H-1	J-2
Producer	Rocketdyne	P&W	Aerojet	Aerojet	Rocketdyne	Rocketdyne	Rocketdyne	Rocketdyne
Launch vehicle	Space shuttle	Centaur	Titan II/IV first stage	Titan II/IV second stage	Delta	Saturn V first stage	Saturn IB first stage	Saturn IB second stage/ Saturn V second stage/ Saturn V third stage
Status	Active	Active	Active	Active	Active	Inactive	Inactive	Inactive
Propellants	LO <sub>2</sub> /LH <sub>2</sub>	LO <sub>2</sub> /LH <sub>2</sub>	N2O4/Aerozine 50	N2O4/Aerozine 50	LO <sub>2</sub> /RP-1	LO <sub>2</sub> /RP-1	LO <sub>2</sub> /RP-1	LO <sub>2</sub> /LH <sub>2</sub>
Thrust, sea level, lbf	394 000	—	446 000 (dual thrust chamber system)	—	200 000	1 522 000	206 145	—
Thrust, vacuum, lbf	488 800	20 800	529 000	100 000	237 000	1 748 000	230 170	230 000
Specific impulse, sea level, s	365.1	—	254.0	263.3	255.0	265.4	264.9	—
Specific impulse, vacuum, s	452.9	449.0	302.0	318.0	302.0	304.1	295.8	427.0
Mixture ratio	6.00	5.50	1.90	1.86	2.25	2.27	2.23	5.50
Total flow rate, lb/s	1079.3	46.3	1751.7	314.5	784.8	5734.7	778.1	538.6
Fuel flow rate, lb/s	154.2	7.1	604.0	110.0	241.8	1753.7	240.9	82.9
Oxidizer flow rate, lb/s	925.1	39.2	1147.6	204.5	542.9	3981.0	537.2	455.8
Nozzle stagnation, psia	3100	575	827	827	700	982	652	763
C-star (engine), ft/s	7507	7696	5611 (each)	5597	10 707	5297	5509	—
Area ratio	77.5	84	15	49.2	12	16	8	27.5
Throat area, in. <sup>2</sup>	81.2	19.3	182.0 (each)	65	373.08	961.4	204.35	—
Length, in.	168	70/90	150	110.62	149	220.4	101.61	133
Exit diameter, in.	96.00	46.00	86.25 (each)	64.00	76.00	143.50	45.62	80.50
Exit area, in. <sup>2</sup>	—	1618.7	5842.6	5842.6	5842.6	15 400.0	1634.8	—
Powerhead diameter, in.	99.8	—	85	64	—	104.75	—	—
Dry weight, lb	7004	370	4583	1284	2444	18 616	2003	3480
Turbine drive	Fuel rich Preburner	Heated hydrogen (expander)	Gas generator	Gas generator	Gas generator	Gas generator	Gas generator	Gas generator
Start method	Tank head	Tank head	Solid propellant cartridge	Solid propellant cartridge	Solid propellant cartridge	Tank head	Solid propellant cartridge	Gas bottle

## Pressure-fed

This system consists of a **thrust chamber assembly**, associated ducting and valves necessary for control, pressurized tankage, and the pressurization system for the tankage. This system is widely utilized for satellite attitude control, orbital transfer, and as auxiliary propulsion for most major launch vehicles. Pressure-fed systems are perhaps the simplest of all propulsion systems, but are performance limited because of the weight penalty associated with increasing chamber pressures. As pressures increase, tank wall thickness and the mass of the gases needed to maintain tank pressures increase. Tank pressures are set by chamber pressure plus pressure losses in the cooling circuit (if any), injector, valves, and ducting. In most pressure-fed applications, combustion chambers are passively cooled (i.e., film-cooled or radiative/**ablative**). The space shuttle utilizes pressure-fed systems for the orbital maneuvering system and the reaction control system. A schematic of a simple pressure-fed system is given in [Fig. 176.1](#).

**Figure 176.1** Pressure-fed propulsion system schematic.



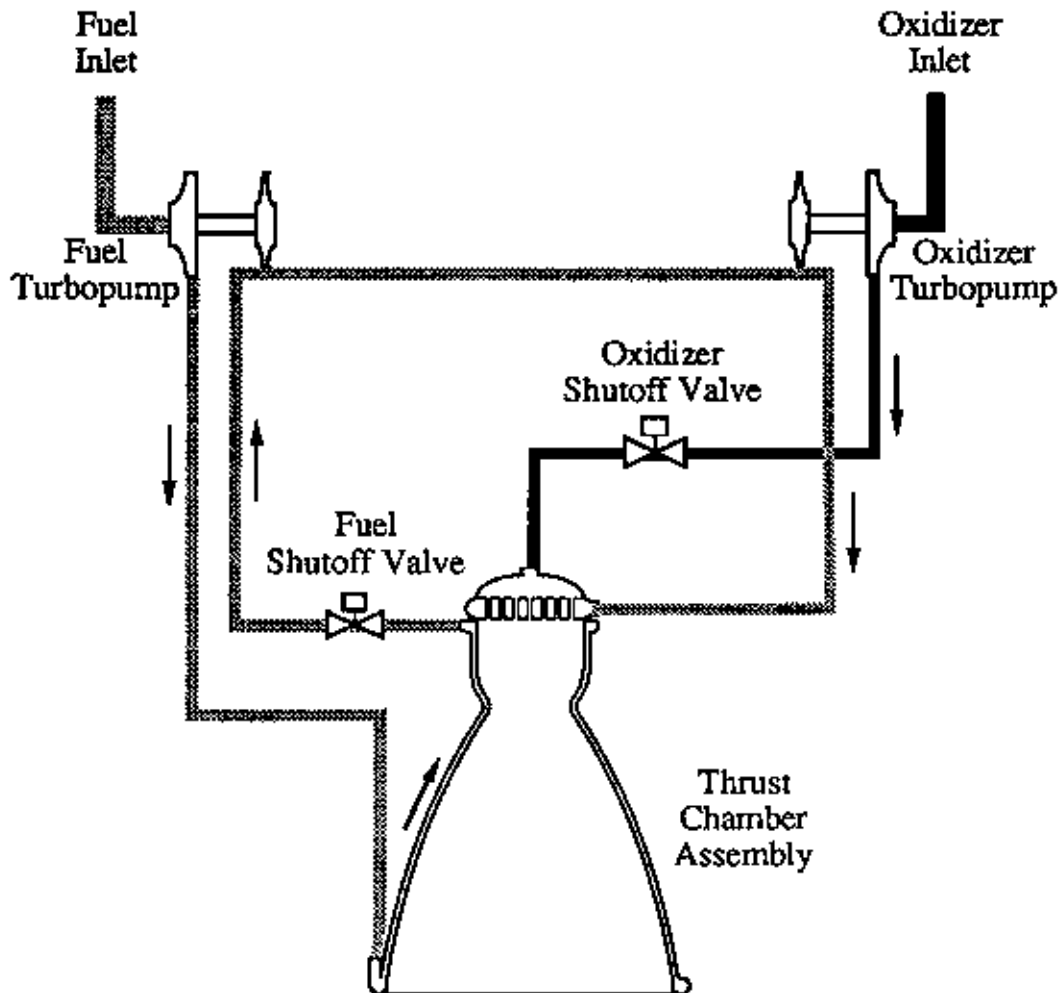
## Expander

This is the simplest of the turbopump-fed systems primarily because the power source for the turbines is the thrust chamber cooling circuit. Only the thrust chamber requires an ignition system. Pump discharge pressures are set by chamber pressure plus pressure losses in the cooling circuit, turbine, injector, valves, and ducting. The combustion chamber is **regeneratively cooled**. In some applications, extensible radiation-cooled nozzle extensions are used to increase area ratio while maintaining a short stowed length. Expander cycles are limited in the combustion chamber pressure that can be attained because the energy available to drive the turbine(s) is obtained from the combustion chamber cooling circuit. For applications that require operation at sea level, this reduces the area ratio that can be achieved without **side loads**. Nozzle flow separation is discussed



later. The RL10 engine is utilized by the Centaur upper stage for the Atlas-Centaur and Titan-Centaur launch vehicles. A schematic of a simple expander system is given in [Fig. 176.2](#) .

**Figure 176.2** Expander engine system schematic.

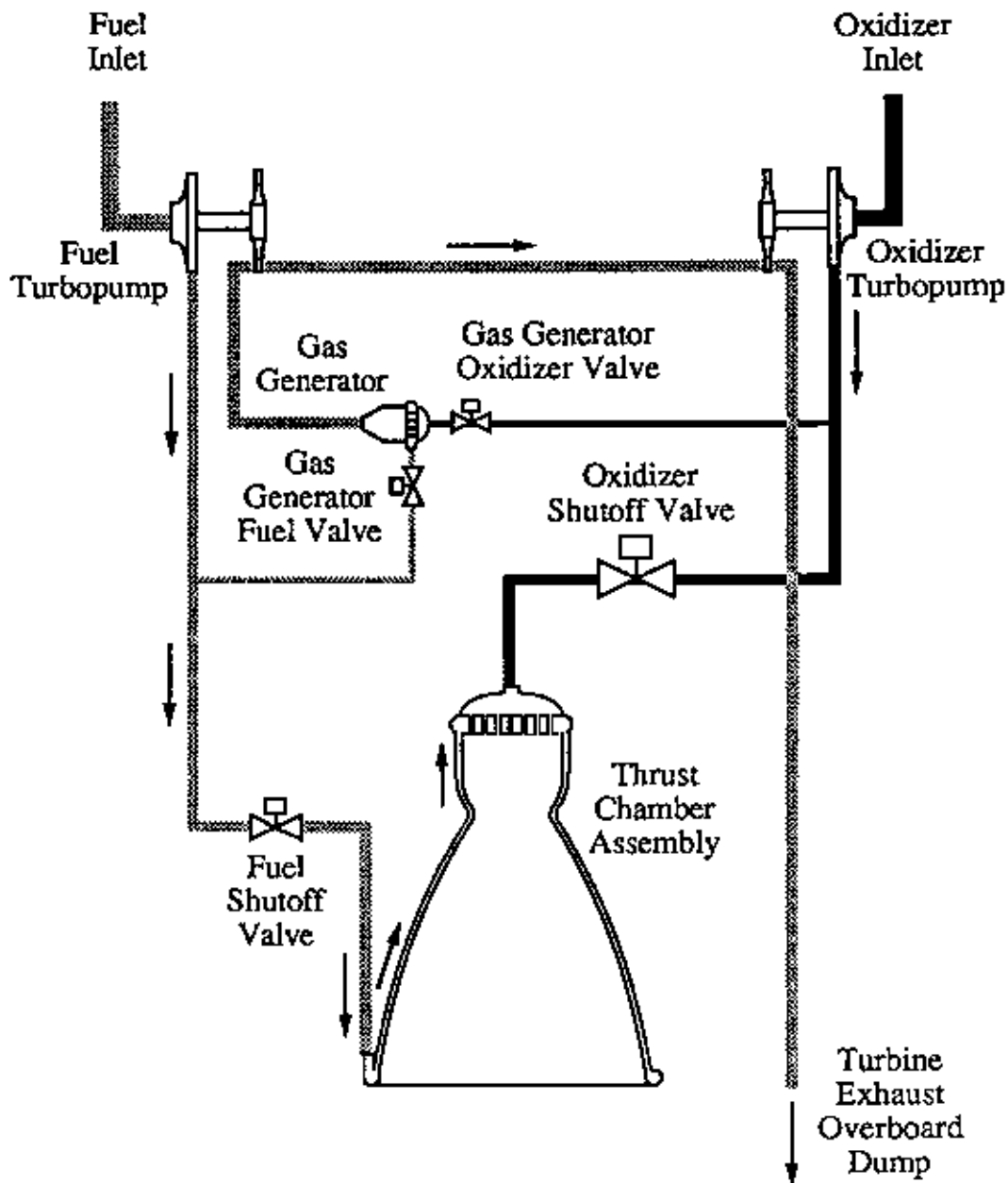


## Gas Generator

This is the most common engine cycle in use today. Turbine power is derived from a separate combustor or gas generator (GG) which utilizes the same propellants as the main system. This hot gas is routed through the turbopump turbines and is dumped overboard. Pump discharge pressures are set by chamber pressure plus pressure losses in the cooling circuit, injector, valves, and ducting. Because the gas generator is parallel to the main chamber, turbine pressure losses do not impact pump discharge pressure in most designs. This highlights one of the disadvantages of this cycle. The gas generator propellants are not used in the main chamber to produce thrust. Some concepts use GG gases for cooling nozzle extensions, but the thrust added is minimal. Gas

generators are operated at relatively low mixture ratios because turbine temperatures must be maintained in the 1000 to 2000 degrees Rankine range. The main combustor mixture ratio is biased higher to offset this parasitic flow. The combination of poor thrust efficiency of the GG gases and main chamber mixture ratio bias results in a specific impulse penalty. The gas generator cycle was utilized on the F-1 and J-2 engines of the Saturn V launch vehicle and is currently in use on the Delta, Atlas, and Titan launch vehicles. A schematic of a simple gas generator system is given in Fig. 176.3.

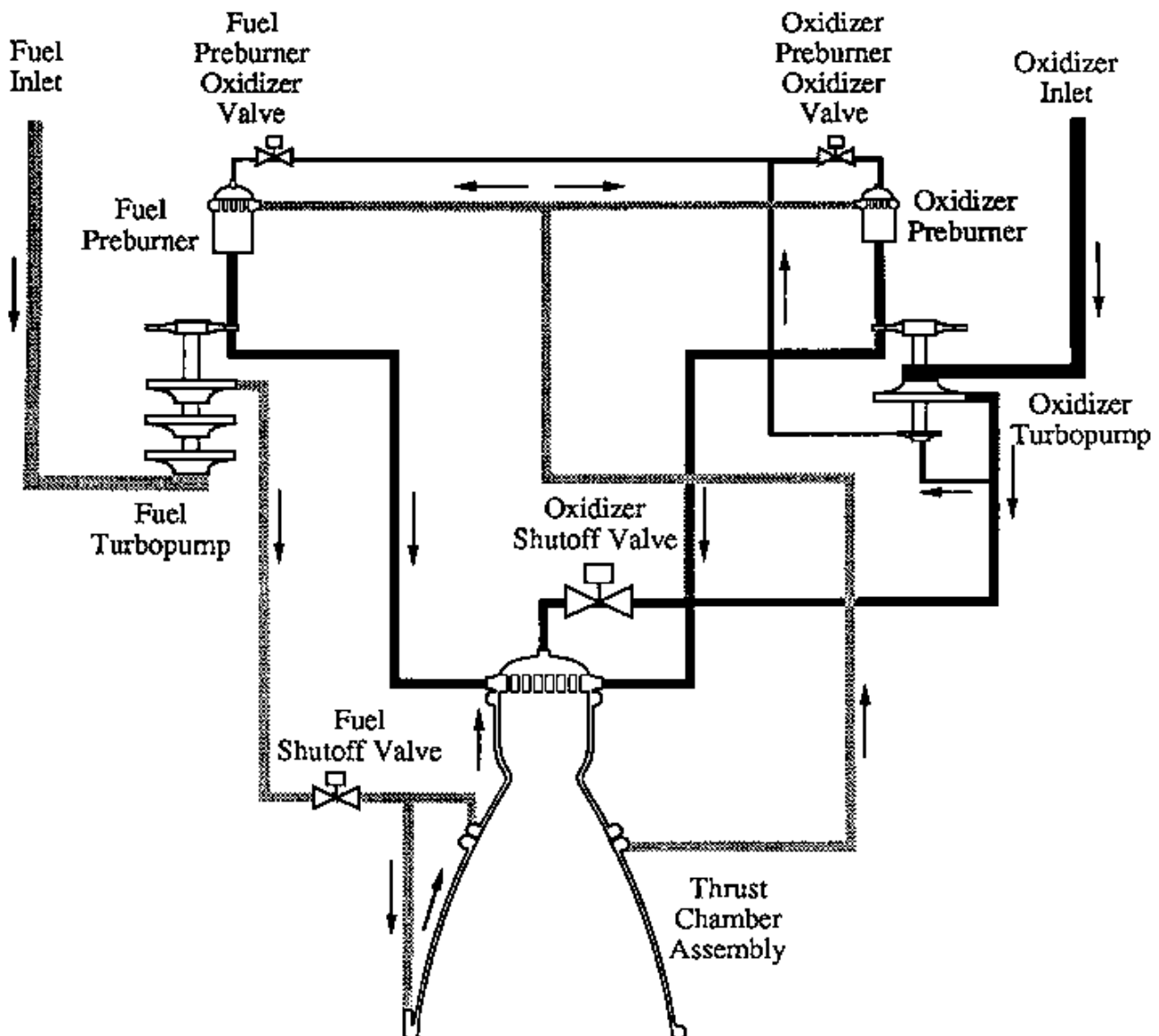
**Figure 176.3** Gas generator engine system schematic.



## Staged Combustion

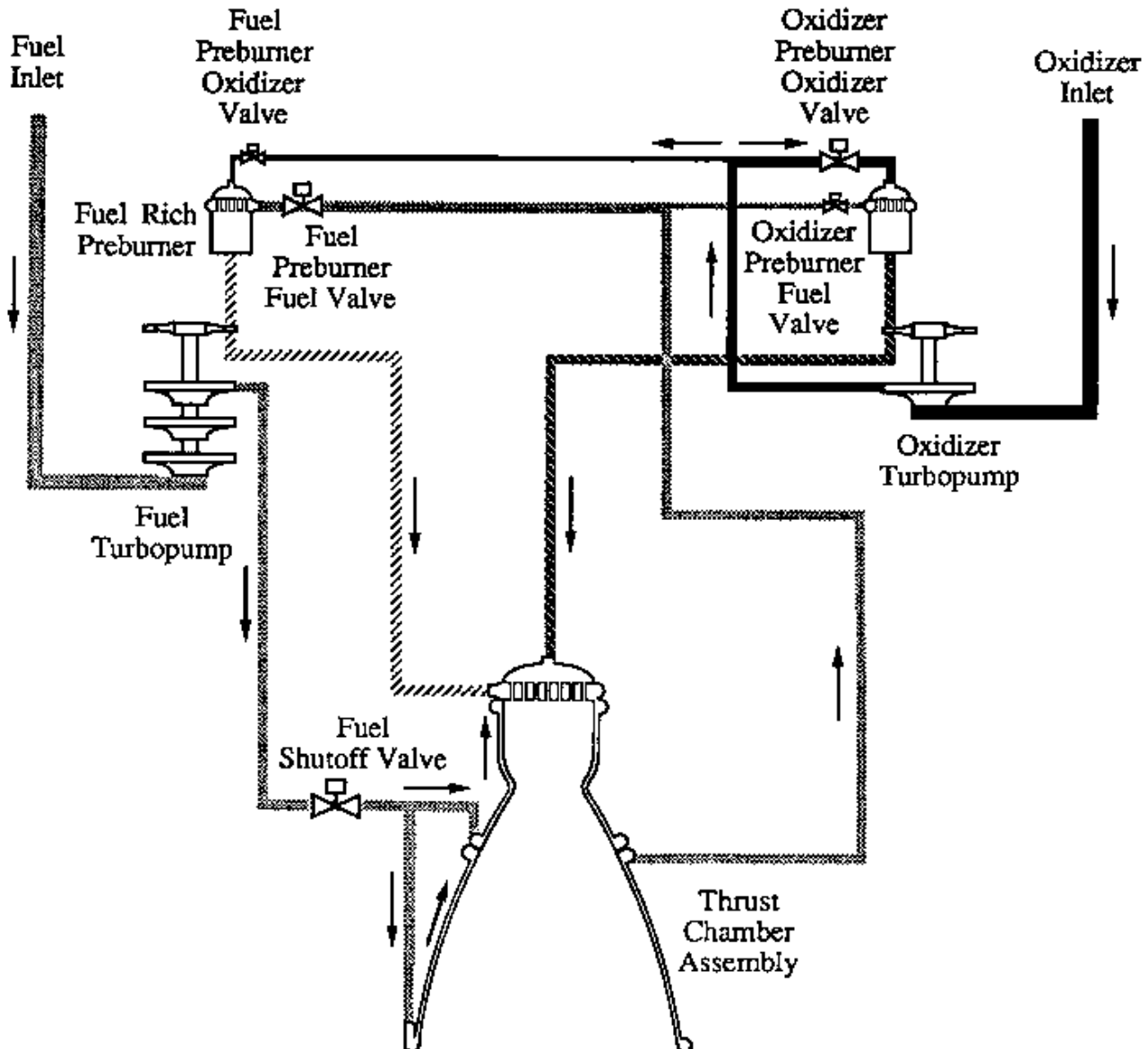
The staged combustion cycle provides the highest performance of conventional chemical rocket engines. Turbine power is derived from a separate combustor or preburner which also utilizes the same propellants as the main system. In bipropellant systems, the hot gas is routed through the turbopump turbines to the main injector where it is mixed with the other propellant and is combusted in the main chamber. Pump discharge pressures are set by chamber pressure plus pressure losses in the cooling circuit, turbine, injector, valves, and ducting. Thrust chambers are regeneratively cooled. Staged combustion cycle engines developed in the U.S. have utilized a fuel-rich preburner. Several rocket engine systems developed in Russia have utilized an oxidizer-rich preburner. In the former case, the fuel-rich hot gases are mixed with oxidizer in the main chamber. In the latter, oxidizer-rich hot gases are mixed with fuel in the main chamber. The staged combustion cycle utilizes all propellants in the main combustion chamber, which provides maximum performance. A schematic of a simple staged combustion system is given in [Fig. 176.4](#).

**Figure 176.4** Staged combustion cycle engine system schematic.



A variant of the staged combustion cycle is the full flow cycle, in which the oxidizer pump turbine is driven with an oxidizer-rich preburner and the fuel pump is driven with a fuel-rich preburner. This cycle offers some simplification in turbomachinery design because of the simplified seal design between the pump end and the turbine end, and a significant reduction in turbine temperatures because all the propellants can be utilized in the turbine drive circuits. This concept is currently under study as a candidate engine system for the next generation launch vehicle. A schematic of a simple full flow staged combustion system is given in Fig. 176.5 .

**Figure 176.5** Full flow staged combustion cycle engine system schematic.

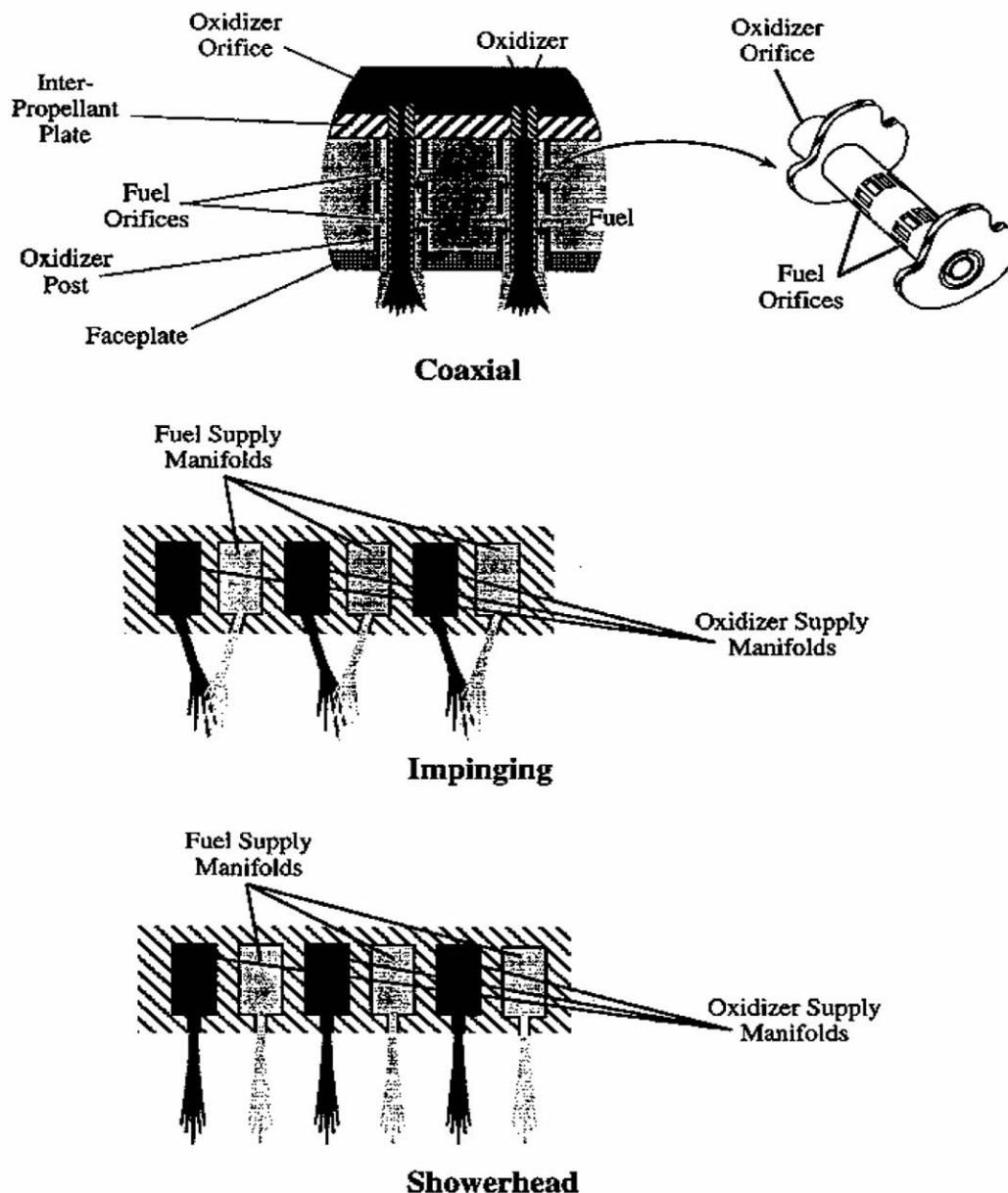


## 176.3 Major Components

### Main Injector

The purpose of the injector is to introduce propellants into the combustion chamber in a controlled manner, to atomize the propellants, and to mix the propellants at the proper mixture ratio in a homogenous manner. Mixture ratio variations across the injector face are one of the most common problems that the designer will encounter, and these maldistributions lead to combustion efficiency losses. In some cases, maldistributions are deliberately introduced. To enhance the durability of combustion chambers, a film of fuel is injected at the outer circumference of the injector. In order to produce a stable combustion process, baffle elements which are cooled with fuel are commonly used. The most common injector concepts are coaxial, showerhead, and impinging. These are illustrated in Fig. 176.6.

**Figure 176.6** Injector concepts.

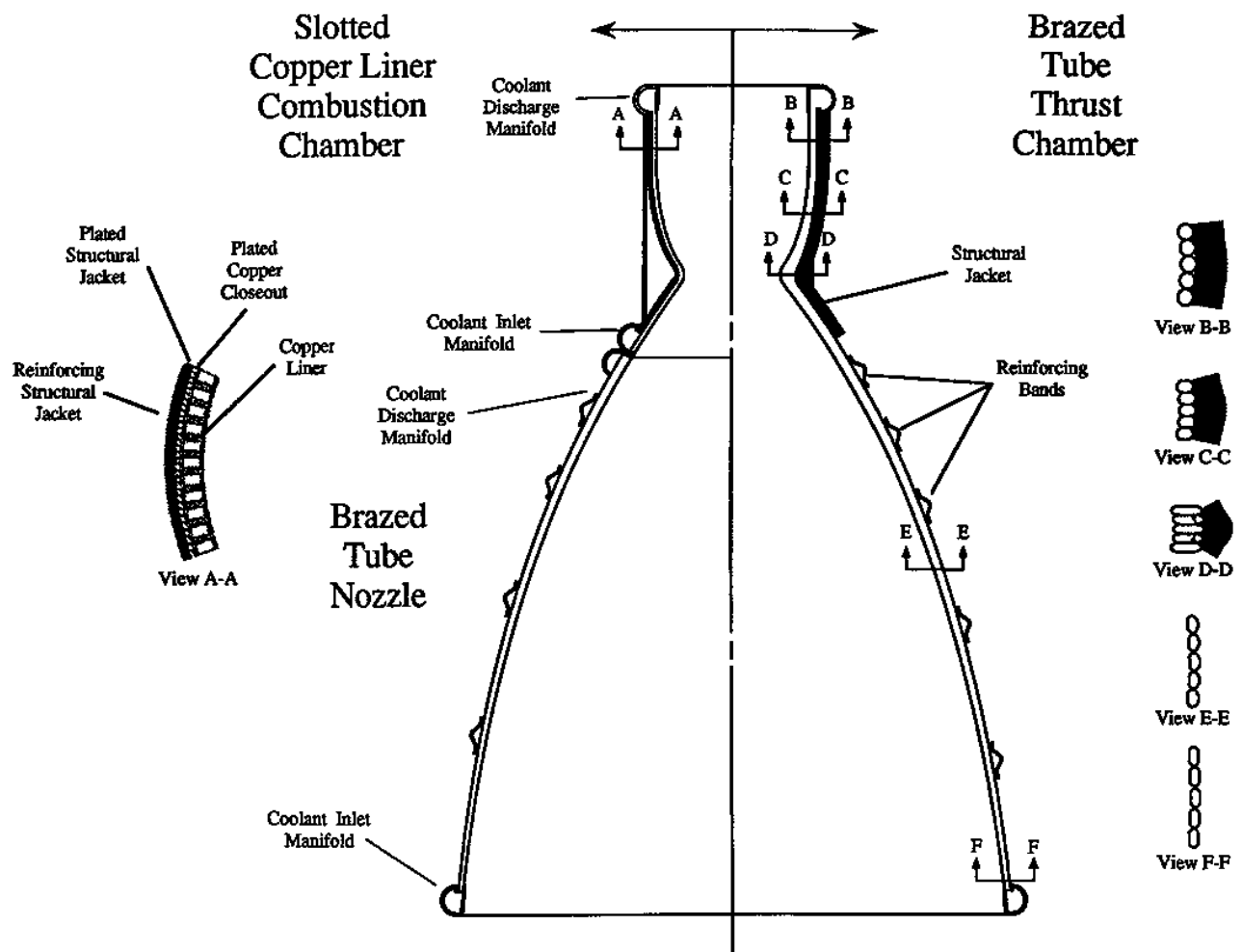


The coaxial injector consists of a series of concentric tubes into which the oxidizer is introduced through a center tube and the fuel introduced through the annular area formed by a second tube. This type of injector is commonly used in oxygen-hydrogen engines. Impinging and showerhead injectors, which are commonly used in oxygen-kerosene and storable propellant engines, consist of a series of two sets of orifices. One injects the oxidizer, and the other the fuel. The number of orifices, injection velocities, and injection angles are selected to provide consistent atomization and mixing of propellants. Impinging injectors slant the orifices to impinge the two propellant streams against each, enhancing mixing. The design utilized by the conventional impinging injector consists of a series of concentric copper rings containing the injection orifices. The rings alternate between oxidizer and fuel. The outer ring is generally a fuel ring and contains a set of smaller orifices that control film coolant for the combustion chamber. Separate manifolding routes propellants to each set of orifices.

## Thrust Chamber

A number of design solutions have been utilized in thrust chambers, varying from passively cooled ablatives to a number of regeneratively cooled concepts. In some applications, the thrust chamber is composed of two separate components. The upper portion—including the throat region and a portion of the expansion region—is commonly called a combustion chamber. The lower portion—consisting of the remainder of the expansion region—is called a nozzle. Regeneratively cooled thrust chamber designs include brazed tube bundles, copper with milled channels, and steel with milled channels. The bundled tube concept utilizes: steel tubes, pressed to vary the shape necessary for formation of the overall thrust chamber shape; a structural shell in the combustion chamber region with a number of straps spaced along the thrust chamber length for additional strength; and necessary manifolding for inlet and discharge coolant flow. For higher pressure applications (greater than approximately 1800 psia), the heat load produced by the combustion process exceeds the capability of brazed tube designs. For these applications, a copper liner is required in the high heat flux region. This configuration consists of a slotted copper liner, structural jacket, and manifolding. [Figure 176.7](#) illustrates these two thrust chamber concepts.

**Figure 176.7** Two thrust chamber concepts.







#### MANNED MANEUVERING UNIT

Astronaut Bruce McCandless II is a few meters away from the cabin of the earth-orbiting Space Shuttle *Challenger* in this 70 mm photograph taken on February 7, 1984. McCandless is one of the two 41-B mission specialists who participated in this historical extravehicular activity (EVA). This spacewalk represented the first use of a nitrogen-propelled, hand-controlled device called the



Manned Maneuvering Unit (MMU), which allows for much greater mobility than that afforded previous spacewalkers, who had to use restrictive tethers.

The MMU is a self-contained backpack with nitrogen gas propulsion that allows orbiter crews to move outside the payload bay to other parts of the orbiter or to other spacecraft. The MMU latches to the spacesuit (Extravehicular Mobility Unit, EMU) backpack and can be donned and doffed by an astronaut unassisted.

MMU controls follow the layout familiar to spacecraft crews: the left-hand controller governs fore–aft, right–left, and up–down translations, while the right-hand controller handles roll, pitch, and yaw motions. The controllers may be used singly or in combination to give a full range of movement within the operating logic of 729 command combinations, including attitude hold.

Thrust impulses are from 24 dry nitrogen gas thrusters each with 7.56 newtons thrust. Two 25-by-76 centimeter (9.8-by-30 inch) Kevlar filament-wrapped aluminum nitrogen tanks each hold 5.9 kilograms (13 pounds) of nitrogen when fully charged. Two 16.8 volt, 752 watt-hour silver zinc batteries supply MMU electrical power, enough for one six-hour EVA. The nitrogen tanks could be recharged in less than 20 minutes at the payload bay MMU service rack.

Built by Martin Marietta, Denver, CO, the MMU is 1.2 m (49.4 in.) high, 81 cm (32.5 in.) wide, and 1.1 m (44.2 in.) deep with control arms extended. The MMU weighs 136 kg (300 lb) when charged with nitrogen. With a spacesuited crewman and consumables added, on-orbit mass is about 335 kg (740 lb). (Photo courtesy of National Aeronautics and Space Administration.)

## Turbomachinery

The turbomachinery design process of liquid rocket engines is very similar to a normal pump/turbine design, except for two critical areas. The first is the critical need to minimize weight. This is perhaps the greatest difference. As stated earlier, the power density of the space shuttle main engine turbopump is 83 horsepower per pound of turbopump weight. The second difference is the dynamic and steady state environments that rocket engines require. Although a number of turbojet engines operate at turbine temperatures significantly higher than most rocket engines, they attain the steady state operating point in a matter of minutes, not in one to four seconds as do rocket engine turbines. This produces severe thermal strains that tax the ability of materials to sustain. Other environments that provide problems in some materials are oxygen and hydrogen. Particle impact, fretting, and rubbing in an oxygen environment can lead to disastrous fires. Susceptibility of materials to hydrogen embrittlement reduces the variety of materials available for the designer or requires platings to protect materials. Another environment to which rocket engine turbomachinery is susceptible is rotor dynamics, which is considerably more critical than in conventional rotating machinery because of the reduced weight of rocket turbopumps. Structural design considerations, including explanation of the processes utilized in the SSME, can be found in *Structural Design/Margin Assessment* [Ryan, 1993].

## 176.4 System Preliminary Design Process

A number of theoretical thrust chamber performance computer models are readily available that provide the basic performance parameters needed to support a conceptual design. The most common is a Finite Area Combustor Theoretical Rocket Performance Program, commonly referred to as the One-Dimensional Equilibrium (ODE) Program referenced in *Computer Program for Calculation of Complex Chemical Equilibrium Compositions and Applications, Supplement I-Transport Properties* [Gordon et al., 1984]. This model generates performance data for various

propellant combinations as a function of mixture ratios, combustion chamber pressures, and nozzle expansion ratios. A sample set of data for liquid oxygen (LO<sub>2</sub>)/liquid hydrogen propellants is given in Table 176.2.

**Table 176.2** Theoretical Performance of Oxygen-Hydrogen Combustor

$P_c$	MR	C-star	$\varepsilon$	$P_e$	$C_f$	$P_c$	MR	C-star	$\varepsilon$	$P_e$	$C_f$	$P_c$	MR	C-star	$\varepsilon$	$P_e$	$C_f$
100.0	5.0	7680	10	1.29	1.760	1000	5.0	7795	10	1.29	1.745	3000	5.0	7829	10	36.26	1.741
			25	0.37	1.860				25	0.37	1.841				25	10.45	1.835
			50	0.15	1.917				50	0.15	1.895				50	4.10	1.889
			100	0.06	1.962				100	0.06	1.938				100	1.61	1.931
			200	0.02	1.996				200	0.02	1.971				200	1.61	1.963
	5.5	7547	10	1.37	1.774		5.5	7692	10	1.37	1.757		5.5	7741	10	0.63	1.751
			25	0.40	1.882				25	0.40	1.859				25	11.08	1.851
			50	0.16	1.944				50	0.16	1.917				50	4.40	1.908
			100	0.06	1.993				100	0.06	1.963				100	1.75	1.953
			200	0.02	2.032				200	0.02	2.000				200	0.69	1.989
	6.0	7408	10	1.46	1.786		6.0	7575	10	1.46	1.769		6.0	7637	10	39.74	1.762
			25	0.43	1.901				25	0.43	1.877				25	11.74	1.867
			50	0.17	1.969				50	0.17	1.939				50	4.71	1.928
			100	0.07	2.022				100	0.07	1.989				100	1.89	1.977
			200	0.03	2.066				200	0.03	2.029				200	0.76	2.016
	6.5	7266	10	1.53	1.794		6.5	7448	10	1.53	1.780		6.5	7522	10	41.63	1.733
			25	0.47	1.917				25	0.47	1.894				25	12.43	1.883
			50	0.19	1.990				50	0.19	1.961				50	5.03	1.947
			100	0.08	2.049				100	0.08	2.015				100	2.04	2.000
			200	0.03	2.097				200	0.03	2.059				200	0.83	2.042
	7.0	7127	10	1.58	1.799		7.0	7316	10	1.58	1.788		7.0	7396	10	43.58	1.782
			25	0.50	1.928				25	0.50	1.908				25	13.20	1.898
			50	0.21	2.007				50	0.21	1.980				50	5.38	1.967
			100	0.08	2.071				100	0.08	2.038				100	2.20	2.024
			200	0.03	2.124				200	0.03	2.086				200	0.90	2.070
500.0	5.0	7767	10	6.20	1.749	2000	5.0	7818	10	6.20	1.742	4000	5.0	7836	10	48.21	1.740
			25	1.79	1.845				25	1.79	1.837				25	13.90	1.834
			50	0.70	1.900				50	0.70	1.890				50	5.45	1.887
			100	0.27	1.943				100	0.27	1.933				100	2.14	1.929
			200	0.11	1.977				200	0.11	1.966				200	0.83	1.962
	5.5	7654	10	6.56	1.762		5.5	7724	10	6.56	1.753		5.5	7751	10	50.39	1.750
			25	1.91	1.865				25	1.91	1.853				25	14.72	1.849
			50	0.76	1.924				50	0.76	1.911				50	5.84	1.906
			100	0.30	1.971				100	0.30	1.957				100	2.32	1.951
			200	0.12	2.008				200	0.12	1.993				200	0.92	1.987
	6.0	7529	10	6.93	1.769		6.0	7616	10	6.93	1.762		6.0	7652	10	52.67	1.760
			25	2.05	1.877				25	2.05	1.867				25	13.90	1.864
			50	0.82	1.939				50	0.82	1.928				50	5.45	1.925
			100	0.33	1.989				100	0.33	1.977				100	2.14	1.973
			200	0.13	2.029				200	0.13	2.016				200	0.83	2.012
	6.5	7397	10	7.30	1.784		6.5	7496	10	7.30	1.775		6.5	7539	10	55.10	1.771
			25	2.19	1.901				25	2.19	1.887				25	16.46	1.880
			50	0.89	1.969				50	0.89	1.952				50	6.66	1.945
			100	0.36	2.025				100	0.36	2.005				100	2.70	1.997
			200	0.15	2.069				200	0.15	2.048				200	1.09	2.038
	7.0	7262	10	7.62	1.791		7.0	7368	10	7.62	1.784		7.0	7416	10	43.58	1.780
			25	2.34	1.915				25	2.34	1.902				25	13.20	1.895
			50	0.96	1.988				50	0.96	1.971				50	5.38	1.964
			100	0.39	2.048				100	0.39	2.029				100	2.20	2.019
			200	0.16	2.097				200	0.16	2.075				200	0.90	2.065

The following process outlines a methodology of determining the initial set of overall system requirements for a liquid rocket engine. A preliminary set of top-level engine requirements must be established by the vehicle systems designer for a booster engine. For this exercise, they are as follows:

Propellants: Liquid Oxygen/Liquid Hydrogen

Sea-Level Thrust: 400000 pounds force

Mixture Ratio: 6.0:1.0

In some instances, a minimum specific impulse value is specified. However, in most cases, the design value should be selected as the result of vehicle-engine trade studies, along with engine weight, recurring costs, and nonrecurring costs.

One of the first choices to be made by the engine designer is the value of combustion chamber pressure. This choice is also the result of a series of trades. For this exercise, 2000 psia has been chosen and is a good first approximation for the optimum value when recurring costs are one of the more important parameters. This value will provide good performance, while simplifying turbomachinery to two pump stages with moderate turbine temperatures. An initial assumption needs to be made as to engine cycle. For a booster application, either a gas generator cycle or a staged combustion cycle usually provides optimum performance. For this exercise, the staged combustion cycle is selected. This simplifies the initial set of calculations in that the engine flow rate and thrust chamber flow rate are approximately identical. Table 176.3 provides a typical ODE output for a thrust chamber operating at a chamber pressure of 2000 psia and a mixture ratio of 6.0. The first parameter to select is area ratio. From previous vehicle trade studies, the optimum nozzle exit pressure ( $P_e$ ) for a first stage or booster vehicle is approximately 6.5 psia, the optimum for a single-stage-to-orbit vehicle is approximately 4.0 psia, and the optimum for a parallel burn core stage (booster and core stages ignite at sea level) is 2.5 psia. Standard atmospheric conditions for sea level and various altitudes can be found in *Terrestrial Environment (Climatic) Criteria Guidelines for Use in Aerospace Vehicle Development, 1993 Revision* [Johnson, 1993].

**Table 176.3** Theoretical Performance of Oxygen-Hydrogen Combustor at a Chamber Pressure of 2000 psia and a Mixture Ratio of 6.0

	Chamber	Throat	Exit	Exit	Exit	Exit	Exit	Exit	Exit	Exit	Exit	Exit
PINF/P	1.00	1.74	74.83	128.69	188.56	253.32	322.26	394.88	470.81	631.58	715.98	802.86
$P_e$ , atm	136.09	78.262	1.8186	1.0575	0.72172	0.53723	0.42231	0.34464	0.28906	0.21548	0.19008	0.16951
$T_e$ , K	3571.28	3362.47	1992.35	1810.36	1688.69	1598.55	1527.58	1469.4	1420.31	1340.93	1308	1278.43
$\rho$ , g/cc	6.2908-3	3.8815-3	1.5692-4	1.0044-4	7.3493-5	5.7792-5	4.7541-5	4.0334-5	3.4998-5	2.7634-5	2.4990-5	2.2801-5
$H$ , cal/g	-235.74	-515.64	-1948.34	-2093.47	-2187.58	-2255.9	-2308.88					
$U$ , cal/g	-759.64	-1003.92	-2229.01	-2348.45	-2425.4	-2481.02	-2524	-2558.69	-2587.56	-2633.49	-2652.24	-2668.94
$G$ , cal/g	-15092.1	-14503.3	-10236.4	-9624.47	-9212.44	-8905.8	-8663.53	-8464.38	-8295.98	-8022.83	-7909.25	-7807.12
$S$ , cal/(g)(K)	4.16	4.16	4.16	4.16	4.16	4.16	4.16	4.16	4.16	4.16	4.16	4.16
$M$ , mol wt	13.546	13.685	14.106	14.11	14.111	14.111	14.111	14.111	14.111	14.111	14.111	14.111
(DLV/DLP)T	-1.02183	-1.01646	-1.00018	-1.00006	-1.00002	-1.00001	-1.00001	-1.00000	-1.00000	-1.00000	-1.00000	-1.00000
(DLV/DLP)P	1.3823	1.3057	1.0054	1.0018	1.0007	1.0004	1.0002	1.0001	1.0001	1.0000	1.0000	1.0000
CP, cal/(g)(K)	1.8834	1.7168	0.8179	0.784	0.7651	0.7518	0.7414	0.7328	0.7254	0.7131	0.7079	0.7031
GAMMA (\$)	1.1455	1.1465	1.2106	1.2199	1.226	1.2307	1.2346	1.2379	1.2409	1.2461	1.2484	1.2505
Son vel, m/s	1584.6	1530.4	1192.3	1140.8	1104.5	1076.7	1054.1	1035.3	1019.1	992.2	980.9	970.5
Mach number	0.000	1.000	3.175	3.456	3.659	3.819	3.951	4.065	4.164	4.333	4.406	4.474
Performance Parameters												
AE/AT	1	10	15	20	25	30	35	40	50	55	60	
CSTAR, ft/s	7616	7616	7616	7616	7616	7616	7616	7616	7616	7616	7616	7616
CF	0.659	1.631	1.699	1.741	1.771	1.794	1.813	1.828	1.852	1.862	1.871	
CF-VAC	1.235	1.765	1.815	1.847	1.870	1.888	1.902	1.913	1.931	1.939	1.946	
IVAC, lb-s/lb	292.2	417.7	429.6	437.2	442.6	446.8	450.1	452.8	457.1	458.9	460.5	
ISP, lb-s/lb	156.1	386	402.1	412.1	419.3	424.7	429.1	432.7	438.4	440.7	442.8	
$P_e$ , psia	1150.1	26.7	15.5	10.6	7.9	6.2	5.1	4.2	3.2	2.8	2.5	

PINF = 2000.0 psia.

O/F = 6.0000.

Another consideration is side loads on the nozzle. For an overexpanded nozzle ( $P_e < P_a$ ), unsymmetrical flow separation results if significant dynamic loads are applied to the nozzle (pressure times surface area forces). These side loads can ultimately destroy a nozzle and, at a minimum, result in significant weight increases. Side loads are computed using both empirical techniques and detailed nozzle fluid flow with computational fluid dynamics techniques. Calculating the locations within the nozzle where flow separation will occur is very difficult and side loads can usually be quantified accurately with test data.

As the application for this exercise is a booster, an exit pressure of approximately 6.5 psia is desired. Using Table 176.3, an expansion ratio of 30 yields an exit pressure of 6.2 psia, which is close enough for a first approximation of the engine characteristics. Using the initial vehicle thrust and mixture ratio requirements, theoretical values of  $C^*$  and  $C_F$  obtained from Table 176.3, and assumed values of  $C^*$  efficiency of 0.99 (readily obtainable with oxygen/hydrogen coaxial tube injectors) and  $C_F$  efficiency of 0.97, various engine parameters can be computed using Eqs. (176.3), (176.5), (176.6), (176.8), and (176.9) and are shown below in order of computation:

$F_{vac}$	400 000 lb
Propellants	Oxygen/hydrogen
MR	6.0
Cycle	Staged combustion
$P_{ns}$	2000 psia
$\varepsilon$	30:1
$C_F$	1.8311
$C^*$	7539.8 ft/s
$A_i$	109.22 in <sup>2</sup>
$A_e$	3276.7 in <sup>2</sup>
$\dot{w}_t$	932.27 lbm/s
$\dot{w}_f$	133.18 lbm/s
$\dot{w}_o$	799.09 lbm/s
$I_{sp_{vac}}$	429.1 s
$F_{sl}$	351 846 lbf
$I_{sp_{sl}}$	377.41 s

## 176.5 Conclusion

---

The science of rocketry has enabled some of humankind's greatest achievements, ranging from instantaneous global communications and accurate weather forecasting via geostationary satellites to trips to the moon. NASA's space shuttle is one of the most complex flying machines ever built and is the only partially reusable launch vehicle. While several countries have rockets capable of

carrying a variety of payloads, the U.S. and Russia are the only countries with spacecraft that can transport a crew to and from orbit. France and China have expendable launch vehicles now in use, and Japan is developing another.

The U.S. is on the threshold of a next generation launch system. NASA aerospace engineers and industry experts are exploring new concepts, including the first fully reusable launch vehicle. Developing rockets for 21st century missions promotes enhanced technologies to meet new challenges, including balancing design requirements between operability, performance, weight, and cost. The next generation of spaceship will open new doors to the space frontier.

## Defining Terms

**Ablation:** A passive cooling technique in which heat is carried away from a vital part by absorption into a nonvital part, which may melt or vaporize and then fall away, taking the heat with it.

**Combustion chamber:** A device—which includes a throat region—to mix, burn, and control propellants.

**Injector:** A device to distribute and inject propellants into the combustion chamber.

**Nozzle:** A device used to accelerate the combusted gases.

**Propellant:** Fuel [the chemical(s) the rocket burns] and an oxidizer (oxygen compounds) to ignite the fuel.

**Sea level:** Standard atmospheric conditions at an altitude of zero feet.

**Side loads:** Unsymmetrical loads put on a nozzle because of internal flow separation of overexpanded gases.

**Regeneratively cooled:** A cooling technique in which propellants, usually fuel, are utilized to remove heat from the inner wall of a combustor in a heat exchange process.

**Thrust chamber assembly:** An assembly consisting of the main injector, combustion chamber, and nozzle. Depending upon fabrication techniques, the combustion chamber and nozzle can be separate components or combined into a single component.

**Vacuum:** Conditions where atmospheric pressure can be considered to be 0.0 psia.

## References

Gordon, S., McBride, B., and Zeleznik, F. 1984. *Computer Program for Calculation of Complex Chemical Equilibrium Compositions and Applications, Supplement I<sup>3</sup>/<sub>4</sub>Transport Properties*. NASA Technical Memorandum 86885, National Aeronautics and Space Administration, Office of Management, Scientific and Technical Information Program.

Johnson, D. 1993. *Terrestrial Environment (Climatic) Criteria Guidelines for Use in Aerospace Vehicle Development, 1993 Revision*. NASA Technical Memorandum 4511, National Aeronautics and Space Administration, Office of Management, Scientific and Technical

Information Program.

Ryan, R. 1993. *Structural Design/Margin Assessment*. NASA Technical Paper 3410, National Aeronautics and Space Administration, Office of Management, Scientific and Technical Information Program.

Sutton, G. P. 1992. *Rocket Propulsion Elements*, 6th ed. John Wiley & Sons, New York.

## **Further Information**

- American Institute of Aeronautics and Astronautics (AIAA)
- American Society of Mechanical Engineers (ASME)
- National Space Society
- Marshall Space Flight Center—Central Technical Library (Phone: 205-544-4524)

Hale, F. J. "Aircraft Performance and Design"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Aircraft Performance and Design

---

## 177.1 Aircraft Forces and Subsystems

The Atmosphere • The Aerodynamic Forces • The Propulsion Subsystems • The Weight Fractions

## 177.2 Level Flight

## 177.3 Climbing Flight

## 177.4 Turning Flight

### Francis Joseph Hale

North Carolina State University

Two major considerations in the performance and design of an aircraft are its range and maneuverability. *Range* deals with how far the aircraft can fly with a given fuel load and given payload with specified flight conditions. *Maneuverability* treats turning rates, rates of climb, and acceleration. A performance analysis starts with a knowledge of the physical characteristics of an aircraft, either actual or hypothetical, and determines how the aircraft will fly under various flight conditions. The converse of the performance analysis is the design of an aircraft to meet a set of operational requirements and constraints. Design, however, is an iterative process, and the first few tries, which yield major aircraft characteristics, may be referred to as conceptual or feasibility designs. The performance analysis and preliminary design processes are often intertwined in a series of iterations and it may be difficult to distinguish between the two.

The flight path and behavior of an aircraft are determined by the interaction between its characteristics and those of the environment in which it is operating—namely, the atmosphere. The aircraft characteristics can be categorized as its physical characteristics, such as shape, mass, volume, and surface area; the characteristics of the propulsion, guidance, and control subsystems; and the structural characteristics, such as loading and temperature limitations and the stiffness (rigidity) of the structure. The environment affects the flight of an aircraft through the surface forces and the field force, the acceleration of gravity, which appears as *weight* and is a function of the mass of the aircraft. The surface forces are the aerodynamic forces (lift, drag, and side force), which are strongly dependent upon the shape and surface area of the aircraft and the properties of the atmosphere. The side force also strongly influences the performance of the propulsion systems.

Aircraft operational units are still principally English and will be given preference over SI units in this article with distance in nautical miles (nmi), airspeed in knots (kt) and feet per second (ft/s), force and weight in pounds (lb), and pressure in lb/ft<sup>2</sup>, where 1 nmi = 1.15 mi = 1.84 km; 1 kt = 1 nmi/h = 1.15 mi/h = 1.84 km/h = 0.51 m/s; 1 ft/s = 0.59 kt; 1 lb = 4.448 N = 0.4535 kg; and 1 lb/ft<sup>2</sup> = 47.87 Pa.



## 177.1 Aircraft Forces and Subsystems

### The Atmosphere

Although the standard atmosphere comprises a number of concentric layers surrounding the earth, only the first two layers of the lower atmosphere are of interest. The *troposphere* starts at sea level and is characterized by a decreasing ambient temperature. At 36 089 ft (11 000 m) above sea level (the *tropopause*), the temperature becomes and remains essentially constant in the *stratosphere* until reaching an altitude of 82 021 ft (25 000 m). In the troposphere, as the atmospheric temperature decreases, so does the local speed of sound  $a$ , thus affecting the **Mach number**. The sonic velocity remains constant in the isothermal stratosphere. Table 177.1 is an abbreviated listing, as a function of the altitude  $h$ , of the two atmospheric ratios of most interest—namely, the *density ratio*  $\sigma$  ( $\rho/\rho_o$ ), where  $\rho$  is the atmospheric density at altitude  $h$  and  $\rho_p$  is the sea level density ( $\rho = \rho_o$ ), and the *sonic ratio*  $a^*$  ( $a/a_o$ ). The standard-day sea level values of  $\rho_o$  and  $a_o$  are given below Table 177.1.

**Table 177.1** Standard Atmosphere Ratios

Altitude (1000 ft)	0	10	20	30	36	40	50	60
Density ratio $\sigma$	1.000	0.738	0.533	0.374	0.297	0.246	0.152	0.094
Sonic ratio $a^*$	1.000	0.965	0.929	0.821	0.867	0.867	0.867	0.867

$$\rho_o = 23.769 \cdot 10^{-4} \text{ lb-s/ft}^2; a_o = 1116 \text{ ft/s}$$

A commonly used exponential approximation for the density ratio  $\sigma$  is

$$\sigma = \rho/\rho_o = \exp(-h/\beta) \quad (177.1)$$

where  $\beta$  is an empirical factor with a value of 23 800 ft (7254 m). Because the atmosphere is assumed to be an ideal gas, variations in the actual temperature and pressure will produce appropriate changes in the actual density. Consequently, on a hot day or with a below-standard barometric pressure, the density will be lower than the standard value for that altitude and the aircraft and propulsion system will perform as though at a higher altitude.

### The Aerodynamic Forces

In coordinated flight, the two principal aerodynamic forces are the *lift*  $L$  and the *drag*  $D$ . The lift is perpendicular to the velocity (*airspeed*) of the aircraft and has the primary function of compensating for the weight of the aircraft. The drag is parallel to the airspeed and resists the motion of the aircraft. The wing is the major source of lift and drag and is designed to maximize the lift while minimizing the drag.

Wing parameters of importance with respect to performance are (1) the wing span  $b$ , the distance from wing tip to wing tip; (2) the average chord  $c_w$ , the distance from the front (leading edge) of the wing to the back (trailing edge); and (3) the wing area  $S$ , the area of one side of the wing to

include the area included by the fuselage. The ratio of the wing span to the average wind chord is the **aspect ratio**  $AR$  of the wing, a measure of the narrowness of the wing. The following relationships are useful:

$$S = bc_w; \quad AR = b/c_w = b^2/S \quad (177.2)$$

The lift and drag forces can be obtained from the following expressions:

$$L = \frac{1}{2}\rho_o\sigma V^2 SC_L = qSC_L; \quad D = \frac{1}{2}\rho_o\sigma V^2 SC_D = qSC_D \quad (177.3)$$

where  $q$  is the dynamic pressure and  $C_L$  and  $C_D$  are the dimensionless lift and drag coefficients.

The **drag polar** is an expression of  $C_D$  as a function of  $C_L$  and is an important performance and design parameter. The parabolic drag polar is an approximation that can be used with many subsonic (and thin-winged) aircraft configurations and is written as

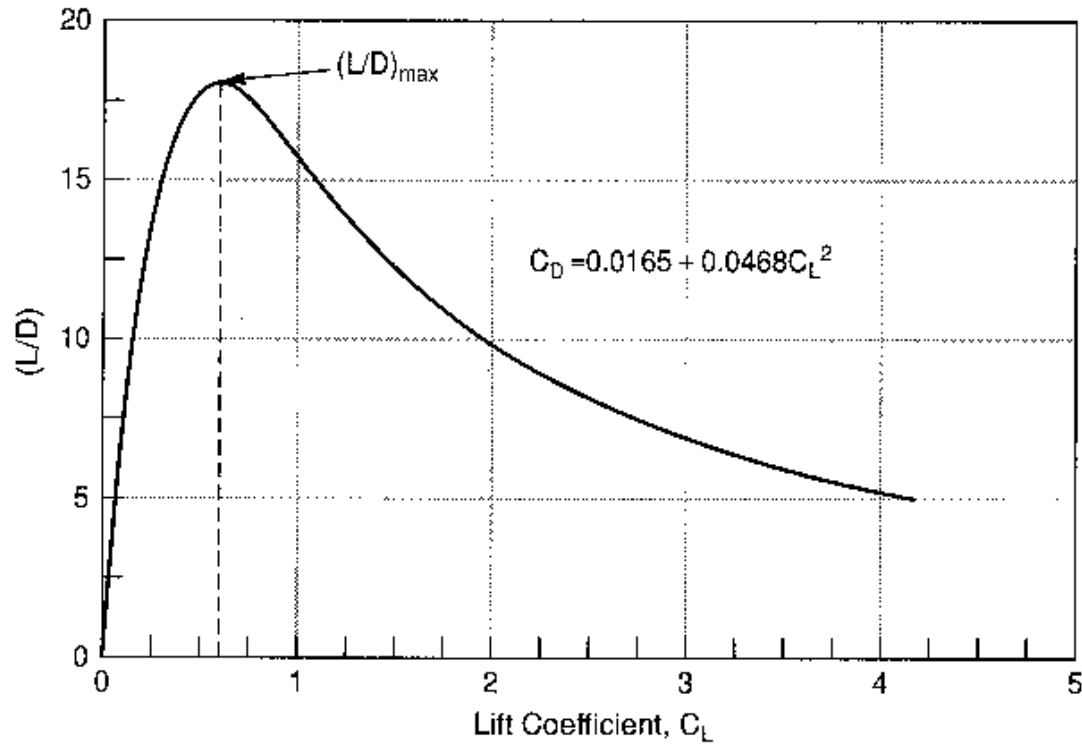
$$C_D = C_{DO} + KC_L^2 \quad \text{where } K = 1/(\pi e AR) \quad (177.4)$$

$C_{DO}$  is the *zero-lift drag coefficient* of the aircraft,  $KC_L^2$  represents the *drag-due-to-lift coefficient*, and  $e$  is the *Oswald span efficiency* (on the order of unity and less). The **lift-to-drag ratio** ( $L/D$ ) or ( $C_L/C_D$ ) represents the aerodynamic efficiency of an aircraft. It has a maximum value that is a design characteristic of the aircraft. This value cannot be exceeded, although the aircraft may, and often does, fly at lower values. The expression for a parabolic drag polar is

$$(L/D)_m = 1/[2(KC_{DO})^{1/2}] \quad (177.5)$$

A large  $(L/D)_m$  calls for low values of  $K$  (high  $AR$  and large  $e$ ) and  $C_{DO}$ . [Figure 177.1](#), a typical plot of the  $(L/D)$  ratio for a specific aircraft and specific parabolic drag polar shows the variation in  $(L/D)$  as  $C_L$  varies. Typical values of  $(L/D)_m$  for several classes of aircraft are: 35 for sailplanes, 18 for Mach 0.8 transports, 7 for supersonic aircraft, and 3 for helicopters.

**Figure 177.1** Typical variation of  $L/D$  with  $C_L$  for a specified drag polar.



## The Propulsion Subsystems

Current aircraft engines are **air breathers** that produce thrust or power, or a combination thereof. The turbojet engine represents a thrust producer, whereas an internal combustion engine in combination with a propeller (a *piston-prop*) represents a power producer. Turbofans, unducted fans, turboprops, and propfans combine the characteristics of both to varying degrees. Turbofans and unducted fans are usually described in turbojet terms and turboprops and propfans in piston-prop terms.

The actual performance and functional relationships of an aircraft engine should be obtained from power plant charts. However, to a first approximation, the *thrust*  $T$  (lb) produced by a turbojet can be considered to be independent of the airspeed and, for a given throttle (percent rpm) setting, to be directly proportional to the atmospheric density, so that

$$T(h) \cong T_o \sigma \quad (177.6)$$

where  $T_o$  is the thrust at sea level. Furthermore, the *fuel consumption rate* (lb/h) is proportional to the thrust, so that

$$dW_f/dt = cT \quad (177.7)$$

where  $c$ , the *thrust specific fuel consumption* (tsfc), has the units of lb/h/lb or h<sup>-1</sup>. Although  $c$  varies somewhat with airspeed and altitude, it may, again to a first approximation, be assumed constant for all altitudes and airspeeds. For current turbine engines, the maximum *thrust-to-engine weight ratio* ( $T/W_e$ )  $\cong$  5–6. Piston-prop engines can be *aspirated* or *turbocharged*, and produce *shaft power* ( $HP$ ), measured in units of horsepower (hp), which is assumed to be independent of the airspeed but to vary with altitude and power setting. The decrease with altitude of the  $HP$  of an aspirated engine is similar to that for the turbojet, but the  $HP$  of a turbocharged engine with a constant power setting remains constant until the *critical altitude* (15–20 000 ft) is reached and then decreases with altitude. The approximations for the density relationships are

$$HP \cong (HP)_0 \sigma \text{ (aspirated)} \quad \text{and} \quad HP \cong (HP)_0 (\sigma/\sigma_{cr}) \text{ (turbocharged)} \quad (177.8)$$

where  $\sigma_{cr}$  is the density ratio at the critical altitude. Figure 177.2 shows the qualitative relationship of the  $HP$  with altitude of the two types of engines. The variation of turbojet thrust is similar to that of the aspirated engine. The propeller converts the shaft horsepower into the *thrust power*  $P$ , which is equal to the product of the thrust and airspeed  $TV$ . The shaft power, thrust power, and thrust are related by the expressions:

$$P = TV = k\eta_p(HP); \quad T = k\eta_p(HP)/V \quad (177.9)$$

where  $\eta_p$  is the propeller efficiency (on the order of 80 to 85% for a well-designed, constant speed propeller) and  $k$  is a conversion factor with a value of 326 when  $V$  is in knots. Notice that for a given horsepower, the thrust power of a piston-prop is independent of the airspeed. However, the available thrust is inversely proportional to  $V$ , decreasing as the airspeed increases, whereas the thrust of a turbojet is constant and the thrust power ( $TV$ ) increases with  $V$ . Because of these differences, propeller and jet aircraft fly differently for best performance. The fuel consumption rate (lb/h) of a piston-prop is proportional to the  $HP$  being delivered, so that

$$dW_f/dt = c^*(HP) = c^*TV/(k\eta_p) \quad (177.10)$$

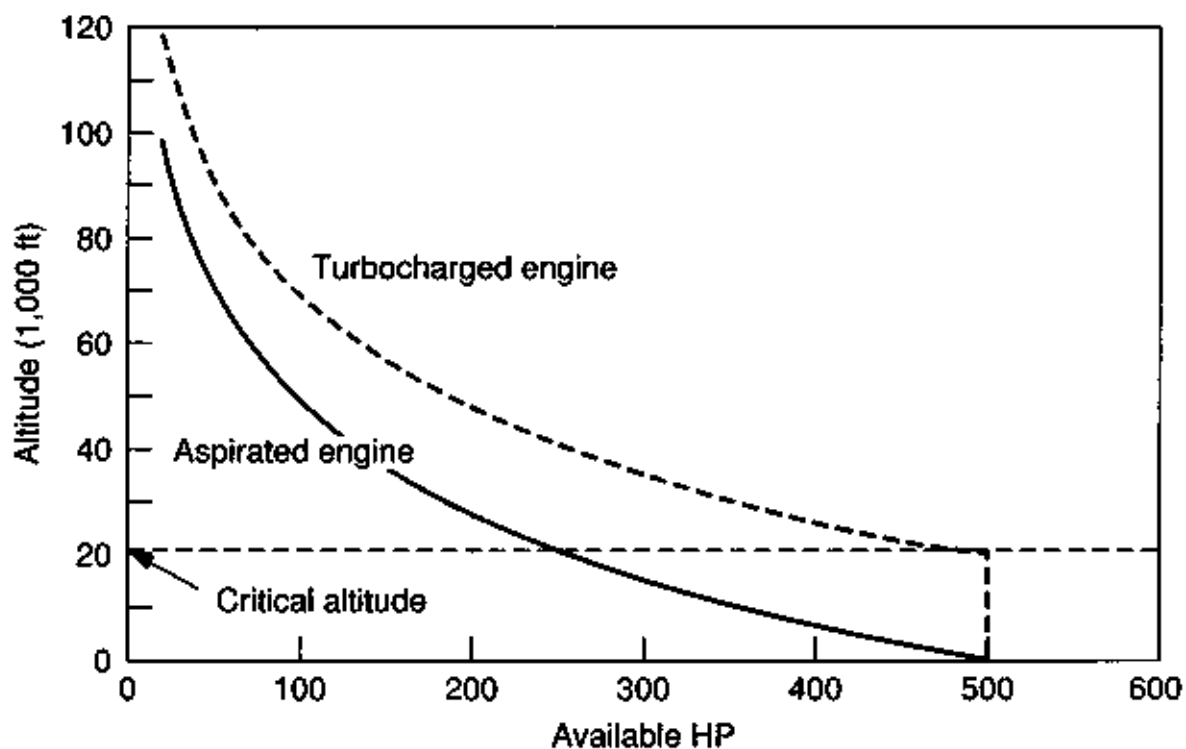
where  $c^*$ , the *horsepower specific fuel consumption* (hpsfc) with units of lb/h/hp, has the same variations as the tsfc and will also be assumed to be constant. Current piston-prop engines are relatively small (less than 1000 hp) because they are the heaviest of all the engines with  $(HP)_m/W_e \cong 0.5$  hp/lb of engine weight. The specific fuel consumption is an extremely important performance parameter. Some typical values, all expressed as an equivalent tsfc (lb/h/lb), are

Rocket engines	10
Ramjets	3
Turbojets (with afterburner)	2.5

Turbojets	0.9–1.0
High-bypass turbofans	0.6–0.8
Turboprops	0.5–0.6
Piston-props	0.4–0.5

The order of this list indicates the relative airspeed regime of the flight vehicles in which these engines are used. Piston-props are used in aircraft with airspeeds on the order of 220 kt or less; turboprops have higher speeds up to Mach 0.7 or so; turbofan engines for airspeeds up to M 0.85; and so on up the speed ladder with turbojets (and very low-bypass turbofans) in supersonic aircraft, ramjets for M 3.0 and higher, and rocket engines in missiles and space boosters. Furthermore, the piston-prop engine is the least expensive and the heaviest of the engines, and the cost increases and the weight decreases as the list is ascended.

**Figure 177.2** Variation of piston-prop *HP* with altitude, both aspirated and turbocharged.



## The Weight Fractions

*Weight* may well be the most important consideration in the design and performance of an aircraft.

Every extra pound of weight is accompanied by an increase in the wing area and in the thrust and fuel required, all leading to a further increase in the aircraft weight and adversely affecting the performance and costs (both initial and operating) of the aircraft. **Weight fractions** are useful in performance and design analyses and are obtained by expressing the gross (total) weight of the aircraft  $W_o$  as the sum of the weights of the major components and then dividing each weight by the gross weight. Using a simple weight breakdown made up of the *structural weight*  $W_s$ , the *engine weight*  $W_e$ , the *fuel weight*  $W_f$ , and the *payload weight*  $W_{PL}$ , the *gross weight*  $W_o$  of the aircraft is the sum of these component weights. The weight fractions are obtained by dividing through by  $W_o$  to obtain

$$1 = W_s/W_o + W_e/W_o + W_f/W_o + W_{PL}/W_o \quad (177.11)$$

With this breakdown,  $W_s$  includes not only the weight of the structure, but also the weight of everything not included in the other categories. It includes the weight of all the equipment and landing gear, for example, and even the weight of the flight and cabin crews when appropriate. Aircraft manufacturers often lump  $W_s$  and  $W_e$  together into the *operational empty weight* and combine  $W_f$  and  $W_{PL}$  into the *useful load*. In Eq. (177.11), the sum of the individual fractions must always be unity. For a turbojet,  $W_e/W_o = (T_m/W_o)/(T_m/W_e)$ , and, for a piston-prop,  $W_e/W_o = (HP_m/W_o)/(HP_m/W_e)$ . Values of  $T_m/W_o$  are on the order of 0.25–0.35 for large high-subsonic transports and on the order of 0.1 hp/lb of aircraft weight for the  $HP/W_o$  of the typical piston-prop.

Although the determination of  $W_s$  is quite complex, order of magnitude values for  $W_s/W_o$  can be used to give a feel for its size and significance. As  $W_o$  increases,  $W_s/W_o$  decreases because, for example, minimum volume requirements for cabin and cargo space and fixed equipment weights have a higher impact on the structural weight fractions of the smaller aircraft. For large subsonic transports,  $W_s/W_o$  values are on the order of 0.45, but smaller aircraft, such as fighters and general aviation aircraft, can have values in excess of 0.55. Note that the structural weight does not include engine weight.

The **payload ratio** (the payload weight fraction) is of major interest inasmuch as it relates the payload weight to the gross weight of the aircraft. For example,  $W_{PL}/W_o = 0.1$  means that ten pounds of aircraft weight is required for each pound of payload (or excess weight). As  $W_{PL}/W_o$  is equal to unity minus the sum of the other weight fractions, any increase in any of these three either decreases the payload weight fraction, thus increasing  $W_o$  or decreasing the payload, or decreases  $W_f/W_o$ , thus decreasing the range. For example, the increase of  $T_m/W_o$  to convert a **CTOL** aircraft to a **VSTOL** aircraft can dramatically reduce the range and payload, or increase the gross weight.

## 177.2 Level Flight

---

For any aircraft in unaccelerated level flight,

$$L = W; \quad T = D; \quad dX/dt = V \quad (177.12)$$

The weight, however, does not remain constant during flight. As fuel is used, the weight decreases at a rate that is the negative of the fuel consumption rate [Eq. (177.7)], so that for a turbojet  $dW/dt = -cT$ . The *instantaneous range* (mileage), the nmi per lb of fuel, is equal to  $dX/dt$  divided by  $dW/dt$  and, with the relationship from Eq. (177.13) that the required thrust  $T = W/(L/D)$ , can be expressed as

$$dX/dW_f = V/cT = V(L/D)/cW \quad (177.13)$$

The last expression of Eq. (177.13) shows that the best mileage (maximum range per lb of fuel) of an aircraft is obtained when the product of  $V$  and  $(L/D)$ , which are coupled, is maximized, and  $c$  and  $W$  are individually minimized. Because  $W$  decreases along the flight path, the mileage improves with time and distance.

Although the mileage at a particular point or instant of time is of interest, the range for a given amount of fuel and set of flight conditions and the amount of fuel required to fly a specified range (from point 1 to point 2) are of greater interest. With  $c$  assumed constant, integrating the mileage over the cruise flight path while holding  $V$  and  $C_L$  [and thus  $(L/D)$ ] constant, yields the *Breguet range equation* in the form:

$$X = [V(L/D)/c] \ln MR = [V(L/D)/c] \ln[1/(1 - \zeta)] \quad (177.14)$$

where the **mass ratio**  $MR = W_1/W_2$  and the *cruise-fuel weight fraction*  $\zeta = \Delta W_f/W_1 = (MR - 1)/MR$ . In order to keep both  $V$  and  $C_L$  constant along the flight path as fuel is used, the altitude must increase so as to keep  $W/\sigma$  constant [see Eq. (177.15)]. The required flight path angle, however, is sufficiently small ( $< 1$  deg) so that the level flight approximation is still valid. Such flight is known as *cruise-climb*. It gives the maximum range for a given fuel load and, for long flights in a controlled area, can be approximated by a series of constant altitude legs with appropriate altitude increases (*stepped altitude flight*).

The airspeed  $V$  and  $C_L$  [and thus  $(L/D)$ ] are related by expressions from Eq. (177.12) and Eq. (177.3)—namely,

$$V = \sqrt{\frac{2(W/S)}{\rho_o \sigma C_L}}; \quad C_L = \frac{2(W/S)}{\rho_o \sigma V^2} = \frac{(W/S)}{q} \quad (177.15)$$

Equation (177.15) shows that for a given  $C_L$ , a high airspeed  $V$  calls for a high **wing loading**  $(W/S)$  and a small  $\sigma$  (a high altitude) for good range performance. Consequently, turbojet aircraft *must* fly high and fast to achieve their best range performance. Reduced flight times and flight above weather are bonuses. The best-range airspeed (and maximum range) of a particular turbojet is 60% greater at 30 000 ft than at sea level, and 80% higher at 35 000 ft. For long-range subsonic transports,  $W/S$  values are on the order of 115–130 lb/ft<sup>2</sup>. Shorter-range and fighter aircraft have lower values to meet requirements for shorter runway lengths or for greater maneuverability.

For a piston-prop, the relationships of Eq. (177.13) still apply, but now the thrust power  $TV$  is of primary interest and equilibrium requires that the thrust power be equal to the drag power

( $TV = DV$ ). The required thrust power  $P \equiv TV = WV/(L/D)$ . With  $dW/dt = -dW_f/dt = -c^*TV/(k\eta_p)$ , the mileage (nmi/lb) of a piston-prop is

$$dX/dW_f = k\eta_p V/(c^*TV) = k\eta_p(L/D)(c^*W) \quad (177.16)$$

Notice that the mileage of a piston-prop is explicitly independent of the airspeed, which must, however, be that required to achieve the desired  $(L/D)$ . Good mileage still calls for a high  $(L/D)$  and low values of  $c^*$  and  $W$  as for the turbojet. Holding  $C_L$  [and thus  $(L/D)$ ] constant yields the piston-prop version of the *Breguet range equation*:

$$X = [326\eta_p(L/D)/c^*] \ln MR = [326\eta_p(L/D)/c^*][\ln 1/(1 - \zeta)] \quad (177.17)$$

with  $V$  in kt. Although the range is explicitly independent of  $V$ ,  $V$  must have the value appropriate to the cruise  $(L/D)$ . Increasing  $V$  by increasing the altitude or  $W/S$  will not increase the range but will decrease the flight time (and increase the power required). Although the maximum range occurs when  $(L/D)$  is at its maximum value, the associated airspeed, especially with an aspirated engine, is usually much lower than the maximum airspeed of the aircraft and the required power for cruise is below the best operating point for the engine. Consequently, cruise of a piston-prop is customarily at 75% of the maximum power available to reduce flight time and favor the engine. With respect to other power plants, whereas a turbojet is a single-flow engine that produces only jet thrust, a *turbopan* is a multistage engine that uses a turbine to drive a multibladed ducted fan that produces thrust power in addition to jet thrust. It has characteristics of both turbojet and propeller aircraft, and its performance (and specific fuel consumption) lies in the region between the two but closer to the turbojet. The turbojet range equation is used.

A *turboprop*, on the other hand, is primarily a power producer, with little or no jet thrust, that uses the piston-prop equations and is replacing the piston-prop because it is lighter and capable of higher speeds. It has a lower  $C_{DO}$ , but a higher sfc, than the piston-prop. *Derating* a turboprop results in an altitude variation of the power that resembles that of a turbocharged piston-prop. The maximum airspeed of a turboprop with a conventional propeller is limited by the drop in  $\eta_p$  at the higher airspeeds. The *propfan* uses a double row of multibladed, variable camber propellers to increase the operating airspeed. The unducted fan (UDF) with an ultrahigh bypass ratio is still driven directly from the turbine, but resembles the propfan in appearance and approaches its performance. The *absolute ceiling* of an aircraft occurs when the maximum available thrust of a turbojet is equal to the required thrust,  $T_m = W/(L/D)_m$ , or when the maximum available power of a piston-prop is equal to the required power,  $(HP)_m = WV/[k\eta_p(L/D)_m]$ . At the ceiling, the rate of climb is zero and the aircraft cannot turn without losing altitude.

## 177.3 Climbing Flight

The steady-state climbing flight equations are

$$T - D - W \sin \gamma = 0; \quad L - W \cos \gamma = 0; \quad R/C = dh/dt = V \sin \gamma \quad (177.18)$$

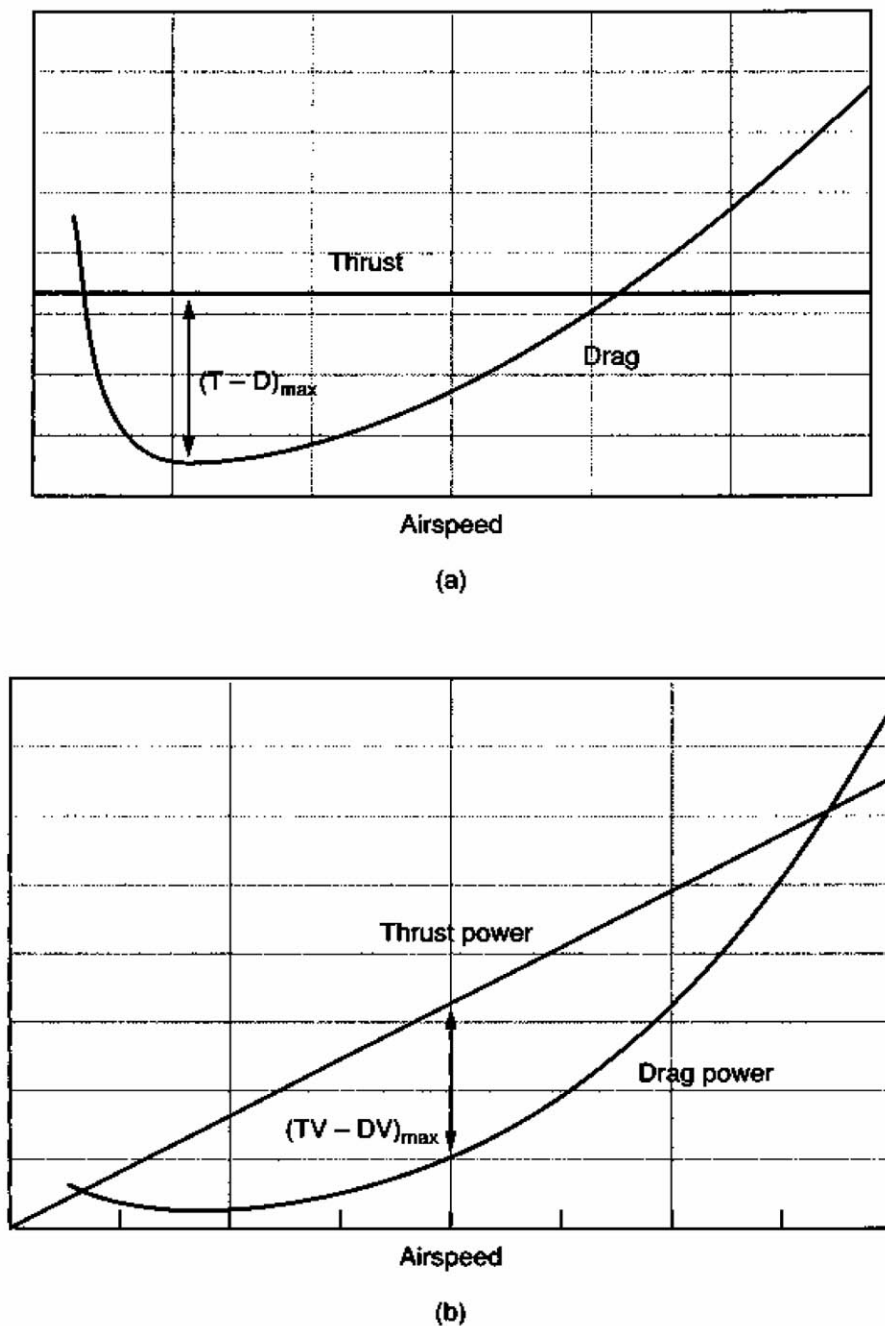
where  $\gamma$ , the *climb (flight path) angle*, is the angle between the horizon and the airspeed vector. The two climb equations of interest are

$$\begin{aligned} \sin \gamma &= (T - D)/W = (T/W) - [\cos \gamma/(L/D)]; \\ R/C &= V \sin \gamma = (TV - DV)/W \end{aligned} \quad (177.19)$$



The first expression in Eq. (177.19) shows that  $\gamma$  is determined by the *excess thrust* (the thrust not needed to overcome the drag) per unit weight, whereas the  $R/C$  is determined by the amount of *excess power* per unit weight. The two climb programs of special interest are *steepest climb*,  $\gamma_m$ , and *fastest climb*,  $(R/C)_m$ . Both have their largest values at sea level, where the excess thrust and power are at a maximum, and both decrease with altitude until  $\gamma$  and  $R/C$  both go to zero at the absolute ceiling. Using the expressions of Eq. (177.19), the climb performance can be obtained either graphically (as sketched in Fig. 177.3 for a turbojet) for various altitudes, weights, power settings, and velocities, or by analytic methods.

**Figure 177.3** Climb relationships for a typical turbojet: (a) climb angle  $\gamma$ ; (b) rate of climb  $R/C$ .



Climb angles for conventional aircraft are not large, as can be seen by dropping the drag term in Eq. (177.20) to obtain the inequality that  $\sin \gamma_m < (T_m/W)$ , which shows that the maximum climb angle for a high-performance turbojet transport, with  $T_m/W = 0.33$ , will be less than  $19^\circ$ . Obviously,  $\gamma$  for  $(R/C)_m$  will be less. As expected, the higher the available thrust for a turbojet and the higher the available power for a piston-prop, the better the climb performance is. In general, the turbojet outclimbs the piston-prop and the climb airspeeds are higher.

## 177.4 Turning Flight

One aspect of maneuverability is the ability to turn. The equations governing steady state turns in the horizontal plane are

$$T = D; \quad L = W / \cos \phi; \quad \dot{\chi} = \frac{g \tan \phi}{V} \quad (177.20)$$

where  $\phi$  is the *bank angle* (deg),  $\dot{\chi}$  is the *turning rate* (rad/s),  $V$  is the airspeed (ft/s), and  $g$  is the acceleration of gravity ( $32.2 \text{ ft/s}^2$ ). The *load factor* is defined as  $n = L/W$  ( $g$ 's). From Eq. (177.20) and Eq. (177.3),

$$\begin{aligned} n &= \frac{L}{W} = \frac{1}{\cos \phi} = \left( \frac{T}{W} \right) (L/D); \\ V &= \sqrt{\frac{2n(W/S)}{\rho_o \sigma C_L}} = \sqrt{\frac{2(W/S)}{\rho_o \sigma C_L \cos \phi}} \end{aligned} \quad (177.21)$$

When  $n > 1 g$ , as in a turn, the airspeed is less than that for wings level flight with the same power setting and the stall speed also decreases. It can be shown that the best turning performance (tightest turn and fastest turn) for both types of aircraft occurs at low speed (on the edge of a stall) and at low altitude while using maximum thrust or power.

### Defining Terms

**Air breathers:** Engines that obtain oxygen from the atmosphere.

**Aspect ratio:** Wingspan divided by the average chord length.

**CTOL:** Conventional (horizontal) takeoff and landing aircraft.

**Drag polar:** Drag coefficient–lift coefficient relationship.

**Lift-to-drag ratio ( $L/D$ ):** Aerodynamic efficiency of an aircraft.

**Mach number:** Airspeed divided by local sonic velocity.

**Mass ratio  $MR$ :** Aircraft weight at start of cruise divided by aircraft weight at end of cruise.

**Payload ratio:** Payload weight divided by gross weight of an aircraft.

**V/STOL:** Aircraft capable of both vertical and short takeoffs and landings.

**Weight fraction:** Component weight divided by aircraft weight.

**Wing loading (W/S):** Aircraft weight divided by wing area (lb/ft<sup>2</sup>).

## References

Anderson, J. D. 1985. *Introduction to Flight*, 2nd ed. McGraw-Hill, New York.

Hale, F. J. 1984. *Introduction to Aircraft Performance, Selection, and Design*. John Wiley & Sons, New York.

Houghton, E. L. and Brock, A. E. 1970. *Aerodynamics for Engineering Students*. Edward Arnold, London.

McCormick, B. W. 1979. *Aerodynamics, Aeronautics, and Flight Mechanics*. John Wiley & Sons, New York.

Raymer, D. P. 1989. *Aircraft Design: A Conceptual Approach*. AIAA, Washington, DC.

## Further Information

*Aerospace America* is a monthly publication of the American Institute of Aeronautics and Astronautics (AIAA) that presents current technical and political activity as well as a look at the future. For subscription and membership information, call (202) 646-7400.

*Aerospace Engineering* is a monthly publication of the Society of Automotive Engineers (SAE) that features articles on technical specialties and new technology. For subscription and membership information, call (412) 776-4841, fax: (412) 776-9765.

*Air & Space* is a bimonthly publication of the Smithsonian Institution that contains a mix of popular and quasi-technical articles written for the general public, but with items of interest for the more technical reader. For subscription and membership information, contact Air & Space/Smithsonian, P.O. Box 53261, Boulder, CO, 80322-3261.

*Aviation* is a bimonthly publication devoted to stories about aircraft (mostly historical), and about incidents of historical interest, including WWII aircraft, pilots, and tales. For subscription information, contact Aviation, P.O. Box 368, Mount Morris, IL, 61054-7739.

*Aviation Week & Space Technology* is a weekly publication by McGraw-Hill that is considered by many readers to be the most important and authoritative aerospace publication of both broad and detailed interest. For subscription information, contact Aviation Week & Space Technology, P.O. Box 503, Highstown, NJ, 08520-9899.

Wallace T. Fowler. "Spacecraft and Mission Design"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Spacecraft and Mission Design

---

## 178.1 Spacecraft Environments

## 178.2 Fundamental Principles

## 178.3 Spacecraft/Mission Categories

Launch Vehicles • Sounding Rockets • Earth-Orbiting Spacecraft • Lunar Spacecraft • Interplanetary Spacecraft • Entry Vehicles • Landers • Returning Spacecraft

## 178.4 Spacecraft Subsystems

Structure Subsystem • Power Subsystem • Attitude Control Subsystem • Telecommunications Subsystem • Propulsion Subsystem • Thermal Control Subsystem • Spacecraft Mechanisms • Launch Vehicles • Upper Stages • Entry/Landing Subsystems • Subsystem Integration and Redundancy

## 178.5 Spacecraft/Mission Design Process

### Wallace T. Fowler

*University of Texas, Austin*

The design of a spacecraft and the design of its mission are closely related parts of the same iterative process. The mission concept usually comes first, but new missions for existing spacecraft designs have also been developed. The key to any successful design is the development of a good set of requirements. In any case, the mission concept is often changed by design factors, and the spacecraft design is driven by mission requirements. Effort spent in defining and refining requirements is crucial to good design.

The design process starts with either mission concept or spacecraft and arrives at a complete design. However, every design process is unique, and no design is optimal. Factors will always arise that do not seem to fit. When this happens, the mission requirements or the design process should be modified to accommodate or eliminate the unique factors.

The match between spacecraft and mission is determined by a number of factors. The two most important preliminary design factors are the mass of the mission payload and the amount of velocity change,  $\Delta V$ , that must be provided to the payload to carry out the mission. These two factors dictate much of the mission and spacecraft design, being the primary drivers for the mass of the booster assembly. Other primary design factors are the duration of the mission and the destination to which the payload is being sent. A two-day mission on which the spacecraft stays near the earth will require a very different hardware/mission plan combination than will a multiyear mission to an outer planet. If a spacecraft is manned, its design and mission planning are greatly complicated. Factors of safety that are acceptable for unmanned flight are unacceptable for manned flight. Life support and safety equipment, as well as backups for such systems, dominate the design. One factor that should not be ignored during the design of a spacecraft/mission

combination is operations. A detailed operations plan is as important to a successful space mission as is good hardware and a good mission plan. Inclusion of operations planning in the initial spacecraft and mission design process, including planning for contingencies, has the potential for lowering mission costs, increasing the chances of mission success, and lowering the potential for problems caused by unforeseen operator/hardware interactions.

## 178.1 Spacecraft Environments

---

Every spacecraft must be able to survive in several environments. All those built on earth must be able to survive the earth's atmosphere or be protected from it. The earth's standard atmosphere is described in **Chapter 177**, and atmosphere models for other planets are available in NASA literature. The launch environment, characterized by vibration, noise, g-loads, aerodynamic loads, transition from air to vacuum, and so on, is a major test for spacecraft. The space environment presents another set of problems for the designer. Hard vacuum, radiation, and temperature extremes are common to all missions. Spacecraft that fly through or beyond the Van Allen radiation belts experience more severe radiation hazards than those that stay inside the Van Allen belts. For manned spacecraft, extended flight beyond the protection of the Van Allen belts implies special crew radiation shielding. Orbital debris is also a hazard that must be considered.

Vehicles that reenter the earth's atmosphere or enter the atmosphere of another planet must contend with entry aerodynamics and the accompanying loads and heating. For landers, abrasive dust can be a problem. Finally, if a spacecraft has a crew, its internal environment must support life while its structure protects the habitable volume. Good discussions of the role of environment in spacecraft design can be found in Fortescue and Stark [1992] and Griffin and French [1991].

## 178.2 Fundamental Principles

---

The primary factor linking the spacecraft and the mission scenario is linear momentum change, and this is the key driver in overall sizing of the spacecraft and its booster system. Except for missions dominated by attitude control requirements, the mission scenario that requires the lowest total linear momentum change will be the most efficient in terms of propellant. Thus, in designing mission and spacecraft, it is important to (1) explore many mission scenarios to ensure that the total momentum change required is as low as possible; (2) use staging to discard unneeded mass and reduce the propellant required during each mission phase; (3) employ engine/propellant combinations that produce large momentum changes with relatively small amounts of propellant (i.e., propellants with high exhaust velocities relative to the vehicle); (4) keep all spacecraft component masses as small as possible; and (5) use unmanned spacecraft whenever possible to keep vehicle masses small. It is also important to use space-qualified hardware when available (and appropriate) to reduce cost and risk.

The ability of a propulsion system to impart momentum changes to a spacecraft is the velocity of the exhaust with respect to the spacecraft. The key equation is

$$c = I_{sp} g \quad (178.1)$$

where  $c$  is exhaust velocity relative to the vehicle,  $I_{sp}$  is specific impulse, and  $g$  is the acceleration of gravity at the surface of the earth.  $c$  is in m/s (ft/s),  $g$  is in m/s<sup>2</sup> (ft/s<sup>2</sup>), and  $I_{sp}$  is in units of seconds. [Note: the value used for  $g$  in Eq. (178.1) does not change when the spacecraft is in a non-earth gravity field].

Propellant combinations (fuel and oxidizer) are often listed in propulsion references with a corresponding  $I_{sp}$  value. These values imply an appropriate engine in which to combine the propellants efficiently. In spacecraft design studies, trade-offs between propellant storability and  $I_{sp}$  are common.  $I_{sp}$  values for most large solid rockets are just under 300 seconds. The  $I_{sp}$  value for the space shuttle main engines (liquid oxygen–liquid hydrogen) is more than 400 seconds.

A second basic equation, which relates the mass of the spacecraft assembly at the  $i$ th stage in the mission (the total mass to be propelled at that point in the mission) to propellant mass that produces the velocity change, is

$$\Delta V_i = c_i \ln(m_{oi}/m_{fi}) \quad (178.2)$$

where  $\Delta V_i$  is the velocity increment provided to the spacecraft assembly by the propellant in the  $i$ th stage,  $c_i$  is the exhaust velocity of the propulsion system in the  $i$ th stage relative to the spacecraft assembly,  $m_{oi}$  is the mass of the spacecraft assembly prior to burning the propellant in the  $i$ th stage, and  $m_{fi}$  is the mass of the spacecraft assembly after the  $i$ th stage propellant is expended. The mass of the propellant burned in the  $i$ th stage is the difference between  $m_{oi}$  and  $m_{fi}$ .

Sequential staging is a very important concept. Mass that is no longer needed should be discarded as soon as is practical in order to avoid using propellant to change the momentum of useless mass. The following equation shows the total  $\Delta V$  for a two-stage vehicle.

$$\Delta V = c_1 \ln(m_{o1}/m_{f1}) + c_2 \ln(m_{o2}/m_{f2}) \quad (178.3)$$

When staging occurs, excess hardware (tanks, engines, pumps, structure) is dropped after a stage burns out and before the next stage begins its burn. Thus,  $m_{f1}$  will be larger than  $m_{o2}$ . This difference is one of the most important factors in space vehicle design and performance. A good discussion of staging can be found beginning on page 603 of Wertz and Larson [1991].

Equation (178.2) does not apply directly when the two different types of engines with different  $I_{sp}$  values are being used simultaneously. The space shuttle first stage (with solid rocket boosters and main engines burning simultaneously) and the Delta booster (with strap-on solid rockets around a liquid rocket core vehicle) are examples of this. A simple modification of Eq. (178.2) applies, however. This equation looks like Eq. (178.2) but has a significant difference in the definition of the exhaust velocity. For such vehicle mission stages,

$$\Delta V = c^* \ln(m_o/m_f) \quad (178.4)$$

where  $c^*$  is a weighted combination of the exhaust velocities of the two types of engines being used. Specifically,

$$c^* = f_1^* c_1 + f_2^* c_2 \quad (178.5)$$

where  $c_1$  and  $c_2$  are the exhaust velocities for the two engine types being used, and  $f_1$  and  $f_2$  are the respective mass flow rate fractions for the two engine types (note that  $f_1 + f_2 = 1$ ). Equations (178.4) and (178.5) allow characteristic  $\Delta V$  values to be calculated for vehicle stages involving parallel burning of dissimilar engine types.

## 178.3 Spacecraft/Mission Categories

---

### Launch Vehicles

Launch vehicles and launches are a common element/system for almost all spacecraft. The choice of launch vehicle is determined by the mission, the mass to be boosted to orbit, the required  $\Delta V$ , the compatibility between launcher and payload, and the overall cost of the launch (launcher plus launch services). The *International Reference Guide to Space Launch Systems* [Isakowitz, 1991] is an excellent source of information on launchers. A good rule of thumb to keep in mind is that for eastward launches into low earth orbit (LEO) from a latitude of about  $30^\circ$ , between 9300 m/s (30 500 ft/s) and 9450 m/s (31 000 ft/s) total  $\Delta V$  must be supplied by the booster. (This accounts for gravity and drag losses as well as increasing the potential and kinetic energy of the spacecraft.) The time from launch to orbit insertion will depend on the booster acceleration capabilities and constraints, but will generally be less than 8 minutes.

### Sounding Rockets

Sounding rockets are usually small, and their missions are characterized by short mission duration and short times out of the atmosphere of the earth (or another planet). Their objectives are usually scientific, their equipment is powered by batteries, and they usually possess a recovery system for the instrumentation package (or perhaps the entire vehicle). Sounding rocket payloads usually consist of structure, instrumentation, a control system, a power system, communications, and a recovery system. The mass and  $\Delta V$  requirement for a sounding rocket depend primarily on the mass to be boosted, the propellant combination to be used, and the altitude to which the payload is to be sent.

### Earth-Orbiting Spacecraft

There are many varieties of earth-orbiting spacecraft. Examples include the original *Sputnik* and *Explorer* spacecraft, the space shuttle, and communications satellites such as *Intelsat*. There are resource mappers, weather satellites, military satellites, and scientific satellites. There are many active satellites in orbit around the earth today. Useful satellite lifetimes range from a few hours to years. It is likely that the *Lageos* satellite, a passive laser reflector, will be providing useful data for at least 100 years. Typical earth-orbiting spacecraft systems are the structure, power system, thermal management system, communications, sensors, computation/data storage, propulsion, and possibly a deorbit/entry system. Launch dates, orbits, and missions of a wide variety of



earth-orbiting spacecraft have been summarized in compact form in the *TRW Space Log* [Thompson, 1987]. Especially useful are the summary documents in this series.

## Lunar Spacecraft

A special category of earth-orbiting spacecraft meriting separate mention are lunar spacecraft. Such spacecraft must climb out of the earth's gravity well, requiring a  $\Delta V$  of more than 3100 m/s (10 200 ft/s) to reach the moon starting in LEO. Spacecraft that orbit the moon require about 1070 m/s (3500 ft/s) to enter an orbit about the moon and 2070 m/s (6800 ft/s) to land. Special thermal considerations are necessary for vehicles designed to rest on the lunar surface because of the slow rotation rate of the moon (the lunar day and night are 14 earth days each). Lunar spacecraft have the systems typical of earth-orbiting spacecraft, with the possible exception of the deorbit/entry system.

## Interplanetary Spacecraft

Interplanetary spacecraft have much in common with lunar vehicles, but they also have many unique characteristics. Spacecraft targeted to the outer planets have longer mission times and different power supplies than those that remain in the inner solar system. The inner planets (Venus and Mercury) present particular problems for the designer of landers. The atmosphere of Venus is hot, dense, and corrosive. The surface of Mercury, except near the poles, sees wide swings in temperature. A point on the equator of Mercury sees 88 earth days of sunlight followed by 88 earth days of darkness. Obviously, structure, materials, and thermal systems are important for such spacecraft. Injection onto a minimum-energy transfer to Mars requires 3600 m/s (11 800 ft/s). Orbiting Mars will require another 2100 m/s (7000 ft/s).

## Entry Vehicles

Entry vehicles are spacecraft (or spacecraft segments) designed to use aerodynamic lift and drag to slow or turn the spacecraft, leading to an orbital plane change, lowering of an orbit, or a descent into the planetary atmosphere. Aerothermodynamic considerations (both stagnation point temperature and integrated heat load) and deceleration loading dominate the design of such vehicles. Aerodynamic control surfaces are the primary means of trajectory shaping, which in turn controls the heating and deceleration. Survivability is a primary design consideration. An entry vehicle is a special type of hypervelocity vehicle, usually a glider (see **Chapter 177** for additional information).

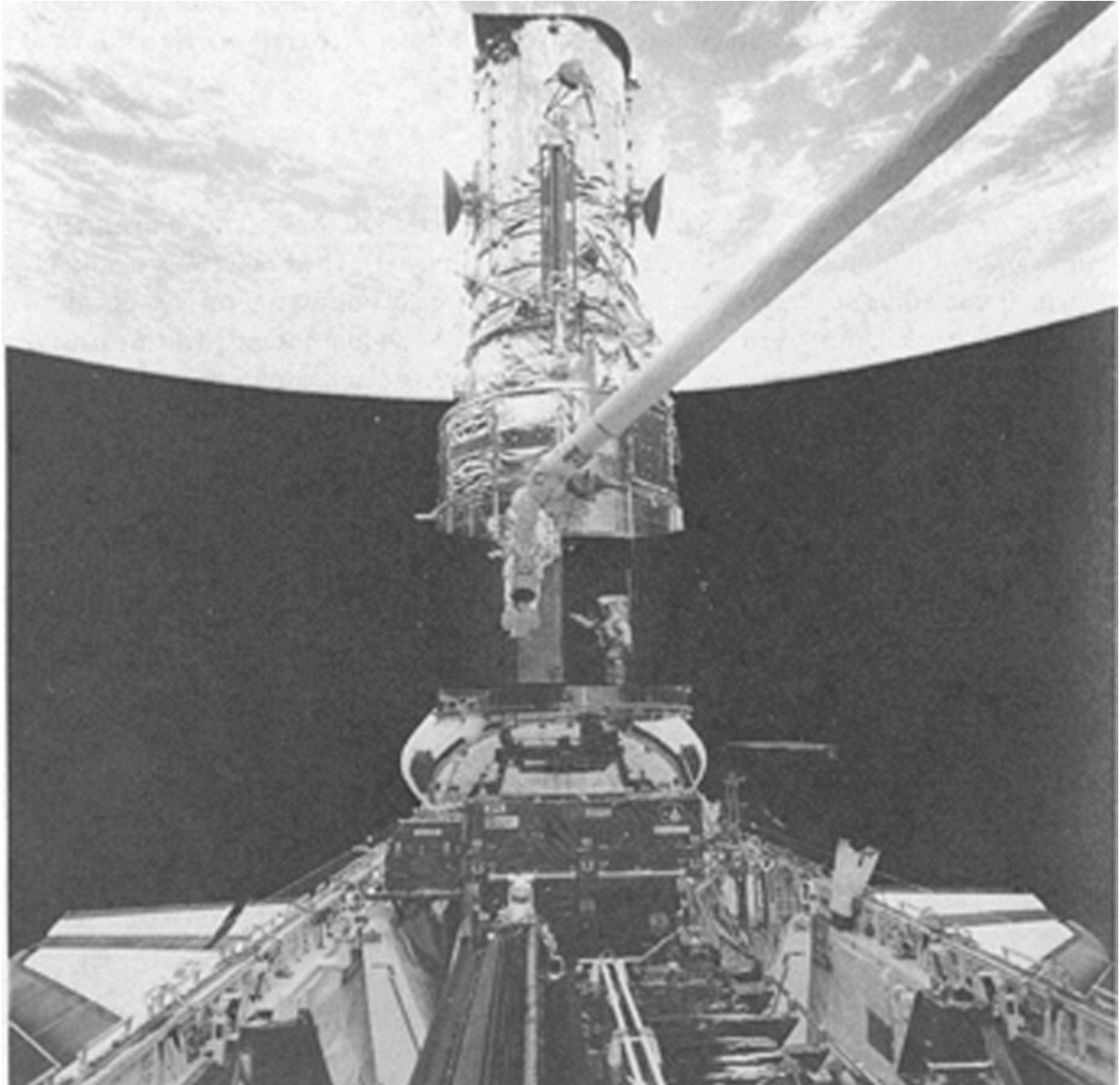
## Landers

Landers are a special class of spacecraft. Some landers are also entry vehicles and some are not, depending on whether the body on which the landing is to take place has an atmosphere. Landings on targets with atmospheres can involve parachutes, cushions, shock absorbers, and so on. Landings on airless bodies usually involve either rocket braking or braking via penetration of the

surface of the target body.

## **Returning Spacecraft**

Spacecraft that return to the planet of origin are a very special case. Vehicles that fly to earth orbit and then deorbit and land on earth, although complex, are much simpler than vehicles that fly from the earth to a target, land, operate, launch, return to earth, enter the atmosphere, and are recovered. For such vehicles, mass is by far the most important driver for all design decisions, and repeated staging is the primary feature of the overall design. Often the redesign of the staging process can greatly simplify an otherwise complex vehicle and mission (the lunar orbit rendezvous used in the Apollo program is an excellent example).



#### HUBBLE TELESCOPE AND COSTAR REPAIR

Astronaut Thomas D. Akers maneuvers inside the bay of the Hubble Space Telescope (HST) which will house the Corrective Optics Space Telescope Axial Replacement (COSTAR). Akers is assisting astronaut Kathryn C. Thornton with the installation of the 640-pound instrument. Thornton, anchored on the end of the Remote Manipulator System (RMS) arm, is partially visible as she prepares to install the COSTAR. The 70 mm scene is backdropped against the earth and the blackness of space. The two spacewalkers participated in the mission's second and fourth (of five) extravehicular activities (EVAs) designed to service the giant telescope. This photograph was made with a handheld Hasselblad camera from *Endeavour's* cabin. (Photo courtesy of National Aeronautics and Space Administration.)

## 178.4 Spacecraft Subsystems

---

Once the overall mission and spacecraft requirements have been defined, the process of defining requirements for various spacecraft subsystems can begin. The integration of the various subsystems into a harmoniously functioning spacecraft can be very difficult. One way to reduce the difficulties in spacecraft integration is to spend sufficient time and effort *early* in carefully and completely defining the interfaces between the subsystems of the spacecraft. Time spent in defining interfaces will repay itself many times as the design progresses. The following is a brief discussion of the major design considerations of the principal spacecraft subsystems. Chetty [1991], Fortescue and Stark [1992], and Griffin and French [1991] provide excellent discussions of various spacecraft subsystems.

### Structure Subsystem

The structure of the spacecraft holds the components together, provides load paths for launch loads, maneuver loads, and so on, and is usually an integral part of the thermal control system or the electrical system for the spacecraft. Design constraints that the structural subsystem must usually meet include stiffness requirements, placing principal inertia axes in preferred directions, sustaining mission loads, and serving as a ground for all spacecraft electrical equipment. Spacecraft structures can be metallic, composite, or ceramic. Structural mass is generally 5 to 20% of the overall spacecraft launch mass.

### Power Subsystem

The type of power system chosen for a spacecraft is strongly dependent on the mission of the spacecraft and the power demands of the equipment to be carried. Short-duration missions often use batteries. Longer-duration missions that go no farther from the sun than Mars can use photoelectric power (solar cells). Long-duration, low-power missions on which the spacecraft go very far from the sun (e.g., *Voyager*) use radioisotope thermoelectric generators (RTGs). Other potential power sources are solar thermionic systems, chemical dynamic systems, and nuclear power systems.

### Attitude Control Subsystem

Most spacecraft require attitude stabilization to point antennas, properly orient solar arrays, point sensors, and orient thrusters. Attitude control consists of sensing the current attitude of the spacecraft and applying torques to reorient the spacecraft to a desired attitude.

### Three-Axis Stabilization

The most complex and precise spacecraft attitude control is three-axis stabilization accomplished using momentum wheels or reaction wheels. Attitude thrusters are also used for three-axis

stabilization, either as the primary attitude control mechanism or to desaturate the momentum wheels. Three-axis stabilization systems can be fast and accurate. For example, the attitude control system on the Hubble Space Telescope, which uses reaction wheels, can point the optical axis of the telescope to within 0.007 arc seconds of the desired direction.

### **Spin Stabilization**

Spin stabilization is much simpler and less precise than three-axis stabilization. Spin-stabilized spacecraft must be rigid and spinning about the principal axis with the largest mass moment of inertia. The first U.S. orbiting satellite, *Explorer 1* (launched 31 January 1958), provided an early demonstration of the need for spinning spacecraft to be rigid. *Explorer 1* was a long slender cylinder spinning about its long axis. It possessed flexible whip antennas that dissipated energy and caused the spacecraft to tumble within a day of orbit insertion.

### **Dual-Spin Spacecraft**

Some spacecraft have a spinning section and a despun section. Such spacecraft must have systems to maintain the attitude of the despun section and to control the direction of the spin vector. Reactions between the spinning and despun sections are major design and operational considerations.

### **Gravity Gradient Stabilization**

Spacecraft can use the gradient of the gravitational field of a planet to provide restoring torques and maintain the spacecraft in an orientation with the spacecraft's long axis pointing along the local vertical. Spacecraft designed to employ gravity gradient stabilization usually feature mass concentrations at the ends of a long boom that lies along the local vertical. Gravity gradient stabilization works best with near-circular orbits and would not work on even moderately elliptic orbits. A pointing accuracy of about  $1^\circ$  with some wobble is achievable with gravity gradient stabilization of earth satellites.

### **Magnetic Stabilization**

Spacecraft attitudes can be stabilized or changed using torques developed by running currents through coils (torque rods) mounted in the spacecraft if the spacecraft is orbiting a body with a significant magnetic field. The primary difficulty with this type of attitude stabilization system is keeping track of the relative orientations of the spacecraft axes and the magnetic field lines as the spacecraft moves along in its orbit. Magnetic stabilization is slow and coarse.

### **Solar Radiation Stabilization**

The pressure of solar radiation falling on a spacecraft can be used to create torques. Spacecraft that employ solar radiation stabilization are large and lightweight.

## **Telecommunications Subsystem**

The primary function of the telecommunications subsystem is to receive, process, and transmit electronic information. Some satellites (communications, intelligence, and weather satellites) are designed to meet one or more telecommunications mission objective. The telecommunications

subsystem often demands a significant amount of electrical power, and this demand often drives the design of the power system. The spacecraft computer/sequencer can be considered part of the telecommunications system or a separate system. Chetty [1991], Fortescue and Stark [1992], and Griffin and French [1991] provide excellent discussions of spacecraft telecommunications.

## **Propulsion Subsystem**

The spacecraft propulsion subsystem will be considered separately from the launch vehicle and upper stages. The propulsion subsystem is responsible for orbit maintenance, small orbit changes, attitude control system desaturation, and so on. Spacecraft onboard propulsion systems use chemically reacting fuels and oxidizers, monopropellants and catalysts, cold gas, and ion thrusters. Chetty [1991], Fortescue and Stark [1992], and Griffin and French [1991] provide excellent discussions of propulsion systems.

## **Thermal Control Subsystem**

Spacecraft thermal control is of utmost importance. Maintaining appropriate thermal environments for the various spacecraft is essential to proper component function and longevity. The primary problem with thermal management of spacecraft is that all excess heat must ultimately be radiated by the spacecraft. For many unmanned spacecraft, passive thermal control involving coatings, insulators, and radiators is ideal. For other spacecraft, active thermal control is necessary. Active systems can employ heaters, radiators, heat pipes, thermal louvers, and flash evaporators. Significant spacecraft mass reductions can sometimes be made if the excess heat from one component is conducted to another component that would otherwise require active heating.

## **Spacecraft Mechanisms**

Some spacecraft are mechanically passive and have no mechanisms in their designs. However, even these spacecraft require mechanisms to deploy them from their boosters. Mechanically active spacecraft can feature a wide variety of mechanisms. There are deployment systems for booms and antennas, docking mechanisms, robot arms, spin/despin systems, sensor pan and tilt mechanisms, gyroscopes, reaction wheels (also considered as part of the attitude control system), landing legs, and airlocks. The types of mechanisms depend on the mission of the spacecraft.

## **Launch Vehicles**

The launch vehicle, or booster, is an integral part of most spacecraft systems. The *International Reference Guide to Space Launch Systems* [Isakowitz, 1991] covers the capabilities of most currently available launch vehicles in detail. The rotation of the earth has a major effect on launch vehicle performance. The eastward motion of the launch site must be considered when computing the payload that a launch vehicle can place in orbit. The launch azimuth and the latitude-dependent eastward velocity of the launch site can be combined to determine how much velocity assist is provided by earth's rotation. For westward launches (to retrograde orbits), the rotation of the earth

extracts a penalty, requiring additional velocity to be provided by the booster to make up for the eastward velocity of the launch site.

## Upper Stages

Launch vehicles are often combined with upper stages to form a more capable launch vehicle. Commonly used upper stages include the *IUS*, the *Centaur*, and several types of PAM (payload assist modules). Upper stages are also carried aloft in the payload bay of the space shuttle to boost spacecraft to higher orbits. Designers should always consider the option of using an available upper stage if the mission requires more  $\Delta V$  than can be provided by the basic launch vehicle.

## Entry/Landing Subsystems

Entry subsystems can be considered either as a separate set of devices or as parts of other subsystems. The heat shield can be reusable or ablative, and could be considered to be part of the thermal control subsystem. The retrorockets, if any, might be considered to be part of the propulsion system. Parachutes, landing legs, cushions, wings, aerosurfaces, a streamlined shape, and landing gear, when needed, belong solely to the category of entry and landing subsystems.

## Subsystem Integration and Redundancy

The interplay among spacecraft subsystems is great, and their integration into an efficient, smoothly operating, and reliable spacecraft is an immense challenge. Redundant system elements can provide added reliability if the risk of failure must be lowered. However, redundancy increases initial costs and usually increases overall mass. It is often good to avoid physically redundant elements, choosing physically different but functionally redundant elements instead. System elements that are to be functionally duplicated should be identified as early in the design process as possible.

## 178.5 Spacecraft/Mission Design Process

---

The spacecraft/mission design process can be outlined as follows:

1. Beginning with a set of mission goals, develop a set of mission and spacecraft requirements that must be met in order to achieve the goals. An excellent discussion of requirements and how to develop them is found in Wertz and Larson [1991, pp. 55–78]. Mission goals may change during the design process, and when this happens, the requirements must be reevaluated.
2. Before beginning a methodical approach to the design, it is almost always helpful to bring the design team together and conduct a brainstorming session. For those who are unfamiliar with brainstorming, Adams [1986] is recommended. The brainstorming session should be expected to identify several nonorthodox candidate scenarios or components for the mission and spacecraft.
3. Develop a conceptual model for the spacecraft and the mission, identifying major systems



and subsystems, and most important, the interfaces between systems, both within the spacecraft itself and between the spacecraft and its support elements on the ground and in space. A preliminary mission scenario and rough timeline should also be developed at this time. These conceptual elements should be developed from the mission and spacecraft requirements. This step will allow the definition of major systems and subsystems and will identify major interactions between mission events and spacecraft hardware elements. Preliminary indications of pointing requirements, required sensor fields of view,  $\Delta V$  requirements, allowable spacecraft mass, and required and desired component lifetimes should come from these considerations. Brown [1992], Fortescue and Stark [1992], Griffin and French [1991], and Isakowitz [1991] give information appropriate to this topic.

4. The role of heritage in spacecraft design is important. Systems and technologies that have worked well on past spacecraft are often good candidates for new spacecraft with similar missions. Examine the recent past for similar missions, spacecraft with similar requirements, and so on. Although sufficiently detailed systems overviews of recent spacecraft are often difficult to obtain, the effort necessary to obtain them is almost always worth it. However, be careful to avoid early adoption of a candidate system from earlier spacecraft to avoid being locked into a system that only marginally meets or does not meet design requirements.
5. Once the first four steps have been taken (some more than once), the design team should develop preliminary spacecraft design candidates (spacecraft/scenario/timeline combinations), concentrating on systems and subsystems and their performance, masses, power requirements, system interactions, thermal input/output, and costs. Several candidate designs should be considered at this stage and their predicted relative characteristics compared. These considerations allow the design team to learn more about the requirements and how best to meet them. It is likely that the design team will recycle through some or all of the design process at this point for one or more of the candidate designs. The possibility of developing a new composite design with the best features of several candidate designs should not be ignored.
6. As candidate designs are refined, the design team should develop criteria for use in choosing among candidate designs. The information on the candidate designs should be organized and presented at a preliminary design review. The feedback from those attending the review is valuable in the identification of strengths and weaknesses of candidate designs.
7. At some point, determined by budget, timeline, technical considerations, and the characteristics of the candidate designs, one design concept must be chosen for development. Once the choice is made, the design team should carry out a complete analysis of the chosen design. This analysis should include a detailed mission scenario and timeline, launch vehicle choice,  $\Delta V$  requirements for every maneuver, hardware choices for every system and subsystem, system interaction analyses, manufacturability considerations, component and overall cost analyses, and failure mode effects analyses. Any remaining system/subsystem hardware choices should be made, and the design should be presented at a critical design review. The feedback from the critical design review will often result in major improvements



in the design. Questions arising at the critical design review will often precipitate another pass through the design process.

## Defining Terms

**Mission scenario:** A sequence of events and times that, in their entirety, meet the objectives of the mission.

**Specific impulse:** The amount of thrust force obtained from a fuel/oxidizer/engine combination when a unit weight of fuel is burned in one second. The units of specific impulse are seconds because the force and weight units cancel.

**Subsystem:** A part of an overall system that can be isolated to some degree. Subsystems interact and cannot actually be isolated, but it is convenient to consider subsystems in sequence rather than all at once.

## References

- Adams, J. L. 1986. *The Care and Feeding of Ideas*<sup>3/4</sup>A Guide to Encouraging Creativity, 3rd ed. Addison-Wesley, Reading, MA.
- American Institute of Aeronautics and Astronautics. 1993. *AIAA Aerospace Design Engineers Guide*, 3rd ed. American Institute of Aeronautics and Astronautics, Washington, DC.
- Brown, C. D. 1992. *Spacecraft Mission Design*. AIAA Education Series, American Institute of Aeronautics and Astronautics, Washington, DC.
- Chetty, P. R. K. 1991. *Satellite Technology and Its Applications*. TAB Professional and Reference Books, Blue Ridge Summit, PA.
- Fortescue, P. and Stark, J. 1992. *Spacecraft Systems Engineering*. John Wiley & Sons, New York.
- Griffin, M. D. and French, J. R. 1991. *Space Vehicle Design*. AIAA Education Series, American Institute of Aeronautics and Astronautics, Washington, DC.
- Isakowitz, S. J. (Ed.) 1991. *International Reference Guide to Space Launch Systems*. AIAA Space Transportation Technical Committee, AIAA, Washington, DC.
- Sellers, J. J. 1994. *Understanding Space*<sup>3/4</sup>An Introduction to Astronautics. McGraw-Hill, New York.
- Thompson, T. D. (Ed.) 1987. *TRW Space Log, 1957-1987*. TRW Space & Technology Group, Redondo Beach, CA.
- Wertz, J. R. and Larson, W. J. (Eds.) 1991. *Space Mission Analysis and Design*. Space Technology Library, Kluwer Academic, Boston.

## Further Information

An excellent and inexpensive source of information on the orbital mechanics necessary for mission planning is Bate, R. R., Mueller, D. D., and White, J. E. 1971. *Fundamentals of Astrodynamics*. Dover, New York.

Details on the locations of the planets, their moons, the orbit of the earth about the sun, values of astrodynamical constants, and many other items pertinent to mission planning are found in *The*

*Astronomical Almanac*. Nautical Almanac Office, Naval Observatory, published yearly by the U.S. Government Printing Office, Washington, DC; and Seidelmann, P. K. (Ed.) 1992.

*Explanatory Supplement to the Astronomical Almanac*. University Science Books, Mill Valley, CA.

An excellent overview of the interplay between spacecraft, launch vehicle, trajectory, and operations is given in Yenne, B. (Ed.) 1988. *Interplanetary Spacecraft*. Exeter, New York.

Furr, A. K. "Safety"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



These two men, wearing required safety gear, are removing 55 gallon drums containing hazardous material from an abandoned hazardous waste site in New Jersey. The regulations and standards related to the handling and use of these materials originate and are enforced by government agencies such as the Occupational Safety and Health Administration (OSHA) and the Environmental Protection Agency (EPA). (Photo by Steve Delany. Courtesy of the Environmental Protection Agency.)

# XXVII

## Safety

---

### A. Keith Furr

*Virginia Polytechnic Institute & State University, Retired*

#### 179 Hazard Identification and Control *M. Rahimi*

Hazard Identification Physical Hazards Chemical Hazards Airborne Contaminants Noise Fire Hazards An Engineering Approach to Hazard Control Hazard Analysis and Quantification

#### 180 Regulations and Standards *A. K. Furr*

Engineering Practices Summary

PRIOR TO THE 20TH CENTURY, safety was largely a matter for the individual and, aside from some of the more obvious concepts, was not well understood. Medicine was only beginning to develop into the sophisticated discipline it is today. Few industries had safety programs. There was no significant regulatory protection for individuals, either in the workplace or elsewhere, and the environment was similarly unprotected. During the period prior to World War II, the growth of labor unions stimulated the creation of safety programs and the beginnings of regulatory actions.

After the war, safety and health issues took on increasing importance. A great amount of research was initiated to study safety management techniques, hazard evaluations, hazard amelioration and abatement, health effects of exposure to materials, manufacturing standards, and equipment design. In the early 1970s, safety and health issues received a tremendous boost by the passage of the Occupational Health and Safety Act. The concern of the public about the effects of exposure to hazardous substances in the air, water, and food generated a demand for regulations in these areas. As a result, there are now many regulatory acts which affect virtually every part of an individual's life. Engineers and other professionals must take these regulations into account in all of their activities.

Much of the most recent concern for safety involves an individual's health. Unfortunately, this is the most difficult area in which to establish meaningful regulations. In order to determine a substance's carcinogenic properties, studies must be done which are extremely expensive and may be invalid due to the necessity of using animal models and deriving the effects of normal exposure from data based on high levels of exposure. Statistical epidemiological studies are also difficult except in extreme cases due to problems of establishing statistically significant experimental and control cohorts. Much effort is still required in this area.

Unfortunately, due to the complexity of safety and health issues and regulations, and the feeling on the part of some that compliance is too costly and time consuming, many architects, engineers, and management personnel are not qualified to take advantage of all the information that has been garnered on safety-related factors in their work. Where safety factors and regulations are interpreted by reasonable persons, safety management can be a very cost-effective tool, raising productivity and improving employee morale. Industries which are not concerned about the

environment are finding that they are not welcome in a community. Unfortunately, the "not in my back yard" position has affected other, more responsible firms due to overreaction to these problem situations. It is incumbent on all professionals to work toward re-instilling public confidence that safety and health issues are a principal concern of their professions.

Rahimi, M. "Hazard Identification and Control"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Hazard Identification and Control

---

### 179.1 Hazard Identification

#### 179.2 Physical Hazards

Human Impact Injuries • Trip, Slip, and Fall • Mechanical Injuries

#### 179.3 Chemical Hazards

Hazardous/Toxic Substances

#### 179.4 Airborne Contaminants

#### 179.5 Noise

#### 179.6 Fire Hazards

#### 179.7 An Engineering Approach to Hazard Control

#### 179.8 Hazard Analysis and Quantification

Preliminary Hazard Analysis • Failure Mode, Effects and Criticality Analysis • Hazard and Operability Study • Fault Tree Analysis • Event Tree Analysis

### Mansour Rahimi

*University of Southern California*

Injuries and illnesses cost American industry an estimated \$115 billion in 1992 [[Occupational Hazards, 1993](#)]. Yet engineered systems are often designed without an extensive evaluation for their potential **hazards**. A glance at multidisciplinary literature in the field of safety and health shows that a large number of tools and techniques are available for hazard identification, analysis, and control. However, little standardization exists for using these techniques across fields such as engineering, physics, chemistry, sociology, psychology, business, and law. Specifically, design engineers do not make extensive use of safety resources commonly recommended by the safety community, nor are they fully familiar with hazard identification and analysis techniques [[Main and Frantz, 1994](#)]. In fact, as the system under study (e.g., product, machine, production cell, shop floor) becomes more complex in terms of hardware, software, and human, environmental, and organizational variables, the need to employ a more comprehensive hazard evaluation and control methodology becomes evident. In addition, the recent developments in environmentally conscious design require a deeper understanding of hazard consequences within and across the product life-cycle span (i.e., design, development, production, operation, deployment, use, and disposal). Therefore, successful safety programs require significant effort and resources on the part of a multidisciplinary team of investigators with support from all levels of management within a safety-cultured organizational structure [[Hansen, 1993](#)].



## 179.1 Hazard Identification

---

It is imperative that engineers become aware of hazard prevention principles early in their education. Also, the principles and techniques outlined in this chapter work best when applied early in the design of systems. The negative impact of hazards (e.g., accidents, diseases) will be far greater (sometimes irreversible) in the later stages of the system design or operations.

One way of classifying the sources of hazards is by the type of dominant (hazardous) energy used in the operation of the system. This energy is normally traced back to the root cause of an injury or illness. In this chapter, the following hazards are considered: physical, chemical, airborne contaminants, noise, and fire. Other hazards are extensively considered in the reference material, and ergonomic hazards are discussed elsewhere in this handbook. The objective of any hazard control program is to control losses from injuries and damages by identifying and controlling the degree of exposure to these sources of energy.

## 179.2 Physical Hazards

---

Physical hazards are usually in the form of kinematic force of lifting or impact by or against an object. More than 50% of compensable work injuries are included in this category.

### Human Impact Injuries

The severity of an injury depends on the velocity of impact, magnitude of deceleration, and body size, orientation, and position. The kinetic energy formula used to describe the impact injury is

$$E_{\text{ft-lb}} = (Wv^2)/2g \quad (179.1)$$

where  $W$  is the weight of an object or part of the body (lb),  $v$  is the velocity (ft/s), and  $g$  is gravity (ft/s<sup>2</sup>). However, if the impacting surface is soft, the kinetic energy for the impact is

$$E_{\text{ft-lb}} = [W(2sA)]/2g \quad (179.2)$$

where  $s$  is the stopping distance (ft), and  $A$  is the deceleration (ft/s<sup>2</sup>). For example, for both of the above cases, the human skull fracture occurs at 50 ft-lb of kinetic energy. Hard hats are expected to prevent the transfer of this energy to the human skull.

### Trip, Slip, and Fall

These injuries comprise 17% of all work-related injuries. Falls are the second largest source of accidental deaths in the U.S. (after motor vehicle accidents). Jobs related to manufacturing, construction, and retail and wholesale activities are the most susceptible to these types of hazards, comprising about 27% of all worker compensation claims in the U.S. These hazards include slipping (on level ground or on a ladder), falling from a higher level to a lower one (or to the ground), falling due to the collapse of a piece of floor or equipment, and failure of a structural support or walkway. Principles of tribology are being used to study the control mechanisms for these accidents. Tribology is the science that deals with the design and analysis of friction, wear,

and lubrication of interacting surfaces in relative motion. The basic measure of concern is the coefficient of friction (COF), which in its simplest form is the horizontal force divided by the vertical force at the point of relative motion. A COF greater than 0.5 appears to provide sufficient traction for normal floor surfaces. However, a number of other conditions make this hazard evaluation difficult: unexpectedness of surface friction change versus human gait progression, foreign objects and debris on a path, walkway depression, raised projections (more than 0.25 in.), change in surface slope, wet surfaces, improper carpeting, insufficient lighting, improper stair and ramp design, improper use of ladders, guardrails and handrails, and human visual deficiencies (color weakness, lack of depth perception and field of view, inattention, and distraction).

## Mechanical Injuries

The U.S. Occupational Safety and Health Act specifically states that one or more methods of machine guarding shall be provided the operator and other employees to protect them from hazards such as those created by point of operation, ingoing nip points, rotating parts, flying chips, and sparks. Other hazards in this category are: cutting by sharp edges, sharp points, poor surface finishes, splinters from wood and metal parts, shearing by one part of a machine moving across a fixed part (e.g., paper cutters or metal shearers), crushing of a skin or a tissue caught between two moving parts (e.g., gears, belts, cables on drums), and straining of a muscle (overexertion) in manual lifting, pushing, twisting, or repetitive motions.

Another important category is the hazard caused by pressure vessels and explosions. These hazards are generally divided into two types—boilers which are used to generate heat and steam, and unfired pressure vessels which are used to contain a process fluid, liquid, or gas without direct contact of burning fuel. The American Society of Mechanical Engineers (ASME) covers all facets of the design, manufacture, installation, and testing of most boilers and process pressure vessels in the Boiler and Pressure Vessel Code (total of 11 volumes). The primary safety considerations relate to the presence of emergency relief devices or valves to reduce the possibility of overpressurization or explosion. Explosions can be classified on the basis of the length-to-diameter ratio ( $L/D$ ) of the container. If the container has an  $L/D$  of approximately one, the rise in pressure is relatively slow and the overpressurization will cause the container to rupture. In containers with a large  $L/D$  ratio, such as gas transfer pipes and long cylinders, the initial flame propagates, creating turbulence in front of it. This turbulence improves mixing and expansion of the flame area and the speed of travel along the vessel, increasing the pressure by as much as 20 times very rapidly. Other types of explosions are caused by airborne dust particles, boiling liquid and expanding vapor, vessels containing nonreactive materials, deflagration of mists, and runaway chemical reactions.

Explosions may have three types of effects on a human body. First, the blast wave effect which carries kinetic energy in a medium (usually air), though decaying with distance, can knock a person down (overpressure of 1.0 lb/in.<sup>2</sup>) or reach a threshold of lung collapse (overpressure of 11 lb/in.<sup>2</sup>). The second type of effect is thermal, usually resulting from the fire. The amount of heat radiated is related to the size of the fireball and its duration of dispersion. The radiant energy is reduced according to the inverse distance-squared law. Most explosive fireballs reach temperatures about 2400°F at their centers. The third effect is the scattering of the material fragments. All three

of these injury effects are multifactorial and they are very difficult to predict.

## 179.3 Chemical Hazards

---

A Union Carbide plant in Bhopal, India, accidentally leaked methyl isocyanate gas from its chemical process. It left over 2500 dead and about 20 000 injured. There are over 3 million chemical compounds, and an estimated 1000 new compounds are introduced every year.

### Hazardous/Toxic Substances

The health hazards of chemicals can be classified into acute or chronic. The acute ones are corrosives, irritants, sensitizers, and toxic and highly **toxic substances**. The chronic ones are carcinogens, liver, kidney and lung toxins, bloodborne pathogens, nervous system damages, and reproductive hazards. For a reference listing of maximum exposures to these substances, refer to the technical committees of the American Conference of Governmental Industrial Hygienists (ACGIH) and American Industrial Hygiene Association (AIHA). Rules and standards related to manufacture, use, and transportation of these chemicals are promulgated and their **compliance** is enforced by governmental agencies such as the Occupational Safety and Health Administration (OSHA), the Environmental Protection Agency (EPA) and the Department of Transportation (DOT).

### Routes of Entry

There are four ways by which toxic substances can enter into the human body and cause external or internal injuries or diseases [Shell and Simmons, 1990].

1. Cutaneous (on or through the skin)
  - a. Corrosives: damage skin by chemical reaction
  - b. Dermatitis: irritants such as strong acids; sensitizers, such as gasoline, naphtha, and some polyethylene compounds
  - c. Absorbed through skin, but affecting other organs
2. Ocular (into or through the eyes)
  - a. Corneal burns due to acids or alkali
  - b. Irritation due to abrasion or chemical reaction
3. Respiratory inhalation (explained later in section 179.4)
4. Ingestion
  - a. Toxic substances may be ingested with contaminated food
  - b. Fingers contaminated with toxic chemicals may be placed in mouth
  - c. Particles in the respiratory system are swallowed with mucus

## Mechanisms of Injury

Toxic agents cause injury by one or a combination of the following mechanisms.

1. Asphyxiants
  - a. Asphyxia refers to a lack of oxygen in the bloodstream or tissues with a high level of carbon dioxide present in the blood or alveoli
  - b. Gas asphyxiants (e.g., carbon dioxide, nitrogen, methane, hydrogen) dilute the air, decreasing oxygen concentration
  - c. Chemical asphyxiants make the hemoglobin incapable of carrying oxygen (carbon monoxide) or keep the body's tissues from utilizing oxygen from the bloodstream (hydrogen cyanide)
2. Irritants which can cause inflammation (heat, swelling, and pain)
  - a. Mild irritants cause hyperemia (capillaries dilate)
  - b. Strong irritants produce blisters
  - c. Respiratory irritants can produce pulmonary edema
  - d. Secondary irritants can be absorbed and act as systemic poisons
3. Systemic poisons
  - a. Poisons may injure the visceral organs, such as the kidney (nephrotoxic agents) or liver (hepatotoxic agents)
  - b. Poisons may injure the bone marrow and spleen, interrupting the production of blood (benzene, naphthalene, lead)
  - c. Poisons may affect the nervous system, causing inflammation of the nerves, neuritis, pain, paralysis, and blindness (methyl alcohol, mercury)
  - d. Poisons may enter the bloodstream and affect organs, bones, and blood throughout the body (usually with prolonged exposure)
4. Anesthetics
  - a. May cause loss of sensation
  - b. May interfere with involuntary muscle actions causing respiratory failure (halogenated hydrocarbons)
5. Neurotoxins
  - a. Neurotics affect the central nervous system
  - b. Depressants cause drowsiness and lethargy (alcohol)
  - c. Stimulants cause hyperactivity
  - d. Hypnotics which are sleep-inducing agents (barbiturates, chloral hydrate)
6. Carcinogens
  - a. Cancers of the skin at points of contact (tar, bitumen)
  - b. Cancers of internal organs and systems have numerous known and suspected causes (labeled as suspected carcinogens)

7. Teratogenic effects. A substance that may cause physical defects in the developing embryo or fetus when a pregnant female is exposed to the substance for a period of time.

## 179.4 Airborne Contaminants

---

The greatest hazard exists when these contaminants have sizes smaller than  $0.5\ \mu\text{m}$  where they can be directly introduced into the bloodstream through alveolar sacs (e.g., zinc oxide, silver iodide). Particles larger than  $0.5\ \mu\text{m}$  are entrapped by the upper respiratory tract of trachea and bronchial tubes (e.g., insecticide dust, cement and foundry dust, sulfuric acid mist). There are two main forms of airborne contaminants [Brauer, 1990]—particulates (dusts, fumes, smoke, aerosols, and mists), and gases or vapors.

Dusts are airborne solids, typically ranging in size from  $0.1$  to  $25\ \mu\text{m}$ , generated by handling, crushing, grinding, impact, detonation, etc. Dusts larger than  $5\ \mu\text{m}$  settle out in relatively still air due to the force of gravity. Fumes are fine solid particles less than  $1\ \mu\text{m}$  generated by the condensation of vapors. For example, heating of lead (in smelters) vaporizes some lead material that quickly condenses to small, solid particles. Smokes are carbon or soot particles less than  $0.1\ \mu\text{m}$  resulting from incomplete combustion of carbonaceous material. Mists are fine liquid droplets generated by condensation from the gaseous to liquid state, or by the dispersion of same by splashing, foaming, or atomizing.

Gases are normally formless fluids which occupy space and which can be changed to liquid or solid by a change in pressure and temperature. Vapors are the gaseous form of substances which are normally in a liquid or solid state.

The measures for toxicity of the above substances are given in parts per million (ppm) for gases and vapors, and milligrams per cubic meter ( $\text{mg}/\text{m}^3$ ) for other airborne contaminants. The criteria for the degree of toxicity are the threshold limit values (TLVs) based on review of past research and monitory experience [early OSHA standards listed permissible exposure limit (PEL)]. TLVs are airborne concentrations of substances which are believed to represent conditions to which nearly all workers may be repeatedly exposed, eight hours a day, for lifetime employment, without any adverse effects. For acute toxins, short-term exposure levels (STELs) are indicated for a maximum of 15 minutes of exposure, not more than four times a day. Because exposures vary with time, a time-weighted average (TWA) is adopted for calculating TLVs:

$$\text{TLV(TWA)} = \sum (E_i T_i / 8) \quad (179.3)$$

where  $E_i$  is the exposure to the substance at concentration level  $i$  and  $T_i$  is the amount of time for  $E_i$  exposure in an eight-hour shift.

In many environments, there are several airborne substances present at the same time. If the effects of these substances are additive and there are no synergistic reactions, the following formula can be used for the combination of TLVs:

$$X = (C_1/T_1) + (C_2/T_2) + \cdots + (C_n/T_n) \quad (179.4)$$

where  $C_i$  is the atmospheric concentration of a substance and  $T_i$  is the TLV for that substance. If  $X < 1$ , the mixture does not exceed the total TLV; if  $X \geq 1$ , the mixture exceeds the total TLV.

## 179.5 Noise

---

Noise-induced hearing loss has been identified as one of the top ten occupational hazards by the National Institute for Occupational Safety and Health (NIOSH). In addition to hearing loss, exposure to excessive amounts of noise can increase worker stress levels, interfere with communication, disrupt concentration, reduce learning potential, adversely affect job performance, and increase accident potential [Mansdorf, 1993]. Among many types of hearing loss, sensorineural hearing loss is the most common form in occupational environments. Sensorineural hearing loss is usually caused by the loss of ability of the inner ear (cochlea nerve endings) to receive and transmit noise vibrations to the brain. In this case, the middle ear (the bone structures of malleus, incus, and stapes) and the outer ear (ear drum, ear canal, and ear lobe) may be intact.

A comprehensive and effective hearing conservation program can reduce the potential for employee hearing loss, reduce workers compensation costs due to hearing loss claims, and lessen the financial burden of noncompliance with government standards. Current OSHA standards require personal noise dosimetry measurements in areas with high noise levels. Noise dosimeters are instruments which integrate (measure and record) the sound levels over an entire work shift. Noise intensities are measured by the dBA scale, which most closely resembles human hearing sensitivity. For continuous noise levels, OSHA's permissible noise exposure is 90 dBA for an eight-hour shift. If the noise levels are variable, a time-weighted average is computed. For noise levels exceeding the limit values, an employee hearing conservation program must be administered [Mansdorf, 1993, pp. 318-320].

## 179.6 Fire Hazards

---

Fire is defined as the rapid oxidation of material during which heat and light are emitted. The National Fire Protection Association (NFPA) reported that large fire losses in the U.S. for the year 1991 exceeded \$2.6 billion. The most frequent causes of industrial fires are electrical (23%), smoking materials (18%), friction surfaces (10%), overheated materials (8%), hot surfaces (7%), and burner flames (7%). The process of combustion is best explained by the existence of four elements—fuel, oxidizer ( $O_2$ ), heat, and chain reaction. A material with a flash point below 100°F (vapor pressure  $< 40$  lb/in.<sup>2</sup>) is considered flammable, and higher than 100°F is combustible. In order to extinguish a fire, one or a combination of the following must be performed: the flammable/combustible material is consumed or removed, the oxidant is depleted or below the necessary amount for combustion, heat is removed or prevented from reaching the combustible material not allowing for fuel vaporization, or the flames are chemically inhibited or cooled to stop the oxidation reaction.

Fire extinguishers are classified according to the type of fire present. Class A involves solids which produce glowing embers or char (e.g., wood, paper). Class B involves gases and liquids which must be vaporized for combustion to occur. Class C includes Class A and B fires involving

electrical sources of ignition. Finally, Class D involves oxidized metals (e.g., magnesium, aluminum, titanium). In addition to heat, the most dangerous by-products of fires are hazardous gases and fumes, such as CO, CO<sub>2</sub>, acrolein formed by the smoldering of cellulosic materials and pyrolysis of polyethylene, phosgene (COCl<sub>2</sub>) produced from chlorinated hydrocarbons, sulfur dioxide, oxides of nitrogen (NO<sub>x</sub>) resulting from wood products, ammonia (NH<sub>3</sub>) when compounds containing nitrogen and hydrogen burn in air, and metal fumes from electronic equipment. For a complete reference to facility design requirements and exiting requirements, refer to NFPA Code 101.

## 179.7 An Engineering Approach to Hazard Control

---

As it is impossible to design an absolutely "safe" system, the following list is suggested to eliminate the critical hazards and reduce other hazards to an acceptable level (**risk control**). It is important to mention that the effectiveness of the control mechanism is diminished by using the lower items on this list.

1. Identify and eliminate the source of hazardous energy. For example, do not use high-voltage electricity. Or consider the use of noncombustible and nontoxic material in environments with fire potentials. This approach is not always practical or cost-effective.
2. Reduce the degree of hazardous energy. For example, use low-voltage solid state devices to reduce heat buildup in areas with explosion hazards. This approach is practical in some cases, yet costly in some other design applications.
3. Isolate the source of hazard. Provide barriers of distance, shields, and personal protective equipment to limit the harmful effects of the hazardous agents. To control the sequence of events in time and space, a lockout/tagout procedure is recommended.
4. Minimize failure. Include constant monitoring of critical safety parameters (e.g., gas concentrations or radiation levels). The monitoring system should detect, measure, understand and integrate the readings, and respond properly.
5. Install a warning system. Similar to consumer product warnings, all system components should be equipped with warning and proper communication systems. Operators (and the general public) should be warned of the type of hazard present and the means by which information can be obtained in case of an accident. However, too many warning signs and display indicators may confuse the operator.
6. Ensure safe procedures. A common cause of accidents is the inadequacy of procedures and the failure to follow the proper procedures.
7. Provide for backout and recovery. In case of an accident, this defensive step is taken to reduce the extent of injury and damage. This step incorporates one or more of the following actions: (a) normal sequence restoring, in which a corrective action must be taken to correct the faulty operation, (b) inactivating only malfunctioning equipment which is applied to redundant components or temporarily substituting a component, (c) stopping the entire operation to prevent further injury and damage, and (d) suppressing the hazard (e.g., spill containment of highly hazardous substances).

When hazard exposure cannot be reduced through engineering controls, an effort should be made to limit the employee's exposure through administrative controls, such as (a) rearranging work schedules and (b) transferring employees who have reached their upper exposure limits to an environment where no additional exposure will be experienced.

## 179.8 Hazard Analysis and Quantification

A number of **hazard analysis techniques** have been developed to study the hazards associated with a system. For a detailed review, see *Guidelines for Hazard Evaluation Procedures* [American Institute of Chemical Engineers, 1985]; Gressel and Gideon [1991]; Vincoli [1993]. These techniques vary in terms of their hazard evaluation approaches and the degree to which hazard exposures can be quantified. A precursor to any of these techniques is the system, which is divided into its small and manageable components, analyzed for causes and consequences of any number of potential hazards, and then synthesized to consider hazard effects on the whole system. Five hazard analysis techniques are briefly presented here.

### Preliminary Hazard Analysis

Preliminary hazard analysis (PHA) is the foundation for effective systems hazard analysis. It should begin with an initial collection of raw data dealing with the design, production, and operation of the system. The purpose of this procedure is to identify any possible hazards inherent in the system. One example is energy as the source of hazard to explore the multitude of circumstances by which an accident can occur in a system. Table 179.1 demonstrates an actual use of PHA in the design phase of metal chemical vapor deposition. The four main categories of this table are hazard, cause, main effects, and preventive control. The hazard effects and corrective/preventive measures are only tentative indicators of potential hazards and possible solutions.

**Table 179.1** A Sample for the Application of Preliminary Hazard Analysis to the Design of Metal Organic Chemical Vapor Deposition (Only Two of the Hazards Are Listed)

Hazard	Cause	Main Effects	Preventive Control <sup>a</sup>
Toxic gas release	Leak in storage cylinder	Potential for injury and fatality from large release	Develop purge system to remove gas to another tank Minimize on-site storage Provide warning system Develop procedure for tank inspection and maintenance Develop emergency response system
Explosion, fire	Overheat in reactor tube	Potential for fatalities due to toxic release and fire Potential for injuries and fatalities due to flying debris	Design control system to detect overheat and disconnect heater Provide warning system for temperature fluctuation, evacuate reaction tube, shut off input valves, activate cooling system



Adapted from Kavianian, H. R. and Wentz, C. A. 1990. *Occupational and Environmental Safety Engineering and Management*. Van Norstrand Reinhold, New York.

<sup>a</sup> This column is simplified to show the major categories of hazard control techniques.

## Failure Mode, Effects and Criticality Analysis

While PHA studies hazards in the entire system, failure mode, effects, and criticality analysis (FMECA) analyzes the components of the system and all of the possible failures which can occur. This form of analysis identifies items whose failures have a potential for hazardous consequences. In this analysis, each item's function must be determined. Once this is done, the causes and effects of the failure of the components are indicated. Then, the criticality factor of each failure is determined, and a quantified severity rating is given to the factor. Table 179.2 shows an example for FMECA. Because the frequency of each potential occurrence is also an important factor, a risk assessment matrix (depicted in Table 179.3) can be used to codify the risk assignment. A design team (including the safety engineer) can use this FMECA to redesign components or parts of the system to reduce the criticality rating to predetermined acceptable regions (preferably a hazard risk index between one and five).

**Table 179.2** A Sample for the Application of Failure Mode, Effects, and Criticality Analysis to the Metal Organic Chemical Vapor Deposition Process (Only Three System Components Are Listed)

System Component	Failure Mode	Effects	Criticality Ranking <sup>a</sup>
Reactor tube	Rupture	Release of pyrophoric gas causing fire and release of toxic gases	III
Control on reactor heater	Sensor fails; response control system fails	Reactor overheating beyond design specification	II
Refrigeration equipment	Failure to operate	Increase in vapor pressure; cylinder rupture	IV

Adapted from Kavianian, H. R. and Wentz, C. A. 1990. *Occupational and Environmental Safety Engineering and Management*. Van Norstrand Reinhold, New York.

<sup>a</sup> Criticality ranks are based on a scale from I to IV.

**Table 179.3** Hazard Risk Assessment Matrix

Frequency of Occurrence	Hazard Category			
	I Catastrophic	II Critical	III Marginal	IV Negligible
Frequent	1	3	7	13
Probable	2	5	9	16
Occasional	4	6	11	18
Remote	8	10	14	19
Improbable	12	15	17	20

Hazard Risk Index	Suggested Criteria
1-5	Unacceptable
6-9	Undesirable
10-17	Acceptable with review
18-20	Acceptable without review

## Hazard and Operability Study

Hazard and operability study (HAZOP) is one of the most tedious, yet thorough, forms of hazard analysis. It identifies potentially complex hazards in a system. HAZOP examines a combination of every part of the system and analyzes the collected data to locate potentially hazardous areas. The first step is to define the system and all specific areas from which data will be collected. This will help in deciding the HAZOP team. Next, a team of experts in these areas is assembled to analyze the collected data. The team usually consists of experts in the fields of engineering, human behavior, hygiene, organization and management, and other personnel who may have operational expertise related to the specific system being analyzed. Once this is done, an intensive information gathering process begins. All aspects of the system's operation and its human interfaces are documented. The information is then broken down into small information nodes. Each node contains information on the procedure or specific machine being used in the system. Each node is then interconnected logically with other nodes in the system. Each node is also given guide words which help identify its conditions. Table 179.4 gives an example of guide words. Each guide word is a functional representation of subsystem hazard. The team can analyze and determine the criticality or likelihood that this node could produce a hazard. At this point, the HAZOP team will need to determine what course of action to take. This procedure is one of the most comprehensive, yet time consuming, of the analysis tools. It is widely used in large and complex systems with critical safety components, such as petrochemical facilities.

**Table 179.4** HAZOP Data Table for Vacuum Air Vent Node and Reverse Flow Guide Word (Design Intention: To Vent Air into the Sterilizer following a Vacuum Stage)

Guide Word	Cause	Consequence	Recommendation
Reverse flow	Control valve leakage	Ethylene oxide leak into utility room	Air vent should not receive air from utility room, but from exhaust duct to reduce risk from reverse flow leakage of ethylene oxide

Adapted from *Hazard and Operability Study of an Ethylene Oxide Sterilizer*. 1989. National Institute for Occupational Safety and Health, NTIS Publication No. PB-90-168-980. Springfield, VA, p. 27

## Fault Tree Analysis

Fault tree analysis (FTA) uses deductive reasoning to qualitatively and quantitatively depict possible hazards which occur due to failure of the relationships between the system components (e.g., equipment, plant procedures). FTA uses a pyramid style tree analysis to start from a top undesired event (e.g., accident, injury) down to the initial causes of the hazard (e.g., a joint

separation under vibrating forces, erroneous operating procedure). There are four main types of event symbols—a fault event, basic event, undeveloped event, and normal event. A fault event is considered to be an in-between event and never the end event. A basic event is considered to be the final event. An undeveloped event is an event which requires more investigation because of its complexity or lack of analytical data. A normal event is an event which may or may not occur. Each one of these events is joined in the tree by a logic symbol. These gates explain the logic relationship between each of the events. For example, a main event is followed in the tree by two possible basic events. Either event can produce the main event. The logic gate in this case would be an *or* gate, but if both events could have contributed to the main event an *and* gate would be used. FTA's ability to combine causal events together to prevent or investigate accidents makes it one of the most powerful accident investigation and analysis tools.

## Event Tree Analysis

Event tree analysis (ETA) is similar to FTA, with the exception that ETA uses inductive reasoning to determine the undesired events which are caused by an earlier event. ETA uses the same pyramid structure as the previous analysis. However, rather than working from top to bottom, ETA works from left to right, dividing the possibility of each event into two outcomes—true (event happening), or false (event not happening). By taking an initial failure of an event, the tree designer tries to incorporate all possible desired and undesired results of the event. The advantage to this system is that it helps predict failures in a step-by-step procedure. This helps the analyst to provide a solution or a countermeasure at each analysis node under consideration. ETA's main weakness lies with its inability to incorporate multiple events at the same time.

Other techniques such as management oversight, risk tree, "what if" analysis, software hazard analysis, and sneak circuit analysis are discussed in the publications listed in the "Further Information" section.

## Defining Terms

**Compliance:** The minimum set of requirements by which an environment conforms to the local, state, and federal rules, regulations, and standards. According to the seriousness of the violation, a workplace may be cited by OSHA for imminent danger, serious violation, nonserious hazards, and de minimis violations. In addition to penalties (fines of up to \$70000 per violation), other civil and criminal charges may be brought against responsible supervisors and managers.

**Hazard:** A set of (or change in) a system's potential and inherent characteristics, conditions, or activities which can produce adverse or harmful consequences, including injury, illness, or property damage (antonym to *safety*).

**Hazard analysis techniques:** A number of analytical methods by which the nature and causes of hazards in a product or a system are identified. These methods are generally designed to evaluate the effects of hazards and offer corrective measures or countermeasures.

**Mechanical injuries:** A type of physical injury caused by excessive forces applied to human body components, such as cutting, crushing, and straining (ergonomic hazards).

**Risk control:** The process by which the probability, severity and exposure to hazards (per mission and unit of time) are considered to reduce the potential loss of lives and property.

**Toxic substances:** Those substances that may, under specific circumstances, cause injury to persons or damage to property because of reactivity, instability, spontaneous decomposition, flammability, or volatility (including those compounds that are explosive, corrosive, or have destructive effects on human body cells and tissues).

## References

- Brauer, R. L. 1990. *Safety and Health for Engineers*. Van Nostrand Reinhold, New York.
- Gressel, M. G. and Gideon, J. A. 1991. An overview of process hazard evaluation techniques. *Am. Industrial Hyg. Assoc. J.* 52(4): 158-163.
- Guidelines for Hazard Evaluation Procedures*. American Institute of Chemical Engineers, New York.
- Hansen, L. 1993. Safety management: A call for (r)evolution. *Prof. Saf.* 38(11): 16-21.
- Main, B. W. and Frantz, J. P. 1994. How design engineers address safety: What the safety community should know. *Prof. Saf.* 39(2): 33-37.
- Mansdorf, S. Z. 1993. *Complete Manual of Industrial Safety*. Prentice Hall, Englewood Cliffs, NJ.
- Occupational Hazards*, Editorial, 1993, November, p. 6.
- Shell, R. L. and Simmons, R. J. 1990. *An Engineering Approach to Occupational Safety and Health in Business and Industry*. Institute of Industrial Engineers, Norcross, GA.
- Vincoli, J. W. 1993. *Basic Guide to System Safety*. Van Nostrand Reinhold, New York.

## Further Information

Safety science is a broad and multidisciplinary field. The engineering aspects of this field, outlined in this chapter, have been mostly discussed by others in the "References" section. For further readings, refer to the following.

- Kavianian, H. R. and Wentz, C. A. 1990. *Occupational and Environmental Safety Engineering and Management*. Van Nostrand Reinhold, New York.
- Bird, F. E., and Germain, G. L. 1986. *Practical Loss Control Leadership: The Conservation of People, Property, Process, and Profits*. International Loss Control Institute, Loganville, GA.
- Hammer, W. 1993. *Product Safety Management and Engineering*, 2nd ed. American Society of Safety Engineers, Des Plaines, IL.
- Hansen, D. J. (Ed.) 1991. *The Work Environment: Occupational Health Fundamentals*. Lewis Publishers, Chelsea, MI.
- Accident Prevention Manual for Business and Industry: Engineering and Technology*, 10th ed. 1992. National Safety Council, Itasca, IL.

Furr, A. K. "Regulations and Standards"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Regulations and Standards

---

[180.1 Engineering Practices](#)

[180.2 Summary](#)

### **A. Keith Furr**

*Virginia Polytechnic Institute & State University, Retired*

Engineers and corporations have long recognized the need for the establishment of basic minimum standards for safe design and construction of structures and equipment. As a result, most major industrial groups have established associations to set guidelines for these minimum standards. However, it has been left up to the various jurisdictions to adopt any of these standards as binding. For example, fire codes are usually designed and administered locally. At least one state has virtually no fire codes that apply statewide. In other cases, the application of the codes is not universal. In the state where this author lives, only selected categories of structures are covered by the statewide code, while local building codes vary widely in their scope and interpretation depending on the local officials responsible for their administration.

After World War II, the need was recognized for "universally" applicable standards, at least in certain areas. One of the earliest of these national standards was in the field of nuclear energy, a prototype for many of the agencies which have since followed. The responsible federal agency, the Nuclear Regulatory Commission (NRC),■

The names of regulatory agencies used here are the current ones. Many of the agencies have changed names over time.

set the basic standards by which all activities involving radioactive materials created as a result of reactor-related operations were regulated. States were given the option of either allowing the federal agency to supervise the use of these materials or adopting regulations at least as stringent as the federal regulations and administer them themselves. This model has persisted in many of the other regulatory acts passed since then. In the early 1970s, the Occupational Safety and Health Act (OSHA) was passed, covering a very large number of industries. The standards in the OSHA act were, for the most part, incorporated directly from the standards and guidelines established by the industrial associations. Unfortunately, the law does not permit changes to the standards to be made easily, so that the legal standards often lag behind the industrial standards. Recently, there have been revisions to the regulatory provisions to make them more applicable to current levels of safety design.

Other major federal safety and health standards are those of the Environmental Protection Agency (EPA), the Department of Transportation, the Food and Drug Administration, the Department of Energy, the Federal Aviation Administration, and a host of other specialized

agencies. The scope of regulations is a rapidly evolving area. In recent years, federal operations have been brought under both OSHA and EPA regulations from which they were previously excluded. OSHA has taken a very active role in the operations of medical and health-related organizations with the Bloodborne Pathogens Act, and the problems of the disabled are now regulated under the Americans with Disabilities Act (ADA). A major area likely to be changed significantly in the future is noise control. Many engineers agree that the current noise safety program is too lenient. The initial threshold is likely to be changed from 90 dB to 85 dB and the doubling ratio changed from 5 dB to 3 dB. These two changes will require major engineering modifications in many cases. Another potential regulatory area that will have a significant impact on engineering design is indoor air quality, an area insufficiently researched until recently. The material in this chapter will primarily cover areas under the purview of OSHA, EPA, and the NRC.

## **180.1 Engineering Practices**

---

Incorporated in the standards for nuclear safety, and now in most other safety and health regulatory standards, was the concept of achieving compliance with safety standards wherever possible by engineering controls. For example, it should, in principle, be impossible to operate a nuclear reactor unsafely. Approaching criticality too rapidly should be impossible by having interlocks which would shut down operations automatically if this were done or if one of a number of other conditions existed, such as loss of coolant, excessively high temperatures, areas of radioactive materials, excessively high radiation levels in monitored areas, earthquakes, and loss of information in critical circuits. The control mechanisms are supposed to be designed so that redundant failures have to occur to make the operations unsafe; no single equipment failure should render operations unsafe (in the sense that an automatic or controlled shutdown should always occur should a failure occur). Of course, experience has shown that humans can bypass almost any safety feature if they try. The catwalk in the Kansas City hotel fell due to human error. The Chernobyl incident was due to a combination of a poor design, with little safety margin, and operations which were lax in maintenance, training, and function. In some cases, failure occurs due to the lack of sufficient knowledge or information. The collapse of the Tacoma Narrow Bridge in 1940, only four months after completion, is an excellent example. In any case, safety and health standards today require that, wherever feasible, safety engineering principles as identified by the regulations be the first option in complying with the standards.

Many of the earlier industrial standards, including those emphasized in the earliest versions of the OSHA standard, are intended to prevent physical injuries. Bending brakes and hydraulic presses are required to have controls making it impossible for the machines, if used properly, to injure any part of the operator's body. Grinding tools and other devices with exposed rotating components are supposed to be designed to prevent hair or clothing to be grabbed by the rotating component or for materials to be thrown by the tool and impact a person, particularly in the face and eye. The design of portable electrical tools is required to make it impossible for them to produce a shock under normal operation. Ladders, scaffolds, and guardrails are defined in great detail to prevent employees from slipping or falling. Indeed, one of the major weaknesses in the original standards was that they defined safety in such minute detail that they were seen by many

to be concerned with trivia, so that they failed in their mission to increase safety. Nevertheless, even with the original problems, most of which have now been addressed, the passing of OSHA was a major step forward in occupational safety. Not only employees benefit, but so does the general public. If a machine or structure is designed properly, it should be safe for everyone with a legitimate need and training to use the device or facility.

Other areas of the original OSHA act dealt with fire safety by adopting standards of the National Fire Prevention Association on the storage of flammable materials, limiting bulk storage, storage within facilities, and sizes of containers for appropriate classes of liquids. Similar restrictions have been established for compressed gas systems, especially gases with inherent safety problems, such as hydrogen, oxygen, and ammonia. With few changes, these regulations have remained in effect and are well understood by engineers charged with designing appropriate facilities.

Recently changes in the OSHA standard are more concerned with meeting performance specifications, allowing the engineer, owner, or user to decide how best to meet the compliance goal. Meeting safety and health regulatory standards through engineering practices is still required as the first option, with personal protective devices and procedural controls used as options when engineering controls are not practicable. A major crossroads was created a few years ago when OSHA made a distinction between industrial operations from the pilot plant scale upward and those of laboratory-scale operations. It was recognized that the designs appropriate to a large plant where a modest number of chemicals are in use, albeit on a substantial scale, are not appropriate for the laboratory environment, where literally hundreds of different chemicals are used on a rapidly changing basis. Therefore, the OSHA Laboratory Safety Act was added to the basic OSHA act, preempting the OSHA general industry standards for laboratories. Neither section of the act is more permissive than the other in terms of the intended goals, a healthy and safe environment. The two different environments do lead to substantially different engineering approaches to many problems, and engineers specializing in the two areas need to understand the differences.

One of the major issues, even in the early versions of the OSHA standards, concerns the levels of specific airborne contaminants in the workplace. For various reasons, such as toxicity, neurological risk, fetal hazards, and corrosiveness, permissible exposure levels (PELs) were set by OSHA. Theoretically, a person in good health could work an eight-hour work day without harm if the exposure levels were less than these, but could not be allowed to work if they exceeded the PELs based on an eight-hour time-weighted average. In some cases, higher short-term exposure limits (STELs) were established, and in others ceiling (C) limits were established. The initial values were essentially the threshold limit values (TLVs) recommended by the American Conference of Governmental Industrial Hygienists (ACGIH). The latter are reviewed on a yearly basis and have changed significantly. Recently OSHA attempted to change their PELs (which are legally enforceable) to be the same in most cases to the more recent values of the ACGIH (which are guidelines and are not legally enforceable). However, court actions required OSHA to rescind the changes, so the current PELs are in many cases out of date with the current available knowledge. Some states, on their own, did adopt the changed values, so now engineers must be aware of the law in their area. A conservative approach would be to design using the ACGIH guidelines since they normally provide an additional safety level. In neither case should the limits



be considered as black and white. Many factors will modify the sensitivity of a specific individual to a given contaminant; for example, a person may have reduced tolerance due to a current health condition, or natural variations in individual tolerance can occur due to heredity or sex factors. A good design practice is to maintain an action level no more than 50% of the PEL, and preferably less. Many of the newer OSHA standards now incorporate the 50% action level concept.

The importance of designing ventilation systems to remove air contaminants does not seem to be as highly appreciated by some engineers as might be expected. Routine ventilation designs for offices and homes are geared typically for temperature control. Since hot air rises, inlet and exhaust duct work and openings are typically placed in the ceiling. However most airborne contaminants are heavier than air, so general room exhausts for these contaminants should be located close to the floor, or no lower than waist high. The air flow should not be such as to draw or discharge contaminated air indoors through an individual's breathing zone. Various recommendations exist for the amount of fresh air needed depending on the type of occupancy. Typically, in an office environment a value of 20 cubic feet per minute (cfm) of fresh air is recommended. Relatively few older structures meet this requirement. In chemically contaminated areas, such as laboratories, the number of air changes per hour is often used as a criterion, with current levels being 10 to 12 per hour. This has serious implications for laboratories since the normal requirement is for this to be 100% fresh air. In office structures the air is normally partially recycled so that only 15% to 30% may be fresh.

The recycling of air in newer energy-efficient buildings appears to be at least partly responsible for the increasing evidence of health problems in these tight buildings. Recycling air can cause a problem which exists in one part of a building to eventually spread throughout the building. This is especially true when the system is itself the cause of the problem. Most duct systems are not designed for ease of cleaning, and over time become dirty. The dirt and grease which builds up in these systems is an ideal breeding ground for molds and bacteria, to which many individuals are allergic. Not only should high-efficiency filters be used and replaced frequently, but the duct work interior needs to be cleaned periodically. If air quality is brought under regulations, there will be significant engineering requirements on the design of the systems to facilitate cleaning, on filter designs, and on the selection of materials for construction to limit the emission of volatile organic compounds. There is increasing evidence that long-term exposure to some organics, such as formaldehyde, may cause problems to hypersensitive individuals at levels very much below the permissible PELs or TLVs. The National Institutes of Occupational Safety and Health (NIOSH) already are recommending levels substantially below those of both OSHA and the ACGIH for a number of materials due to these low, continuous exposures. Unfortunately, our knowledge of health effects in general and especially those related to long-term exposures at low levels leaves a great deal to be desired, although it is improving. Statistically, the problem with low levels of airborne contaminants is very much like the situation with low levels of radioactivity. The problems may exist, but the experimental evidence is virtually impossible to obtain because of the extremely large number of subjects required for direct measurement.

The most effective and energy-conserving means of controlling contaminants is to capture them at the source. In a factory large-scale scavenging systems may be appropriate, while in laboratories it would be appropriate to require that all work with the potential to release hazardous vapors,

fumes, or gases be done in fume hoods with at least 100 feet per minute (fpm) face velocity.

Once contaminated air has left a facility, it is hoped that it will be diluted to a level that is safe. Air pollution standards have been established which may or may not be sufficient. Problems such as acid rain and ozone depletion are well documented in the literature and have been extensively publicized in the media. In laboratory buildings, for example, engineers until quite recently typically placed rain caps over fume hood exhausts so that the contaminated air was forced back onto the roof of a building and in many cases was recaptured by the building air system. Often, once a laboratory building was completed, subsequent modifications added fume hoods to a duct system which did not have the capacity to handle them. Fire codes typically require laboratories to be at negative pressure with respect to the corridors so that fire within a laboratory cannot spread via the corridors. However, the negative pressure also prevents contaminated air from spreading via the corridors as well. Exhaust ducts for the newer hoods were often allowed to penetrate floors intended to be fire separations. All of these practices violated standards recommended by either the ACGIH, NFPA, or American Society of Heating, Refrigerating and Air Conditioning Engineers (ASHRAE). Since many of the standards have now been adopted by OSHA, a strong national regulatory agency is charged with enforcing the basic standards.

The Environmental Protection Agency has been a major factor in changing engineering practices. One of the most significant standards recently enacted dealt with underground petroleum storage tanks. A large percentage of organizations with a substantial number of employees, including industries, transportation departments, hospitals, and universities, have either heating oil tanks or gasoline tanks which are buried. Tanks which have been buried for 20 years or more have an increasing probability of leaking as they age. The EPA regulation requires all of these tanks above a certain capacity to be protected from leaking into the environment by 1998. Typically, this will involve replacing a tank with a double-shelled unit equipped with appropriate sensors to ensure that leaking does not occur. If environmental contamination has already occurred, the owner is required to clean up the contaminated site. This law alone has generated immense growth in engineering firms specializing in testing, abating, and replacing these sites.

Although addressed in **Chapter 89**, the related issue of prevention of ground contamination from landfills will be briefly mentioned here. In the last five years, a very large number of landfills have been closed due to the enactment of strict standards on landfills. Although no responsible person would deliberately dispose of any kind of hazardous waste routinely at a municipal landfill, it is estimated that 3 to 5% of all municipal solid waste could be characterized as hazardous waste, primarily from households and small business establishments. To lengthen the life of existing landfills, and to reduce the cost of new ones, recycling programs have been growing dramatically and have given rise to their own engineering disciplines. Alternative means of hazardous and solid waste disposal have been developed to reduce or eliminate the burial of wastes which could remain hazardous for thousands of years. Various incineration methods have been developed, which must comply with EPA's clean air standards as well as being subject to the "not-in-my-back-yard" (NIMBY) syndrome. Dozens of techniques have been developed specifically for disposal of medical wastes, which are regulated by state programs of varying stringency. No matter what engineering approach is involved in any of these environmental issues (and others not mentioned),

the net result must be to reduce the level of contaminants in the soil or air below a regulatory defined level either by destruction or by modification. Underground burial has been deemed unacceptable for most hazardous materials, with the notable exception of radioactive isotopes, for whose destruction no practical means exist. The major regulatory act which sets standards for the disposition of materials with hazardous properties is the Resource Conservation and Recovery Act (RCRA). Since most facilities do not process their hazardous waste locally, packaging and transportation of these materials must also meet the requirements of the Department of Transportation. These standards have changed significantly within the past five years and now are generally consistent with comparable international requirements.

Another major EPA standard is the community-right-to-know standard. Under this act, communities must be made aware of potential industrial chemical hazards in their areas.

The original standards for making buildings accessible to the handicapped were less successful than desired, so the Americans With Disabilities Act was passed in the early 1990s. This standard is much more stringent and carries the potential for much more severe penalties. Not only are the familiar wheelchair ramps and modified bathroom facilities required, but a number of other engineering changes are specified as well. Safety equipment, such as safety showers and eyewash fountains, must now meet specifications to ensure that a disabled person can use them. Fire alarms must provide a high-intensity visual alarm as well as an audible alarm. Doors must provide means to prevent a disabled person from being trapped or injured by the opening and closing mechanism. Instructions on public signs, including safety signs, must be provided in braille where appropriate. Meetings which could be attended by the disabled must ensure that the information can be received by a disabled person. This could include audio and or visual equipment for the hearing or visually impaired, or possibly a sign-language interpreter. The availability of these resources must be made known at the time the meeting is announced. Places of refuge must be identified to which disabled persons could retreat while awaiting rescue in the event of a fire or other incident. All new facilities must be designed to comply with the standard, and older ones will have to be brought toward full compliance.

The Food and Drug Administration not only establishes standards for nutritional information and for such things as drugs, pesticides, and residues of these materials on foods, but also regulates or provides guidelines for the design of x-ray machines, lasers, microwave units, and other generators of electromagnetic emissions. Most of their safety standards are intended as requirements for the manufacturer and effectively require engineering safeguards to be incorporated in these devices which a user cannot easily bypass. For example, industrial x-ray cameras are required to be interlocked so that the intense radiation areas within the target area cannot be accessed while the beam is on. The radiation levels within this area are such that a serious radiation burn to the hands can occur within 30 seconds. Requirements for the design of x-ray units for diagnostic tests now mandate that the unit define the exposed target area by light so that areas other than the one of interest are not exposed. Similar restrictions are imposed on lasers which have sufficient power to cause blindness if a beam were to reach the eye directly or by non-diffuse reflection. The standards also include guidelines for the safe use of the equipment by the operators, although these are recommendations only and are not legally binding. The ACGIH has established TLVs for laser exposures as well as for chemical exposures.

The NRC does not regulate radiation exposures due to x-rays or the use of radioactive materials not related to by-product materials (materials made radioactive by reactor operations). However, dose rates for x-rays are generally equivalent to those for radiation from radioactive materials. Also, states typically regulate the radioactive materials not covered by the NRC in a comparable fashion. In January 1994 the NRC regulations were changed in several significant ways. For example, exposure limits, which had been primarily concerned with external exposures, now must incorporate exposures due to ingestion, inhalation, and possible percutaneous absorption; certain external limits were changed, including those related to the extremities, which now include the area below the knee as well as the areas beyond the elbows, since there are no critical organs within those bodily components; and women who may be pregnant (and so declare) must be limited to 10% of the whole body exposure permissible for other monitored employees, in order to limit the exposure to the fetus. The NRC has also stepped up the stringency with which it monitors compliance, and in recent years it has become commonplace for the NRC to impose substantial fines on industrial users, hospitals, and universities.

Also in 1994 the EPA activated a law which had been on the books since 1989 but had not been enforced. This standard limits the dispersion of radioactive materials into the atmosphere if it is possible that they may be taken into the body directly or through the food chain. This standard has a potential impact on the research use of radioactive materials since much of the common applications of radioactive materials are in research institutions, where radioactive materials are emitted in fume hood exhausts. A radioisotope in very common use is  $^{125}\text{I}$ . This isotope is readily taken up by the body and is concentrated in the thyroid. For operations that are limited due to this regulation, it is apt to be due to this isotope.

There are other standards and guidelines which have a significant impact on safety and health. Among these are the OSHA Hazard Communication Act (29 CFR 1910.1200) for industrial users of hazardous materials, and its companion for laboratory employees (OSHA 29 CFR 1450). Both require extensive training of employees and provision for responses to emergencies. The previously mentioned Bloodborne Pathogens Act (OSHA 29 CFR 1919.1030) is a very strong standard. All of these incorporate significant engineering requirements to protect employees through either provision of emergency measures or uses of protective equipment.

The National Institutes of Health (NIH) Guidelines for Research Involving Recombinant DNA Molecules represent a significant standard which is, in effect, a regulation although it is labeled a guideline. Research programs which do not comply with these guidelines will not receive funding from NIH. Characteristics of the facility in which the research is done, the procedures which are followed, and the experimental subjects are all carefully defined by the standard. Further, any genetically engineered products of these researches must receive explicit approval prior to being placed on the market. In a similar vein, the Centers for Disease Control has published standards governing microbiological research practices based on concerns that not only research workers might contract a contagious disease. This care is not due to a high probability that contagion may spread beyond the facility, as in fact this has not been known to occur in this country, but is designed to ensure that it does not. A disturbing trend is that our disease-fighting capabilities are rapidly decreasing in efficacy as infectious organisms become immune to previously effective treatments.

## 180.2 Summary

---

The preceding material represents a very brief overview of the regulatory environment. Any one of the topics mentioned would justify a fully detailed exposition in a book, if not an entire book, to be covered adequately. There are topics which were not mentioned at all that are of comparable importance to those covered. Safety and health programs, once developed only voluntarily by enlightened organizations, due to their recognition that it was cost-effective, are now virtually completely defined by regulations and standards applicable to the substantial majority of employees. Engineering subdisciplines have developed to provide the supportive infrastructure for the regulations. This is a rapidly developing field. Many of the regulations cited did not exist five years ago, and every one of them has undergone significant changes in that time.

There is a tendency to believe that regulations stifle productivity. This may be so if the regulations are applied inflexibly, without common sense. A sensible safety program will without doubt improve productivity. One of the country's most profitable companies is E. I duPont, which pioneered many of the safety programs later incorporated into standards. They are generally regarded as having one of the very best safety and health programs in industry. They are not doing it for their health or to be nice. They know their program is cost-effective.

## References

No date is given for the citations of regulatory standards. These undergo continual revision. The most current version should always be used.

Department of Health and Human Services, Centers for Disease Control, and National Institutes of Health. *Biosafety in Microbiological and Biomedical Laboratories*. U.S. Government Printing Office, Washington, DC.

Department of Transportation. Hazardous Materials, Substances, and Waste Regulations. Title 49, Code of Federal Regulations. U.S. Government Printing Office, Washington, DC.

Environmental Protection Agency. Hazardous Waste Regulations. Title 40, Code of Federal Regulations. U.S. Government Printing Office, Washington, DC.

*Industrial Ventilation: A Manual of Recommended Practice*. American Conference of Governmental Industrial Hygienists, Cincinnati, OH.

Nuclear Regulatory Commission. Radiation Safety Regulations. Title 10, Code of Federal Regulations. U.S. Government Printing Office, Washington, DC.

Occupational Health and Safety Administration. Title 29, Code of Federal Regulations. U.S. Government Printing Office, Washington, DC.

*Threshold Limit Values for Chemical Substances and Physical Agents and Biological Exposure Indices*. American Conference of Governmental Industrial Hygienists, Cincinnati, OH.

## Further Information

For information on the topic of regulations in safety, there is, of course, no complete substitute for the regulations themselves. They are being changed on almost a daily basis. In addition, the regulations themselves are being interpreted by the administering agencies. Each agency publishes an annual update, and the *Federal Register* is available as printed text and is now available on CD-ROM and as an on-line electronic service. The following two references provide a useful means of keeping track of the changes and their interpretations. The second one lists equipment which is available to comply with the standards, and sources for the equipment.

Bureau of National Affairs, Rockville, MD: Publications covering occupational safety, environmental safety and health, chemical safety, and hazardous waste management, providing frequent regular updates.

*Best's Safety Directory*. A. M. Best Company, Oldwick, NJ.

Au, T. "Engineering Economics and Management"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



"An Act to Promote the Progress of Useful Arts," was the first patent statute in the United States. It was signed by George Washington on April 10, 1790 and became law. The intention of the Patent Act was to protect inventor's rights to their inventions. Initially, each inventor was required to submit a drawing, description, and a miniature working model built to scale. Three cabinet heads examined each application and granted patents "for any such useful art, manufacture, engine, machine, or device as deemed sufficiently useful and important." The system quickly became so cumbersome that it was dissolved three years later.

A second Patent Act was passed in 1793 which eliminated the examination procedure and established a simple registration system and a flat \$30 fee. This system automatically granted patents; however, the merit of the invention was deferred. Eli Whitney's 1794 patent for the cotton gin made him famous, but he



realized few financial rewards because the new patent system failed to protect the right of the inventor. This system remained in place for 43 years.

The Patent Act of July 4, 1836 reinstated the American patent system and provided protection for investors and their inventions. In a speech given in 1859, Abraham Lincoln said the U.S. patent system "added the fuel of interest to the fire of genius." Today, the Patent and Trademark Office operates under the same examination system.

The photograph is of the old Patent Office Building around 1917. The Patent Office remained in this building until 1932. Today, the U.S. Patent and Trademark Office is located at the Crystal Plaza in Arlington, Virginia. (Photo courtesy of the U.S. Patent and Trademark Office.)

# XXVIII

## Engineering Economics and Management

---

**Tung Au**

*Carnegie Mellon University*

**181 Present Worth Analysis** *W. D. Short*

Calculation • Application • Other Considerations

**182 Project Analysis Using Rate-of-Return Criteria** *R. G. Beaves*

Net Present Value • Internal Rate of Return • Overall Rate of Return • Project Investment Base • Scale-Adjusted ORR • Project Life Differences • Conclusion

**183 Project Selection from Alternatives** *C. Hendrickson and S. McNeil*

Problem Statement for Project Selection • Steps in Carrying Out Project Selection • Selection Criteria • Applications • Conclusion

**184 Depreciation and Corporate Taxes** *T. Au*

Depreciation as Tax Deduction • Tax Laws and Tax Planning • Decision Criteria for Project Selection • Inflation Consideration • After-Tax Cash Flows • Evaluation of After-Tax Cash Flows • Effects of Various Factors

**185 Financing and Leasing** *C. Fazzi*

Debt Financing • Equity Financing • Leasing

**186 Risk Assessment** *R. T. Ruegg*

Expected Value (EV) Analysis • Mean-Variance Criterion (MVC) and Coefficient of Variation (CV) • Risk-Adjusted Discount Rate (RADR) Technique • Certainty Equivalent (CE) Technique • Simulation Technique • Decision Analysis

**187 Sensitivity Analysis** *H. E. Marshall*

Sensitivity Analysis Applications • Advantages and Disadvantages

**188 Life-Cycle Costing** *W. J. Fabrycky and B. S. Blanchard*

The Life-Cycle Costing Situation • Cost Generated over the Life Cycle • The Cost Breakdown Structure • Life-Cycle Cost Analysis • Cost Treatment over the Life Cycle • Summary

**189 Project Evaluation and Selection** *H. J. Thamhain*

Quantitative Approaches to Project Evaluation and Selection • Qualitative Approaches to Project

Evaluation and Selection • Recommendations for Effective Project Evaluation and Selection

190 **Critical Path Method** *J. L. Richards*

Planning the Project • Scheduling the Project • Controlling the Project • Modifying the Project Schedule • Project Management Using CPM

191 **Patents, Copyrights, Trademarks, and Licenses** *P. A. Beck and C. I. Bordas*

Patents • Copyrights • Trademarks • Licenses

THE SUCCESS OF ANY ENGINEERING PROJECT for production or service is inextricably linked to economics and management issues related to that project. Engineers who act as project managers are often called upon to make informed economic and managerial decisions. This section intends to help readers to understand the basic concepts and analytical tools of engineering economics and management.

Investment in a capital project involves the commitment of resources of an organization in anticipation of greater returns in the future. Based on the generally accepted objective of maximizing the return, a merit measure is identified and a decision criterion is established for evaluation of the merit of a project. The two most common merit measures are present worth (also called net present value) and internal rate of return. The present worth analysis is based on the net value of a project calculated by discounting all cash flows over time to the present according to a predetermined discount rate. The internal rate of return is based on the average annual percentage return over the life of an investment. For either merit measure, a corresponding decision criterion must be used for project selection. More generally, decision rules are introduced for three classes of project selection problems: to accept or reject a proposed project, to select the best among a group of mutually exclusive alternatives, and to select a group of noncompeting projects under budget constraint.

The basic concepts have been extended to cover real issues in the business world. One aspect is depreciation and corporate taxes imposed by tax laws and regulations. Another is the financing decision to obtain funds for a project either through issuing debt and equity securities (bonds and stocks) in the capital market or through leasing to avoid a large capital outlay. The decision maker must also consider the uncertainty of various factors in economic analysis. Risk assessment provides information about the risk exposure inherent in a given decision and takes into consideration the risk attitude of the decision maker. Sensitivity analysis allows the decision maker to measure the impact resulting from changing values of uncertain variables on the investment returns.

Application of principles of economic analysis to engineering production and service requires the consideration of operational requirements, performance and maintenance effectiveness, production quantity, utilization factors, logistic support, and disposal. Life-cycle costing involves such decisions to minimize the discounted cost from the life cycle of a project. Evaluation of complex projects from product development to organizational improvement requires the integration of both analytical and judgmental techniques in order to select and fund the best project. Another analytical tool for planning, scheduling, and controlling time and cost of a project is the critical path method, which utilizes a network-based model to represent the interrelationships of project activities.

The profit of an organization may also be enhanced by royalties derived from intellectual properties such as patents, copyrights, trademarks, and licenses, which are protected by intellectual

property laws. The owner of a copyrighted product can enter into a license agreement in which the licensee can distribute and reproduce the product in return for royalties payable to the licensor-owner.

Short, W. D. "Present Worth Analysis"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Present Worth Analysis

---

### 181.1 Calculation

Analysis Period • Cash Flows • Discount Rate

### 181.2 Application

### 181.3 Other Considerations

Savings Versus Income/Returns • Uncertainty • Externalities

### Walter D. Short

*National Renewable Energy Laboratory*

Evaluation of any project or investment is complicated by the fact that there are usually costs and benefits (i.e., cash flows) associated with an investment that occur at different points in time. The typical sequence consists of an initial investment followed by operations and maintenance costs and returns in later years. Present worth analysis is one commonly used method that reduces all cash flows to a single equivalent cash flow or dollar value (Palm and Qayum, 1985). If the investor did not care when the costs and returns occurred, **present worth** (also known as *net present value*) could be easily calculated by simply subtracting all costs from all income or returns. However, to most investors the timing of the **cash flows** is critical due to the time value of money; that is, cash flows in the future are not as valuable as the same cash flow today. Present worth analysis accounts for this difference in values over time by discounting future cash flows to the value in a base year, which is normally the present (hence the term "present" worth). If the single value that results is positive, the investment is worthwhile from an economic standpoint. If the present worth is negative, the investment will not yield the desired return as represented by the discount rate employed in the present worth calculation.

### 181.1 Calculation

---

The present worth (PW), or net present value, of an investment is

$$PW = \sum_{t=0}^N \frac{C_t}{(1+d)^t} \quad (181.1)$$

where  $N$  is the length of the analysis period,  $C_t$  is the net cash flow in year  $t$ , and  $d$  is the **discount rate**.

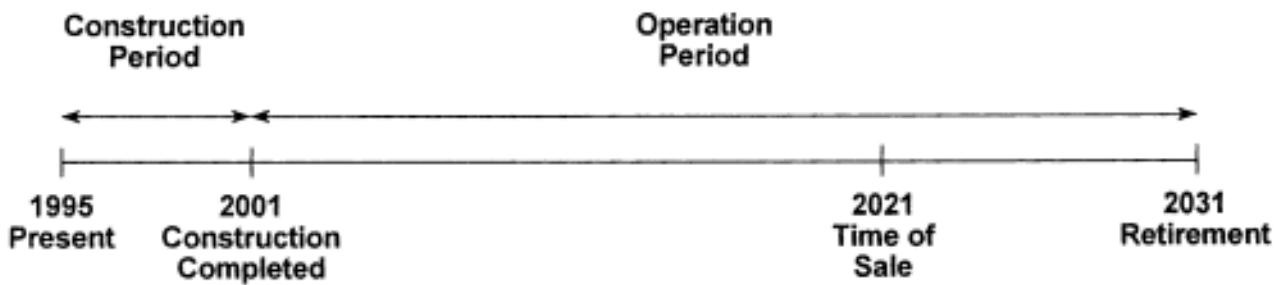
## Analysis Period

The analysis period must be defined in terms of its first year and its length, as well as the length of each time increment. The standard approach, as represented in Eq. (181.1), is to assume that the present is the beginning of the first year of the analysis period [ $t = 0$  in Eq. (181.1)], that cash flows will be considered on an annual basis, and that the analysis period will end when there are no more cash flows that result from the investment (e.g., when the project is completed or the investment is retired) [Ruegg and Petersen, 1987]. For example, the present worth for a refinery that will require 6 years to construct and then last for 30 years (to 2031) can be represented as

$$PW(1995) = \sum_{t=0}^{36} \frac{C_t}{(1+d)^t} \quad (181.2)$$

In this case the analysis period covers both the construction period and the full operation period, as shown in Fig. 181.1. In Eq. (181.2) the parenthetical (1995) explicitly presents the base year for which the present worth is calculated. This designation is generally omitted when the base year is the present.

**Figure 181.1** Analysis period for a refinery.



However, any of these assumptions can be varied. For example, the present can be assumed to be the first year of operation of the investment. In the refinery example the present worth can be calculated as if the present were 6 years from now (2001)—that is, the future worth in the year 2001—with all investment costs between now and then accounted for in a single turnkey investment cost ( $C_0$ ) at that time ( $t = 0$ ). In this case the base year is 2001.

$$PW(2001) = \sum_{t=0}^{30} \frac{C_t}{(1+d)^t} \quad (181.3)$$

The length of the analysis period is generally established by the point in time at which no further costs or returns can be expected to result from the investment. This is typically the point at which the useful life of the investment has expired. However, a shorter lifetime can be assumed with the cash flow in the last year accounting for all subsequent cash flows. For the refinery example the analysis period could be shortened from the 36-year period to a 26-year period, with the last year

capturing the salvage value of the plant after 20 years of operation.

$$PW = \sum_{t=0}^{26} \frac{C_t}{(1+d)^t} \quad (181.4)$$

where  $C_{26}$  is now the sum of the actual cash flows in  $C_{26}$  and the salvage value of the refinery after 20 years of operation. One common reason for using a shorter analysis period is to assess the present worth, assuming sale of the investment. In this case the salvage value would represent the price from the sale minus any taxes paid as a result of the sale.

Finally, the present worth is typically expressed in the dollars of the base year. However, it can be expressed in any year's dollars by adjusting for inflation. For example, in the refinery case, in which the base year is 2001 [Eq. (181.3)], the present worth value expressed in the dollars of the year 2001 can be converted to a value expressed in today's (1995) dollars, as follows:

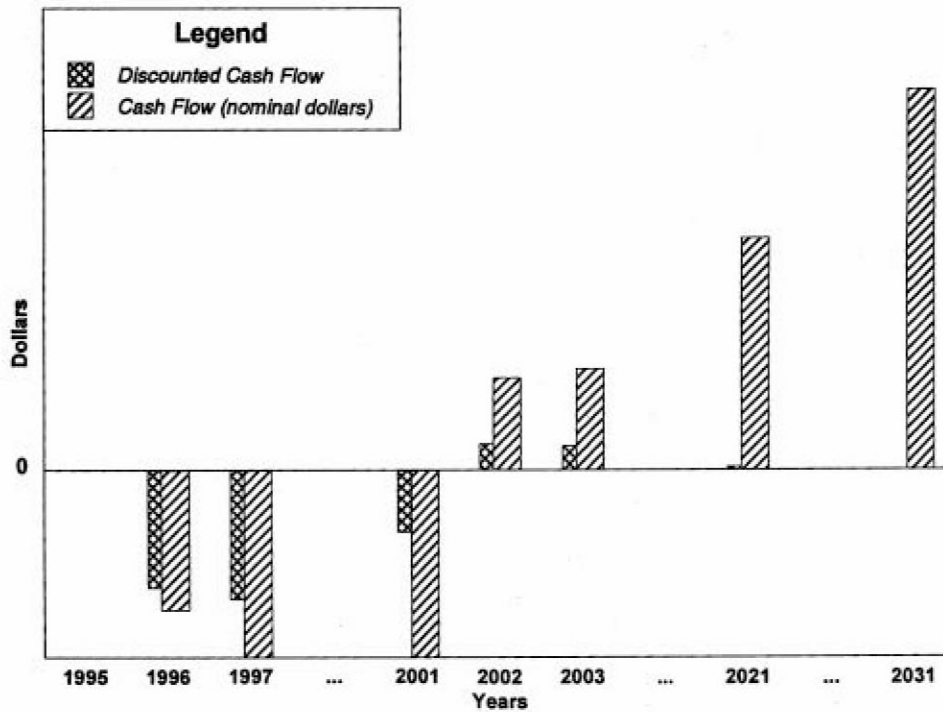
$$PW_{1995\$}(2001) = \frac{1}{(1+i)^{2001-1995}} PW_{2001\$}(2001) \quad (181.5)$$

where  $i$  is the annual **inflation rate** between 1995 and 2001 and the subscripts indicate the year of the dollars in which the present worth is presented.

## Cash Flows

Cash flows include all costs and returns where costs are negative and returns or income are positive (see [Fig. 181.2](#)). The cash flow for each year of the analysis period must include all actual costs and income associated with the investment during that year. Thus, cash flows must include all initial capital costs, all taxes, debt payments, insurance, operations, maintenance, income, and so forth.

**Figure 181.2** Cash flows.



Cash flows are typically expressed in terms of the actual dollar bills paid or received (**nominal** or current **dollars**). Alternatively, cash flows can be expressed in terms of the dollars of a base year (**real** or constant **dollars**). Nominal dollar cash flows,  $C^n$ , can be converted to real dollar cash flows,  $C^r$ , (and vice versa) by accounting for the effects of inflation

$$C_0^r = \frac{C_t^n}{(1+i)^t} \quad (181.6)$$

where  $i$  is the annual inflation rate and  $t$  is the difference in time between the base year ( $t = 0$ ) and the year of the nominal dollar cash flow.

A discounted cash flow is the present worth of the individual cash flow. Discounted cash flow in period  $s$  equals

$$\frac{C_s}{(1+d)^s} \quad (181.7)$$

As shown in [Fig. 181.2](#) (which used a nominal discount rate of 0.2 or 20%), discounting can significantly reduce the value of cash flows in later years.

## Discount Rate

The discount rate is intended to capture the time value of money to the investor. The value used varies with the type of investor, the type of investment, and the opportunity cost of capital (i.e., the



returns that might be expected on other investments by the same investor). In some cases, such as a regulated electric utility investment, the discount rate may reflect the weighted average cost of capital (i.e., the after-tax average of the interest rate on debt and the return on common and preferred stock). Occasionally, the discount rate is adjusted to capture risk and uncertainty associated with the investment. However, this is not recommended; a direct treatment of uncertainty and risk is preferred (discussed later).

As with cash flows, the discount rate can be expressed in either nominal or real dollar terms. A nominal dollar discount rate must be used in discounting all nominal dollar cash flows, whereas a real discount rate must be used in discounting all real dollar cash flows. As with cash flows, a nominal dollar discount rate,  $d_n$ , can be converted to a real dollar discount rate,  $d_r$ , by accounting for inflation with either of the following:

$$1 + d_n = (1 + d_r)(1 + i)$$

$$d_r = \frac{1 + d_n}{1 + i} - 1 \quad (181.8)$$

Since cash flow should include the payment of taxes as a cost, the discount rate used in the present worth calculation should be an after-tax discount rate (e.g., the after-tax opportunity cost of capital).

## 181.2 Application

---

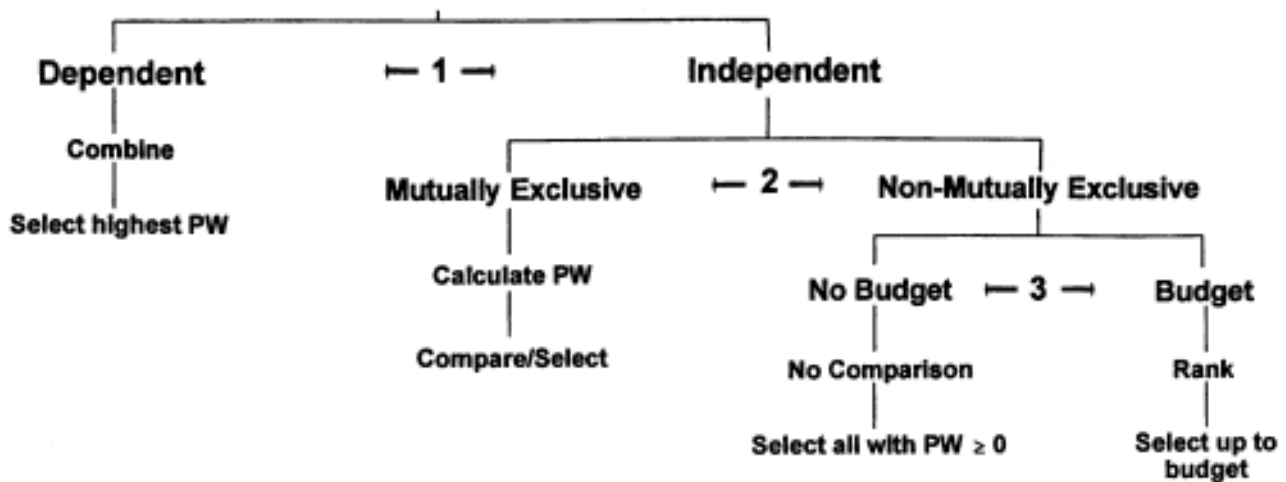
Present worth analysis can be used to evaluate a single investment or to compare investments. As stated earlier, a single investment is economic if its present worth is equal to or greater than zero but is not economic if the present worth is less than zero.

PW > 0. The investment is economic.

PW < 0. The investment is not economic.

The use of present worth to compare multiple investments is more complex. There are several possible situations, as shown in Fig. 181.3, including investments that impact one another (dependence, point 1 on Fig. 181.3), investments that are mutually exclusive (only one investment is possible from a set of investments, point 2 on Fig. 181.3), and/or investments for which there is a budget limitation (point 3 on Fig. 181.3).

**Figure 181.3** Comparison of investments.



If the investments do have an impact on one another (i.e., they are dependent), then they should be considered both together and individually. Investments that impact one another are reevaluated as a package, with new cash flows representing the costs and returns of the combination of investments. The present worth of all individual investments and all combinations of dependent investments should then be compared. This comparison will, by definition, include some mutually exclusive investments (i.e., they cannot all be undertaken because some are subsumed in others). Each set of these mutually exclusive alternatives can be resolved to a single alternative by selecting the combination or individual investment with the highest present worth from each set. For example, if a refinery and a pipeline are being considered as possible investments and the pipeline would serve the refinery as well as other refineries, then the presence of the pipeline might change the value of the refinery products (i.e., the two investments are not independent). In this case the present worth of the refinery alone, the present worth of the pipeline alone, and the present worth of the combined pipeline and refinery should be calculated and the one with the highest present worth selected.

Once each set of dependent investments has been resolved to a single independent investment, then the comparison can proceed. If only one investment can be made (i.e., they are mutually exclusive), then the comparison is straightforward; the present worth of each investment is calculated and the investment with the highest present worth is the most economic. This is true even if the investments require significantly different initial investments, have significantly different times at which the returns occur, or have different useful lifetimes. Examples of mutually exclusive investments include different system sizes (e.g., three different refinery sizes are being considered for a single location), different system configurations (e.g., different refinery configurations are being considered for the same site), and so forth.

If the investments are not mutually exclusive, then one must consider whether there is an overall budget limitation that would restrict the number of economic investments that might be undertaken. If there is no budget (i.e., no limitation on the investment funds available) and the investments have no impact on one another, then there is really no comparison to be performed and the investor simply undertakes those investments that have positive present worth and discards those that do not.

If the investments are independent but funds are not available to undertake all of them (i.e., there is a budget), then there are two approaches. The easiest approach is to rank the alternatives, with the best having the highest benefit-to-cost ratio or savings-to-investment ratio. (The investment

with the highest present worth will not necessarily be the one with the highest rank, since present worth does not show return per unit investment.) Once ranked, those investments at the top of the priority list are selected until the budget is exhausted. Present worth can be used in the second, less desirable approach by considering the total present worth of each combination of investments whose total initial investment cost is less than the budget. The total present worth of the investment package is simply the sum of the present worth values of all the independent investments in the package. That combination with the greatest present worth is then selected.

## 181.3 Other Considerations

---

Three commonly encountered complications to the calculation of present worth are savings versus returns, uncertainty, and externalities.

### Savings Versus Income/Returns

In many, if not most, investments, the positive cash flows or benefits are actual cash inflows that result from sales of products produced by the investment. However, there are also a large number of investments in which the positive cash flows or benefits are represented by savings. For example, the benefit of investing in additional insulation for steam pipes within a refinery is savings in fuel costs (less fuel will be required to produce steam). These savings can be represented in a present worth evaluation of the insulation as positive cash flows (i.e., as income or returns).

### Uncertainty

In the present worth discussion up to this point, it has been assumed that all the input values to the present worth calculation are known with precision. In fact, for most investments there is considerable uncertainty in these values—especially the future cash flows. The preferred method for including such uncertainties in the calculation of present worth is to estimate the probability associated with each possible cash flow stream, calculate the present worth associated with that cash flow stream, and assign the probability of the cash flow to the present worth value. This will produce one present worth value and one probability for each possible cash flow stream (i.e., a probability distribution on the present worth of the investment). This distribution can then be used to find statistics such as the expected present worth value, the standard deviation of the present worth value, confidence intervals, and so forth. These statistics, especially the expected present worth value and confidence intervals, can then be used in the decision process.

### Externalities

In many cases not all costs and benefits are included in the cash flows because they are not easily quantified in terms of dollars or because they do not benefit or cost the investor directly. Such costs and benefits are referred to as *externalities* because they are generally considered external to the

direct economic evaluation (i.e., they are not included in the present worth calculation). For example, the cost of air emissions from a refinery are often not considered in a present worth calculation, even though they impact the local community as well as the employees of the refinery. Such emissions may result in lost work days and more sick pay, as well as the loss of the local community's goodwill to the refinery, making future refinery expansion more difficult. Likewise, emissions may affect the health and quality of life of local residents.

Externalities can be considered qualitatively, along with measures such as the present worth, when evaluating an investment. Or externalities can be explicitly considered within a present worth calculation by estimating their costs and benefits in terms of dollars and including these dollars in the present worth cash flows. In this case these costs and benefits are said to have been *internalized* and are no longer externalities.

## Defining Terms

**Analysis period:** The period during which the investor will consider the costs and benefits associated with an investment. This period is usually determined by the period during which the investor will develop, operate, and own the investment.

**Cash flow:** The dollar value of all costs and benefits in a given period. Cash flows are normally expressed in nominal dollars, but can be expressed in real dollars.

**Discount rate:** The interest rate that represents the time value of money to the investor. For most investors this is the opportunity cost of capital (i.e., the rate of return that might be expected from other opportunities to which the same capital could be applied). Discount rates can include inflation (nominal discount rate) or exclude inflation (real discount rate).

**Inflation rate:** The annual rate of increase in a general price level, frequently estimated as the gross domestic product deflator or the gross national product deflator. Estimates of future value are generally provided by macroeconomic forecasting services.

**Nominal dollars:** Current dollars; dollars in the year the cost or benefit is incurred (i.e., number of dollar bills).

**Present worth:** Net present value. The sum of all cash flows during the analysis period discounted to the present.

**Real dollars:** Constant dollars; value expressed in the dollars of the base year. Real dollars are the value excluding inflation after the base year.

## References

- Palm, T. and Qayum, A. 1985. *Private and Public Investment Analysis*. Southwestern, Cincinnati, OH.
- Ruegg, R. and Marshall, H. 1990. *Building Economics: Theory and Practice*. Van Nostrand Reinhold, New York.
- Ruegg, R. and Petersen, S. 1987. *Comprehensive Guide for Least-Cost Energy Decisions*. National Bureau of Standards, Gaithersburg, MD.

## Further Information

- Au, T. and Au, T. P. 1983. *Engineering Economics for Capital Investment Analysis*. Allyn & Bacon, Boston, MA.
- Brown, R. J. and Yanuck, R. R. 1980. *Life Cycle Costing: A Practical Guide for Energy Managers*. Fairmont Press, Atlanta, GA.
- National Renewable Energy Laboratory. 1993. *A Manual for the Economic Evaluation of Energy Efficiency and Renewable Energy Technologies*. Draft.
- Samuelson, P. A. and Nordhaus, W. D. 1985. *Economics*, 12th ed. McGraw-Hill, New York.
- Stermole, F. J. 1984. *Economic Evaluation and Investment Decision Methods*, 5th ed. Investment Evaluations Corporation, Golden, CO.
- Weston, J. F. and Brigham, E. F. 1981. *Managerial Finance*, 7th ed. Dryden Press, Fort Worth, TX.

Beaves, R. G. "Project Analysis Using Rate-of-Return Criteria"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Project Analysis Using Rate-of-Return Criteria

---

- 182.1 Net Present Value
- 182.2 Internal Rate of Return
- 182.3 Overall Rate of Return
- 182.4 Project Investment Base
- 182.5 Scale-Adjusted ORR
- 182.6 Project Life Differences
- 182.7 Conclusion

**Robert G. Beaves**

*Robert Morris College*

Many decision makers find rate-of-return investment criteria to be more intuitive and therefore easier to understand than net present value (NPV). As a result the internal rate of return (IRR) continues to be widely used despite its legendary quirks and the well-established superiority of NPV. A second rate-of-return criterion, the overall rate of return (ORR), provides more NPV consistency than does the IRR but is not widely known or used. This chapter compares NPV, IRR, and ORR and demonstrates some shortcomings of rate criteria.

## 182.1 Net Present Value

---

A project's **net present value** (NPV) represents *the change in the value of the firm* that occurs if and when the firm implements that project. The NPV of a project is calculated by summing the present values of all cash flows associated with it while preserving negative signs of flows to the project and positive signs of flows from the project. NPV can be defined as follows:

$$\text{NPV} = \sum_{t=0}^n a_t(1+k)^{-t} \quad (182.1)$$

where

- $a_t$  = the cash flow at the time  $t$  (end of period  $t$ )
- $k$  = the firm's opportunity cost; the discount or "hurdle" rate
- $n$  = the number of periods in the project's life.

Implementation of projects with positive NPVs increases firm value, whereas implementation of projects with negative NPVs reduces it. Because management's goal is to maximize firm value, a project should be rejected if its NPV is negative and accepted otherwise. When forced to choose among competing positive-NPV projects, the project that offers the highest NPV should be preferred. It is well established that the NPV criterion provides theoretically correct accept/reject and project-ranking decisions.

## 182.2 Internal Rate of Return

A project's **internal rate of return** (IRR) is that discount rate  $k^*$  for which that project's NPV would be zero. Note that a project's IRR is independent of the firm's actual hurdle rate  $k$ . A project's IRR represents *the average periodic rate of return earned on funds while they are invested in that project*. Because IRR is an average rather than a constant rate, the rate or return generated by a project in any one period need not equal its IRR.

Consider project A in [Table 182.1](#), which has an IRR of 12.48%. Project A's rate of return during the first period (time 0 to time 1) is unknown. The "balance" in project A at time 1 need not be \$1125 (\$700 released to the firm and \$525 remaining in the project) as would be the case if its 12.48% IRR was a constant rate. Although a project's IRR represents the average rate of return earned on funds invested in that project, the amount of funds so invested varies from period to period and is often unknown.

**Table 182.1** Projects A, B, and C\*

	Project A	Project B	Project B–A	Project C
Time				
0	–1000	–1000	0	–2000
1	700	0	–700	1400
2	300	0	–300	600
3	200	1423	1223	400
Method				
IRR	12.48%	12.48%	12.48%	12.48%
NPV	\$34.56	\$69.12		\$69.12
IB	\$1000.00	\$1000.00		\$2000.00
ORR	11.25%	12.48%		11.25%
ORR <sub>2000</sub>	10.63%	11.25%		11.25%

\* $k = 10\%$

A project should be rejected if its IRR is less than the firm's **opportunity cost**  $k$  and accepted otherwise. NPV and IRR provide the same accept/reject decisions for projects that have unique IRRs. The use of IRR for accept/reject decisions is complicated, however, by the fact that some projects have multiple IRRs whereas others have none.

It is well established that ranking competing projects on the basis of their IRRs can provide an incorrect choice from among those projects. Consider projects A and B in [Table 182.1](#). Both projects require initial investments of \$1000 and both have the same IRR, 12.48%. No clear



preference exists between projects A and B if ranked on the basis of their IRRs. Nonetheless, project A's NPV is only \$34.56, whereas project B's is \$62.12. Project B has a high NPV because its entire initial investment and all accumulated returns on that investment remain in the project, earning an average 12.48% rate for three periods. On the other hand, much of project A's initial investment and accumulated returns are released before the end of its 3-period life.

Any two projects can be correctly ranked by examining the IRR of an incremental or "difference" project. Project B–A in [Table 182.1](#) is an incremental project for projects A and B created by subtracting each cash flow of A from the respective cash flow of B. Because the IRR of incremental project B–A exceeds the firm's 10% opportunity cost, project B (the project from which a second project was subtracted) should be preferred over A. If the IRR of the incremental project was less than  $k$ , project A would have been preferred. Because incremental IRR analysis can only provide pair-wise rankings, its use becomes tedious where more than two projects are being ranked. Further, an incremental project may have multiple IRRs or no IRR at all.

## 182.3 Overall Rate of Return

---

A project's **overall rate of return** (ORR) represents the *average periodic rate earned on a fixed investment amount* over the expected life of the project. That fixed investment amount is known as the project's **investment base**. A project's ORR can be defined as:

$$\text{ORR} = \left[ \frac{(\text{IB} + \text{NPV}) \times (1 + k)^n}{\text{IB}} \right]^{1/n} - 1.0 \quad (182.2)$$

where

NPV = the project's NPV, with  $k$  as discount rate

IB = the project's investment base; to be defined in greater detail later.

In calculating a project's ORR, that project's investment base plus any accumulated returns are assumed to earn the project's IRR while invested in it and to earn the firm's opportunity cost when not needed by the project. Thus project B in [Table 182.1](#) has an ORR of 12.48% because, for its entire 3-period life, its \$1000 investment base and all accumulated returns remain invested in the project earning its 12.48% IRR. Project A's ORR is less because significant amounts of its investment base and accumulated returns are released at times 1 and 2 and earn the firm's 10% opportunity cost for the balance of that project's life.

Projects with ORRs greater than or equal to the firm's opportunity cost  $k$  should be accepted, whereas those with lower ORRs should be rejected. As Eq. (182.2) suggests, the ORR always provides accept/reject decisions that are consistent with those provided by NPV. In contrast to the IRR, the ORR is uniquely defined for all projects and is generally a function of the firm's opportunity cost  $k$ . Further, the ORR provides an NPV-consistent ranking of competing projects that have the same scale or investment base IB.

When comparing projects having different investment bases, ORRs adjusted to some common

scale will provide an NPV-consistent ranking. Scale adjustments are required because the ORR, like all rate criteria, does not preserve scale. Consider project C (Table 182.1), which was created by doubling project A's cash flows. Although a project's NPV doubles when its size is doubled, its IRR and ORR are unchanged. Scale is lost in the calculation of any rate of return, an important shortcoming in ranking competing projects.

## 182.4 Project Investment Base

A project's investment base represents the time 0 value of all external funds (i.e., funds not provided by the project itself) needed to finance the project. All projects considered thus far required a single cash inflow (i.e., negative flow), which occurred at time 0. The investment base of such projects is simply the amount of that initial inflow,  $-a_0$ . Because some projects have more complex cash flow patterns, it is necessary to establish the following general rule for determining a project's investment base:

A project's investment base is determined by multiplying the minimum cumulative present value associated with its cash flows times  $-1.0$ .

Consider project D in Table 182.2, for which the firm's opportunity cost is assumed to be 10%. Using 10% as the discount rate, present values and cumulative present values are calculated for each of project D's cash flows. The minimum cumulative present value associated with project D's cash flows is  $-\$1057.85$ . Thus project D's investment base is  $\$1057.85$ . Assuming a 10% opportunity cost for all funds not currently needed by project D,  $\$1057.85$  is the minimum time 0 amount that will fund project D to its termination at time 5.

**Table 182.2** Project D\*

	Cash Flow ( $a_t$ )	Present Value	Cumulative PV
Time			
0	-1000	-1000.00	-1000.00
1	300	272.73	-727.27
2	-400	-330.58	-1057.85
3	700	525.92	-531.93
4	-500	-341.51	-873.44
5	1500	931.38	57.94
Method			
IRR	11.48%		
NPV	\$57.94		
IB	\$1057.85		
ORR	11.18%		

\* $k = 10\%$

Note that the final cumulative present value listed in Table 182.2 is necessarily project D's NPV of  $\$57.94$ . Project D's ORR is calculated by substituting into Eq. (182.2) as follows:

$$\left[ \frac{(57.94 + 1057.85) \times (1.10)^5}{1057.85} \right]^{1/5} - 1.0 = .1118$$

In other words, \$1057.85 committed to project D at time 0 will earn an average return of 11.18% per period for five periods.

## 182.5 Scale-Adjusted ORR

---

ORR as defined in Eq. (182.2) should only be used to compare projects of identical scale (i.e., the same investment base). Assuming projects B and C in [Table 182.1](#) are mutually exclusive, they cannot be directly compared, because project C's 11.25% ORR is earned on an investment base of \$2000, whereas project B's 12.48% ORR is earned on an investment base of only \$1000. A project's ORR can be adjusted to any scale by replacing its investment base (IB) in Eq. (182.2) with the desired scale as follows:

$$\text{ORR}_S = \left[ \frac{(S + \text{NPV}) \times (1 + k)^n}{S} \right]^{1/n} - 1.0 \quad (182.3)$$

where  $\text{ORR}_S$  is the project's ORR adjusted to scale  $S$  and  $S$  is the desired scale to which the project's ORR is being adjusted.

These scale adjustments are based on the assumption that differences in project scales can be invested at the firm's opportunity cost  $k$ . Investing at the firm's opportunity cost increases a project's scale but does not affect its NPV. Where two or more competing projects are being compared, the ORRs of all must be adjusted to a common scale (preferably the largest investment base among the projects). ORRs adjusted to a common scale of \$2000 were calculated in [Table 182.1](#) for projects A, B, and C. Those ORRs reveal that the firm should be indifferent in choosing between projects B and C but should prefer either to project A, whose  $\text{ORR}_{2000}$  is 10.63%. These preferences are consistent with the NPV ranking of projects A, B, and C.

## 182.6 Project Life Differences

---

Comparing projects that have different project lives creates certain problems—no matter what decision criterion is being used. Such comparisons generally require an assumption as to what occurs at the end of each project's life—that is, whether the project is or is not replaced with a similar project. Where all competing projects are one-time expenditures with no replacement, their NPVs can be directly compared even if they have different project lives. Such comparison assumes that funds released by any project earn the firm's opportunity cost  $k$  until the end of the longest-lived project. In contrast, the ORRs of competing one-time projects cannot be directly compared if those projects have different lives. Rate criteria such as ORR and IRR measure average performance per period over a project's life, whereas NPV measures cumulative performance over a project's life.

Consider projects E and F in [Table 182.3](#). If we assume that both projects are one-time

expenditures that will not be replaced at the end of their respective lives, the ORRs of projects E and F are not directly comparable. Project E earns an average rate of 13.05% per period for 3 periods, whereas project F earns an average rate of 11.92% per period for 5 periods. The ORR of any project can be calculated over a common life  $z$  periods longer than its project life as follows, by assuming that all funds can be invested at the firm's opportunity cost  $k$  during those  $z$  additional periods:

$$\text{ORR}_{S, n+z} = \left[ \frac{(S + \text{NPV}) \times (1 + k)^{(n+z)}}{S} \right]^{1/(n+z)} - 1.0 \quad (182.4)$$

where  $\text{ORR}_{S, n+z}$  is the project's ORR adjusted to common scale  $S$  and common life  $n + z$ , and  $n + z$  is the common life over which the project's ORR is being calculated. Adjusted to project F's 5-period life, project E's ORR is 11.82% and is less than project F's ORR of 11.92%. Thus, ORRs calculated over a common 5-period life provide the same ranking of projects E and F as is provided by their NPVs.

**Table 182.3** Projects E and F\*

		Project E	Project F
Time	0	−5000	−4000
	1	2500	−1100
	2	2000	2000
	3	2000	2000
	4		2000
	5		1500
Method	IRR	15.02%	13.15%
	NPV	\$428.25	\$452.93
	IB	\$5000.00	\$5000.00
	ORR	13.05%	11.92%
	ORR <sub>5000,5</sub>	11.82%	11.92%

\* $k = 10\%$

Where competing projects are assumed to be replaced at the end of their initial project lives, project rankings for projects having different lives cannot simply be based on the NPVs of the respective projects. Instead, revised projects are generated for each competing project by extending their cash flow streams to some common terminal point. NPVs and ORRs calculated for such revised projects should provide consistent project rankings.

**Example 3/4 Project Analyses.** A firm has been approached by two different manufacturers who want to sell it machines that offer labor savings. The firm will buy one machine at most because both machines being considered perform virtually the same services and either machine has sufficient capacity to handle the firm's entire needs (i.e., mutually exclusive projects). The firm's

cost accounting department has provided management with projected cash flow streams for each of these machines (see Table 182.4) and has suggested a 15% opportunity cost for these machines. When one considers the IRRs and unadjusted ORRs of machines A and B in Table 182.4, it is clear that both are acceptable, since the IRR and the ORR of each exceeds the firm's 15% opportunity cost. It may appear that machine B should be favored because both its IRR and its unadjusted ORR are higher than those of machine A. Note, however, that machines A and B have different investment bases and different lives. Adjusting machine B's ORR to A's larger scale of \$25 000 and to project A's longer 5-period life lowers it from 17.57 to 16.29%. When adjusted to a common scale and a common life, the ORRs of machines A and B provide the same ranking of those two projects as does NPV.

**Table 182.4** Machines A and B \*

	Machine A		Machine B	
	Cash Flow	Cumulative PV	Cash Flow	Cumulative PV
Time				
0	-25 000	-25 000.00	-12 000	-12 000.00
1	7 000	-18 913.04	-4 000	-15 478.26
2	7 000	-13 620.04	10 000	-7 916.82
3	9 000	-7 702.39	9 000	-1 999.18
4	10 000	-1 984.86	6 000	1 431.34
5	7 000	1 495.38		
Method				
IRR	17.41%		19.10%	
NPV	\$1 495.38		\$1 431.34	
IB	\$25 000.00		\$15 478.26	
ORR	16.34%		17.57%	
ORR <sub>25 000,5</sub>	16.34%		16.29%	

\* $k = 15\%$

## 182.7 Conclusion

Rate-of-return criteria can provide correct accept or reject decisions for individual projects and can correctly rank competing projects if used properly. The concept of a rate of return seems easier to understand than NPV. After all, most investments (stocks, bonds, CDs, etc.) are ranked according to their rates of return. The NPV criterion, however, is the "gold standard" and its characteristics of preserving project scale and measuring cumulative performance make it more convenient to use when ranking projects.

Rate-of-return criteria should be viewed as complementing the NPV by providing project performance information in a slightly different form. The IRR provides the average rate earned per period on funds invested in the project itself while they remain so invested. In contrast, the ORR provides the average rate earned per period on the project's investment base over the entire life of the project, assuming that funds earn the firm's opportunity cost when not invested in the project

itself.

## Defining Terms

**Internal rate of return:** The average rate earned per period on funds invested in the project itself.

**Investment base:** The time 0 value of the funds that must be provided to (invested in) the project during its life.

**Net present value:** The expected increase in a firm's value if and when it implements the project.

**Opportunity cost:** The rate of return the firm believes it can earn on projects of similar risk. Also referred to as the *minimum required return*, the *discount rate*, or the project's *hurdle rate*.

**Overall rate of return:** The average rate earned per period on the project's investment base over the life of the project.

## References

Bailey, M. J. 1959. Formal criteria for investment decisions. *J. Polit. Econ.* 67(6):476–488.

Beaves, R. G. 1993. The case for a generalized net present value formula. *Eng. Econ.* 38(2):119–133.

Bernhard, R. H. 1989. Base selection for modified rates of return and its irrelevance for optimal project choice. *Eng. Econ.* 35(1):55–65.

Hirshleifer, J. 1958. On the theory of optimal investment decision. *J. Polit. Econ.* 66(4):329–352.

Lin, S. A. Y. 1976. The modified rate of return and investment criterion. *Eng. Econ.* 21(4):237–247.

Mao, J. T. 1966. The internal rate of return as a ranking criterion. *Eng. Econ.* 11(1):1–13.

Shull, D. M. 1992. Efficient capital project selection through a yield-based capital budgeting technique. *Eng. Econ.* 38(1):1–18.

Shull, D. M. 1993. Interpreting rates of return: A modified rate-of-return approach. *Financial Pract. Educ.* 3(2): 67–71.

Solomon, E. 1956. The arithmetic of capital budgeting decisions. *J. Bus.* 29(12):124–129.

## Further Information

Au, T. and Au, T. P. 1992. *Engineering Economics for Capital Investment Analysis*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.

Hendrickson, C., McNeil, S. "Project Selection from Alternatives"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Project Selection from Alternatives

---

183.1 Problem Statement for Project Selection

183.2 Steps in Carrying Out Project Selection

183.3 Selection Criteria

Net Present Value • Other Methods

183.4 Applications

183.5 Conclusion

**Chris Hendrickson**

*Carnegie Mellon University*

**Sue McNeil**

*Carnegie Mellon University*

Practical engineering and management requires choices among competing alternatives. Which boiler should be used in a plant? Which computer should be purchased for a design office? Which financing scheme would be most desirable for a new facility? These are practical questions that arise in the ordinary course of engineering design, organizational management, and even personal finances. This chapter is intended to present methods for choosing the best among distinct alternatives.

### 183.1 Problem Statement for Project Selection

---

The economic project selection problem is to identify the best from a set of possible alternatives. Selection is made on the basis of a systematic analysis of expected revenues and costs over time for each project alternative.

Project selection falls into three general classes of problems. Accept-reject problems (also known as a *determination of feasibility*) require an assessment of whether or not an investment is worthwhile. For example, the hiring of an additional engineer in a design office is an accept-reject decision. Selection of the best project from a set of mutually exclusive projects is required when there are several competing projects or options and only one project can be built or purchased. For example, a town building a new sewage treatment plant may consider three different configurations, but only one configuration will be built. Finally, capital budgeting problems are concerned with the selection of a set of projects when there is a budget constraint and many, not necessarily competing, options. For example, a state highway agency will consider many different



highway rehabilitation projects for a particular year, but generally the budget is insufficient to allow all to be undertaken, although they may all be feasible.

## 183.2 Steps in Carrying Out Project Selection

---

A systematic approach for economic evaluation of projects includes the following major steps [Hendrickson, 1989]:

1. Generate a set of project or purchase **alternatives** for consideration. Each alternative represents a distinct component or combination of components constituting a purchase or project decision. We shall denote project alternatives by the subscript  $x$ , where  $x = 1, 2, \dots$  refers to projects 1, 2, and so on.
2. Establish a **planning horizon** for economic analysis. The planning horizon is the set of future periods used in the economic analysis. It could be very short or long. The planning horizon may be set by organizational policy (e.g., 5 years for new computers or 50 years for new buildings), by the expected economic life of the alternatives, or by the period over which reasonable forecasts of operating conditions may be made. The planning horizon is divided into discrete periods—usually years, but sometimes shorter units. We shall denote the planning horizon as a set of  $t = 0, 1, 2, 3, \dots, n$ , where  $t$  indicates different periods, with  $t = 0$  being the present,  $t = 1$  the first period, and  $t = n$  representing the end of the planning horizon.
3. Estimate the **cash flow profile** for each alternative. The cash flow profile should include the revenues and costs for the alternative being considered during each period in the planning horizon. For public projects, revenues may be replaced by estimates of benefits for the public as a whole. In some cases revenues may be assumed to be constant for all alternatives, so only costs in each period are estimated. Cash flow profiles should be specific to each alternative, so the costs avoided by not selecting one alternative (say,  $x = 5$ ) are not included in the cash flow profile of the alternatives ( $x = 1, 2$ , and so on). Revenues for an alternative  $x$  in period  $t$  are denoted  $B(t, x)$ , and costs are denoted  $C(t, x)$ . Revenues and costs should initially be in **base-year** or constant dollars. Base-year dollars do not change with inflation or deflation.

For tax-exempt organizations and government agencies, there is no need to speculate on inflation if the cash flows are expressed in terms of base-year dollars and a MARR without an inflation component is used in computing the net present value. For private corporations that pay taxes on the basis of then-current dollars, some modification should be made to reflect the projected inflation rates when considering depreciation and corporate taxes.

4. Specify the **minimum attractive rate of return (MARR)** for discounting. Revenues and costs incurred at various times in the future are generally not valued equally to revenues and costs occurring in the present. After all, money received in the present can be invested to obtain interest income over time. The MARR represents the trade-off between monetary amounts in different periods and does not include inflation. The MARR is usually expressed as a percentage change per year, so that the MARR for many public projects may be stated as

10%. The value of MARR is usually set for an entire organization based upon the opportunity cost of investing funds internally rather than externally in the financial markets. For public projects the value of MARR is a political decision, so MARR is often called the *social rate of discount* in such cases. The equivalent value of a dollar in a following period is calculated as  $(1 + \text{MARR})$ , and the equivalent value two periods in the future is  $(1 + \text{MARR})^2$ . In general, if you have  $Y$  dollars in the present [denoted  $Y(0)$ ], then the future value in time  $t$  [denoted  $Y(t)$ ] is

$$Y(t) = Y(0)(1 + \text{MARR})^t \quad (183.1)$$

or the present value,  $Y(0)$ , of a future dollar amount  $Y(t)$  is

$$Y(0) = Y(t)/(1 + \text{MARR})^t \quad (183.2)$$

5. Establish the criterion for accepting or rejecting an alternative and for selecting the best among a group of mutually exclusive alternatives. The most widely used and simplest criterion is the **net present value** criterion. Projects with a positive net present value are acceptable. Only one from a group of mutually exclusive alternatives can be chosen. For example, the alternatives might be alternative boilers for a building or alternative airport configurations. From a set of mutually exclusive alternatives, the alternative with the highest net present value is best. The next section details the calculation steps for the net present value and also some other criterion for selection.
6. Perform sensitivity and uncertainty analysis. Calculation of net present values assumes that cash flow profiles and the value of MARR are reasonably accurate. In many cases assumptions are made in developing cash flow profile forecasts. Sensitivity analysis can be performed by testing a variety of such assumptions, such as different values of MARR, to see how alternative selection might change. Formally treating cash flow profiles and MARR values as stochastic variables can be done with probabilistic and statistical methods.

## 183.3 Selection Criteria

---

### Net Present Value

Calculation of net present values to select projects is commonly performed on electronic calculators, on commercial spreadsheet software, or by hand. The easiest calculation approach is to compute the net revenue in each period for each alternative, denoted  $A(t, x)$ :

$$A(t, x) = B(t, x) - C(t, x) \quad (183.3)$$

where  $A(t, x)$  may be positive or negative in any period. Then, the net present value of the alternative,  $\text{NPV}(x)$ , is calculated as the sum over the entire planning horizon of the discounted

values of  $A(t, x)$ :

$$\text{NPV}(x) = \sum_{t=0}^n A(t, x)/(1 + \text{MARR})^t \quad (183.4)$$

## Other Methods

Several other criteria may be used to select projects. Other discounted flow methods include **net future value** [denoted  $\text{NFV}(x)$ ] and **equivalent uniform annual value** [denoted  $\text{EUAV}(x)$ ]. It can be shown [Au, 1992] that these criteria are equivalent where

$$\text{NFV}(x) = \text{NPV}(x)(1 + \text{MARR})^n \quad (183.5)$$

$$\text{EUAV}(x) = \frac{\text{NPV}(x)(1 + \text{MARR})^n}{[(1 + \text{MARR})^n - 1]} \quad (183.6)$$

The net future value is the equivalent value of the project at the end of the planning horizon. The equivalent uniform annual value is the equivalent series in each year of the planning horizon.

Alternatively, benefit-to-cost ratio (the ratio of the discounted benefits to discounted costs) and the internal rate of return [the equivalent MARR at which  $\text{NPV}(x) = 0$ ] are merit measures, each of which may be used to formulate a decision. For accept-reject decisions, the benefit-to-cost ratio must be greater than one and the internal rate of return greater than the MARR. However, these measures must be used in connection with incremental analyses of alternatives to provide consistent results for selecting among mutually exclusive alternatives [see, for instance, Au (1992)].

Similarly, the payback period provides an indication of the time it takes to recoup an investment but does not indicate the best project in terms of expected net revenues.

## 183.4 Applications

To illustrate the application of these techniques and the calculations involved, two examples are presented.

**Example 183.14 Alternative Bridge Designs.** A state highway agency is planning to build a new bridge and is considering two distinct configurations. The initial costs and annual costs and benefits for each bridge are shown in the following table. The bridges are each expected to last 30 years.

	Alternative 1	Alternative 2
Initial cost	\$15 000 000	\$25 000 000
Annual maintenance and operating costs	\$15 000	\$10 000
Annual benefits	\$1 200 000	\$1 900 000
Annual benefits less costs	\$1 185 000	\$1 890 000

**Solution.** The net present values for a MARR of 5% are given as follows:

$$\begin{aligned}\text{NPV}(1) &= (-15\,000\,000) + (1\,185\,000)/(1 + 0.05) + (1\,185\,000)/(1 + 0.05)^2 \\ &\quad + (1\,185\,000)/(1 + 0.05)^3 + \cdots + (1\,185\,000)/(1 + 0.05)^{30} \\ &= \$3\,216\,354\end{aligned}$$

$$\begin{aligned}\text{NPV}(2) &= (-15\,000\,000) + (1\,890\,000)/(1 + 0.05) + (1\,890\,000)/(1 + 0.05)^2 \\ &\quad + (1\,890\,000)/(1 + 0.05)^3 + \cdots + (1\,890\,000)/(1 + 0.05)^{30} \\ &= \$4\,053\,932\end{aligned}$$

Therefore, the department of transportation should select the second alternative, which has the largest net present value. Both alternatives are acceptable since their net present values are positive, but the second alternative has a higher net benefit.

**Example 183.2% Equipment Purchase.** Consider two alternative methods for sealing pavement cracks [McNeil, 1992]. The first method is a manual method; the second is an automated method using a specialized equipment system. Which method should be used? We shall solve this problem by analyzing whether the new automated method has revenues and benefits in excess of the existing manual method.

**Solution.** Following the steps outlined earlier, the problem is solved as follows:

1. The alternatives for consideration are (1) the existing manual method, and (2) the automated method. The alternatives are mutually exclusive because cracks can only be sealed using either the existing method or the new method.
2. The planning horizon is assumed to be 6 years to coincide with the expected life of the automated equipment.
3. The cash flow profile for alternative 2 is given in the following table:

System acquisition costs	\$100 000
Annual maintenance and operating costs	\$10 000
Annual labor savings	\$36 000
Annual savings over costs	\$26 000

The values are estimated using engineering judgment and historical cost experience. We assume that the productivity and revenues for both alternatives are the same and treat labor savings as additional benefits for alternative 2. Therefore, only the net present value for alternative 2, which represents the result of introducing the automated method, need be computed.

4. The MARR is assumed to be 5%. The net present value is computed as follows:

$$\begin{aligned}
\text{NPV}(2) &= 100\,000 + (26\,000)/(1 + 0.05) \\
&\quad + (26\,000)/(1 + 0.05)^2 + \cdots + (26\,000)/(1 + 0.05)^5 \quad (183.7) \\
&= \$12\,566
\end{aligned}$$

5. Using the criterion  $\text{NPV}(2) > 0$ , alternative 2 is selected.

6. To determine the sensitivity of the result to some of the assumptions, consider [Table 183.1](#). The table indicates that additional investment in the automated method is justifiable at the MARR of 5% if the acquisition costs decrease or the labor savings increase. However, if the MARR increases to 10% or the acquisition costs increase, then the investment becomes uneconomical.

**Table 183.1** Energy Price Escalation Rates

Acquisition Cost (\$)	Labor Saving (\$)	Maintenance and Operation (\$)	MARR		
			0.05	0.01	0.15
50 000	36 000	10 000	\$62 566	\$48 560	\$37 156
100 000	36 000	10 000	\$12 566	(\$1 440)	(\$12 844)
150 000	36 000	10 000	(\$37 434)	(\$51 440)	(\$62 844)
50 000	45 000	10 000	\$101 532	\$82 678	\$67 325
100 000	45 000	10 000	\$51 532	\$32 678	\$17 325
150 000	45 000	10 000	\$1 532	(\$17 322)	(\$32 675)

This example illustrates the use of the net present value criteria for an incremental analysis, which assumes that the benefits are constant for both alternatives and examines incremental costs for one project over another.

## 183.5 Conclusion

This chapter has presented the basic steps for assessing economic feasibility and selecting the best project from a set of mutually exclusive projects, with net present value as a criterion for making the selection.

### Defining Terms

**Alternatives:** A distinct option for a purchase or project decision.

**Base year:** The year used as the baseline of price measurement of an investment project.

**Cash flow profile:** Revenues and costs for each period in the planning horizon.

**Equivalent uniform annual value:** Series of cash flows with a discounted value equivalent to the net present value.

**Minimum attractive rate of return (MARR):** Percentage change representing the time value of money.

**Net future value:** Algebraic sum of the computed cash flows at the end of the planning horizon.

**Net present value:** Algebraic sum of the discounted cash flows over the life of an investment project to the present.

**Planning horizon:** Set of time periods from the beginning to the end of the project; used for economic analysis.

## References

Au, T. and Au, T. P. 1992. *Engineering Economics for Capital Investment Analysis*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.

Hendrickson, C. and Au, T. 1989. *Project Management for Construction*. Prentice Hall, Englewood Cliffs, NJ.

McNeil, S. 1992. An analysis of the costs and impacts of the automation of pavement crack sealing. *Proc. World Conf. on Transp. Res.* Lyon, France, July.

Park, C. S. 1993. *Contemporary Engineering Economics*. Addison Wesley, Reading, MA.

## Further Information

A thorough treatment of project selection is found in *Engineering Economics for Capital Investment Analysis*. Many examples are presented in *Contemporary Engineering Economics*.

Au, T. "Depreciation and Corporate Taxes"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Depreciation and Corporate Taxes

---

This article is based on material in chapters 10, 11, 12, and 16 of Au and Au [1992]. The permission of Prentice Hall, Inc. is gratefully acknowledged.

- 184.1 Depreciation as Tax Deduction
- 184.2 Tax Laws and Tax Planning
- 184.3 Decision Criteria for Project Selection
- 184.4 Inflation Consideration
- 184.5 After-Tax Cash Flows
- 184.6 Evaluation of After-Tax Cash Flows
- 184.7 Effects of Various Factors

**Tung Au**

*Carnegie Mellon University*

The government levies taxes on corporations and individuals to meet its cost of operations. Regulations on depreciation allowances are part of the taxation policy. The tax laws promulgated by the federal government profoundly influence capital investments undertaken by private corporations. Economic valuation of the after-tax cash flows of an investment project based on projected tax rates and inflation effects provides a rational basis for accepting or rejecting investment projects.

### 184.1 Depreciation as Tax Deduction

---

**Depreciation** refers to the decline in value of physical assets over their estimated useful lives. In the context of tax liability, **depreciation allowance** refers to the amount allowed as a deduction in computing taxable income, and **depreciable life** refers to the estimated useful life over which depreciation allowances are computed. Historically, an asset could not be depreciated below a reasonable salvage value. Thus, depreciation allowance is a systematic allocation of the cost of a physical asset between the time it is acquired and the time it is disposed of.

The methods of computing depreciation and the estimated useful lives for various classes of physical assets are specified by government regulations as a part of the tax code, which is subject to periodic revisions. Different methods of computing depreciation lead to different annual depreciation allowances and hence have different effects on taxable income and the taxes paid.

Let  $P$  be the historical cost of an asset,  $S$  its estimated salvage value, and  $N$  the depreciable life in



years. Let  $D_t$  denote the depreciable allowance in year  $t$ , and  $T_t$  denote the accumulated depreciation up to and including year  $t$ . Then for  $t = 1, 2, \dots, N$ ,

$$T_t = D_1 + D_2 + \dots + D_t \quad (184.1)$$

An asset's book value  $B_t$  is simply its historical cost less any accumulated depreciation. Then

$$B_t = P - T_t \quad (184.2)$$

or

$$B_t = B_{t-1} - D_t \quad (184.3)$$

Among the depreciation methods acceptable under the tax regulations, the straight-line method is the simplest. Using this method, the uniform annual allowance in each year is

$$D_t = (P - S)/N \quad (184.4)$$

Other acceptable methods, known as *accelerated depreciation methods*, yield higher depreciation allowances in the earlier years of an asset and less in the later years than those obtained by the straight-line method. Examples of such methods are sum-of-the-years'-digits depreciation and double-declining-balance depreciation, which are treated extensively elsewhere [Au and Au, 1992].

Under the current IRS regulations on depreciation, known as the Modified Accelerated Cost Reduction System, the estimated useful life of an asset is determined by its characteristics that fit one of the eight arbitrarily specified categories. Furthermore, the salvage value  $S$  for all categories is assumed to be zero, whereas all equipment with life of ten years or less is assumed to be purchased at mid-year.

## 184.2 Tax Laws and Tax Planning

---

Capital projects are long-lived physical assets for which the promulgation and revisions of tax laws may affect the tax liability. For the purpose of planning and evaluating capital projects, it is important to understand the underlying principles, including the adjustments for the transition period after each revision and for multiyear "carry-back" or "carry-forward" of profits and losses.

The federal income tax is important to business operations because profits are taxed annually at substantial rates on a graduated basis. Except for small businesses, the corporate taxes on ordinary income may be estimated with sufficient accuracy by using the marginal tax rate. **Capital gain**, which represents the difference between the sale price and the book value of an asset, is taxed at a rate lower than on ordinary income if it is held longer than a period specified by tax laws.

Some state and/or local governments also levy income taxes on corporations. Generally, such taxes are deductible for federal income tax to avoid double taxation. The computation of income taxes can be simplified by using a combined tax rate to cover the federal, state, and local income taxes.

Tax planning is an important element of private capital investment analysis because the economic feasibility of a project is affected by the taxation of corporate profits. In making estimates of tax liability, several factors deserve attention: (1) number of years for retaining the asset, (2) depreciation method used, (3) method of financing, including purchase versus lease, (4) capital gain upon the sale of the asset, and (5) effects of inflation. Appropriate assumptions should be made to reflect these factors realistically.

### 184.3 Decision Criteria for Project Selection

---

The economic evaluation of an investment project is based on the merit of the **net present value**, which is the algebraic sum of the discounted net cash flows over the life of the project to the present. The discount rate is the minimum attractive rate of return specified by the corporation.

The evaluation of proposed investment projects is based on the net present value criteria, which specify the following: (1) an independent project should be accepted if the NPV is positive and rejected otherwise; and (2) among all acceptable projects that are mutually exclusive, the one with the highest positive NPV should be selected.

A more general treatment of the net present value decision criteria for economic evaluation of investment projects may include effects of different reinvestment assumptions [Beaves, 1993] and different scales of investment [Shull, 1992].

### 184.4 Inflation Consideration

---

The consideration of the effects of inflation on economic evaluation of a capital project is necessary because taxes are based on then-current dollars in future years. The year in which the useful life of a project begins is usually used as the baseline of price measurement and is referred to as the **base year**. A **price index** is the ratio of the price of a predefined package of goods and service at a given year to the price of the same package in the base year. The common price indices used to measure inflation include the consumer price index, published by the Department of Labor, and the gross domestic product price deflator, compiled by the Department of Commerce.

For the purpose of economic evaluation, it is generally sufficient to project the future inflation trend by using an average annual inflation rate  $j$ . Let  $A_t$  be the cash flow in year  $t$ , expressed in terms of base-year (year 0) dollars, and  $A'_t$  be the cash flow in year  $t$ , expressed in terms of then-current dollars. Then

$$A'_t = A_t(1 + j)^t \quad (184.5)$$

$$A_t = A'_t(1 + j)^{-t} \quad (184.6)$$

In the economic evaluation of investment proposals in an inflationary environment, two approaches may be used to offset the effects of inflation. Each approach leads to the same result if the discount rate  $i$ , excluding inflation, and the rate  $i'$ , including inflation, are related as follows:

$$i' = (1 + i)(1 + j) - 1 = i + j + ij \quad (184.7)$$

$$i = (i' - j)/(1 + j) \quad (184.8)$$

The net present value (NPV) of an investment project over a planning horizon of  $n$  years can be obtained by using the constant price approach as follows:

$$NPV = \sum_{t=0}^n A_t(1 + i)^{-t} \quad (184.9)$$

Similarly, the NPV obtained by using the then-current price approach is

$$NPV = \sum_{t=0}^n A'_t(1 + i')^{-t} \quad (184.10)$$

In some situations the prices of certain key items affecting the estimates of future incomes and/or costs are expected to escalate faster than the general inflation. For such cases the differential inflation for those items can be included in the estimation of the cash flows for the project.

## 184.5 After-Tax Cash Flows

---

The economic performance of a corporation over time is measured by the net cash flows after tax. Consequently, after-tax cash flows are needed for economic evaluation of an investment project. Since interests on debts are tax deductible according to the federal tax laws, the method of financing an investment project could affect the net profits. Although the projected net cash flows over the years must be based on then-current dollars for computing taxes, the depreciation allowances over those years are not indexed for inflation under the current tax laws.

It is possible to separate the cash flows of a project into an operating component and a financing component for the purpose of evaluation. Such a separation will provide a better insight to the tax advantage of borrowing to finance a project, and the combined effect of the two is consistent with the computation based on a single combined net cash flow. The following notations are introduced to denote various items in year  $t$  over a planning horizon of  $n$  years:

$A_t$  = net cash flow of operation (excluding financing cost) before tax

$A_t$  = net cash flow of financing before tax

$A_t = A_t + A_t =$  combined net cash flow before tax  
 $Y_t =$  net cash flow of operation (excluding financing cost) after tax  
 $Y_t =$  net cash flow of financing after tax  
 $Y_t = Y_t + Y_t =$  combined net cash flow after tax  
 $D_t =$  annual depreciation allowance  
 $I_t =$  annual interest on the unpaid balance of a loan  
 $Q_t =$  annual payment to reduce the unpaid balance of a loan  
 $W_t =$  annual taxable income  
 $X_t =$  annual marginal income tax rate  
 $K_t =$  annual income tax

Thus, for operation in year  $t = 0, 1, 2, \dots, n$ ,

$$W_t = A_t - D_t \quad (184.11)$$

$$K_t = X_t W_t \quad (184.12)$$

$$Y_t = A_t - X_t(A_t - D_t) \quad (184.13)$$

For financing in year  $t = 0, 1, 2, \dots, n$ ,

$$I_t = Q_t - A_t \quad (184.14)$$

$$Y_t = A_t + X_t I_t \quad (184.15)$$

where the term  $X_t I_t$  is referred to as the **tax shield** because it represents a gain from debt financing due to the deductibility of interests in computing the income tax.

Alternately, the combined net cash flows after tax may be obtained directly by noting that both depreciation allowance and interest are tax deductible. Then,

$$W_t = A_t - D_t - I_t \quad (184.16)$$

$$Y_t = A_t - X_t(A_t - D_t - I_t) \quad (184.17)$$

It can be verified that Eq. (184.17) can also be obtained by adding Eqs. (184.13) and (184.15), while noting  $A_t = A_t + A_t$  and  $Y_t = Y_t + Y_t$ .

## 184.6 Evaluation of After-Tax Cash Flows

---

For private corporations the decision to invest in a capital project may have side effects on the financial decisions of the firm, such as taking out loans or issuing new stocks. These financial decisions will influence the overall equity-debt mix of the entire corporation, depending on the size of the project and the risk involved.

Traditionally, many firms have used an adjusted cost of capital, which reflects the opportunity

cost of capital and the financing side effects, including tax shields. Thus only the net cash from operation  $Y_t$  obtained by Eq. (184.13) is used when the net present value is computed. The after-tax net cash flows of a proposed project are discounted by substituting  $Y_t$  for  $A_t$  in Eq. (184.9), using after-tax adjusted cost of capital of the corporation as the discount rate. If inflation is anticipated,  $Y'_t$  can first be obtained in then-current dollars and then substituted into Eq. (184.10). The selection of the project will be based on the NPV thus obtained without further consideration of tax shields, even if debt financing is involved. This approach, which is based on the adjusted cost of capital for discounting, is adequate for small projects such as equipment purchase.

In recent years another approach, which separates the investment and financial decisions of a firm, is sometimes used for evaluation of large capital projects. In this approach the net cash flows of operation are discounted at a risk-adjusted rate reflecting the risk for the class of assets representing the proposed project, whereas tax shields and other financial side effects are discounted at a risk-free rate corresponding to the yield of government bonds. An adjusted net present value reflecting the combined effects of both decisions is then used as the basis for project selection. The detailed discussion of this approach may be found elsewhere [Brealey and Myers, 1988].

## 184.7 Effects of Various Factors

---

Various depreciation methods will produce different effects on the after-tax cash flows of an investment. Since the accelerated depreciation methods generate larger depreciation allowances during the early years, the net present value of the after-tax cash flows using one of the accelerated depreciation methods is expected to be more favorable than that obtained by using the straight-line method.

If a firm lacks the necessary funds to acquire a physical asset that is deemed desirable for operation, it can lease the asset by entering into a contract with another party, which will legally obligate the firm to make payments for a well-defined period of time. The payments for leasing are expenses that can be deducted in full from the gross revenue in computing taxable income. The purchase-or-lease options can be compared after their respective net present values are computed.

When an asset is held for more than a required holding period under tax laws, the capital gain is regarded as long-term capital gain and is taxed at a lower rate. In a period of inflation the sale price of an asset in then-current dollars increases, but the book value is not allowed to be indexed to reflect the inflation. Consequently, capital gain tax increases with the surge in sale price resulting from inflation.

**Example 184.1.** A heavy-duty truck is purchased at \$25 000 in February. This truck is expected to generate a before-tax uniform annual revenue of \$7000 over the next six years, with a salvage value of \$3000 at the end of six years. According to the current IRS regulations, this truck is assigned an estimated useful life of five years with no salvage value. The straight-line depreciation method is used to compute the annual depreciation allowance. The combined federal and state income tax rate is 38%. Assuming no inflation, the after-tax discount rate of 8%, based on the

adjusted cost of capital of the corporation, is used. Determine whether this investment proposal should be accepted.

**Solution.** Using Eq. (184.4), the annual depreciation allowance  $D_t$  is found to be 5000 per year, since a useful life of five years is specified and the salvage value is zero, according to the current IRS regulations. Following the mid-year purchase assumption, the actual depreciation allowances for years one through six are as shown in the following table. The actual revenues over six years are used in the analysis.

$t$	$A_t$	$D_t$	$A_t - D_t$	$K_t$	$Y_t$
0	-25 000	—	—	—	-25 000
1	7 000	2500	2500	1710	5 290
2-5	7 000	5000	2000	760	6 240
6	7 000	2500	4500	1710	5 290

Using the adjusted cost of capital approach, the net present value of the after-tax net cash flows discounted at 8% is obtained by substituting  $Y_t$  for  $A_t$  in Eq. (184.9),

$$\begin{aligned} \text{NPV} = & -25\,000 + (6240)(P | U, 8\%, 5) - (6240 - 5290)(P | F, 8\%, 1) \\ & + (5290)(P | F, 8\%, 5) = 2369 \end{aligned}$$

in which  $(P | U, 8\%, 5)$  is the discount factor to present at 8% for a uniform series over 5 years, and  $(P | F, 8\%, 1)$  and  $(P | F, 8\%, 5)$  are discount factors to present at 8% for a future sum at the end of 1 year and 5 years, respectively. Since  $\text{NPV} = 2369$  is positive, the proposed investment should be accepted.

**Example 184.2.** Consider a proposal for the purchase of a computer workstation that costs \$20 000 and has no salvage value at disposal after four years. This investment is expected to generate a before-tax uniform annual revenue of \$7000 in base-year dollars over the next four years. An average annual inflation rate of 5% is assumed. The straight-line depreciation method is used to compute the annual depreciation allowance. The combined federal and state income tax rate is 38%. Based on the adjusted cost of capital of the corporation, the after-tax discount rate, including inflation, is 10%. Determine whether this investment proposal should be accepted.

**Solution.** To simplify the calculation, the assumption of mid-year purchase is ignored. Using Eq. (184.4), the annual depreciation allowance  $D_t$  is found to be \$5000 for  $t = 1$  to 4. This annual depreciation allowance of \$5000 will not be indexed for inflation, according to the IRS regulations.

The annual before-tax revenue of \$7000 in base-year dollars must be expressed in then-current dollars before computing the income taxes. From Eq. (184.5),

$$A'_t = (7000)(1 + 0.05)^t$$

where  $t = 1$  to 4 refers to each of the next four years. The after-tax cash flow  $Y'_t$  for each year can be computed by Eq. (184.13). The step-by-step tabulation of the computation for each year is shown in the following table.

$t$	$A_t$	$A'_t$	$D_t$	$A'_t - D_t$	$K_t$	$Y'_t$
0	-20 000	-20 000	—	—	—	-20 000
1	7 000	7 350	5999	2359	893	6 457
2	7 000	7 718	5000	2718	1033	6 685
3	7 000	8 013	5000	3103	1179	6 924
4	7 000	8 509	5000	3509	1333	7 176

Using the adjusted cost of capital approach, the net present value of the after-tax cash flows discounted at  $i' = 10\%$ , including inflation, can be obtained by substituting the value of  $Y'_t$  for  $A'_t$  in Eq. (184.10) as follows:

$$\begin{aligned} \text{NPV} &= -20\,000 + (6457)(1.1)^{-1} + (6685)(1.1)^{-2} + (6924)(1.1)^{-3} + (7176)(1.1)^{-4} \\ &= 1498 \end{aligned}$$

Since  $\text{NPV} = \$1498$  is positive, the investment proposal should be accepted.

**Example 184.3.** A developer bought a plot of land for \$100 000 and spent \$1.6 million to construct an apartment building on the site for a total price of \$1.7 million. The before-tax annual rental income after the deduction of maintenance expenses is expected to be \$300 000 in the next six years, assuming no inflation. The developer plans to sell this building at the end of six years when the property is expected to appreciate to \$2.1 million, including land. The entire cost of construction can be depreciated over 32 years based on the straight-line depreciation method, whereas the original cost of land may be treated as the salvage value at the end. The tax rates are 34% for ordinary income and 28% for capital gain, respectively. Based on the adjusted cost of capital, the developer specifies an after-tax discount rate of 10%.

**Solution.** Using Eq. (184.4), the annual depreciation allowance  $D_t$  over 32 years is found to be 50 000. Noting that  $P = 1\,700\,000$  and  $T_t = (6)(50\,000) = 300\,000$ , the book value of the property after six years is found from Eq. (184.2) to be \$1.4 million.

Ignoring the assumption of mid-year purchase to simplify the calculation and assuming no inflation, the after-tax annual net income in the next six years is given by Eq. (184.13):

$$Y_t = 300\,000 - (34\%)(300\,000 - 50\,000) = 215\,000$$

The capital gain tax for the property at the end of 6 years is

$$(28\%)(2\,100\,000 - 1\,400\,000) = 196\,000$$

Using the adjusted cost of capital approach, the net present value of after-tax net cash flows in

the next six years, including the capital gain tax paid at the end of six years discounted at 10%, is

$$\begin{aligned}\text{NPV} &= -1\,700\,000 + (215\,000)(P | U, 10\%, 6) \\ &\quad + (2\,100\,000 - 196\,000)(P | F, 10\%, 6) \\ &= 311\,198\end{aligned}$$

in which  $(P | U, 10\%, 6)$  is the discount factor to present at 10% for a uniform series over six years, and  $(P | F, 10\%, 6)$  is the discount factor to present at 10% for a future sum at the end of six years. Since  $\text{NPV} = \$311\,198$  is positive, the proposed investment should be accepted.

## Defining Terms

**Base year:** The year used as the baseline of price measurement of an investment project.

**Capital gain:** Difference between the sale price and the book value of an asset.

**Depreciable life:** Estimated useful life over which depreciation allowances are computed.

**Depreciation:** Decline in value of physical assets over their estimated useful lives.

**Depreciation allowance:** Amount of depreciation allowed in a systematic allocation of the cost of a physical asset between the time it is acquired and the time it is disposed of.

**Net present value:** Algebraic sum of the discounted cash flows over the life of an investment project to the present.

**Price index:** Ratio of the price of a predefined package of goods and service at a given year to the price of the same package in the base year.

**Tax shield:** Gain from debt financing due to deductibility of interests in computing the income tax.

## References

- Au, T. and Au, T. P. 1992. *Engineering Economics for Capital Investment Analysis*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.
- Beaves, R. G. 1993. The case for a generalized net present value formula. *Eng. Economist*. 38(2):119–133.
- Brealey, R. and Myers, S. 1988. *Principles of Corporate Finance*, 3rd ed. McGraw-Hill, New York.
- Shull, D. N. 1992. Efficient capital project selection through a yield-based capital budgeting technique. *Eng. Economist*. 38(1):1–18.



## **Further Information**

Up-to-date information of federal tax code may be found in the following annual publications:

*Federal Tax Guide*, Prentice Hall, Englewood Cliffs, NJ.

*Standard Federal Tax Reporter*, Commerce Clearing House, Chicago, IL.

Fazzi, C. "Financing and Leasing"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Financing and Leasing

---

### 185.1 Debt Financing

Term Loans • Mortgage Loans • Bonds

### 185.2 Equity Financing

Preferred Stock • Common Stock

### 185.3 Leasing

Operating Lease • Financial Lease • Sale and Leaseback • Accounting Treatment of Leases • Tax Treatment of Leases • Evaluating Lease Financing

### Charles Fazzi

*Robert Morris College*

In other chapters a firm's investment decision has been discussed in terms of capital budgeting. This chapter focuses on the other major decision that a firm must face: where to obtain the funds to invest in these capital projects. The decision of where to raise the money is called the financing decision. The market for funds is divided into the money market and the capital market. The money market includes short-term debt securities, that is, securities that will mature in one year or less. The capital market is the market for longer-term borrowings, that is, sources of cash with a time horizon of more than one year. The capital market is further subdivided into an intermediate capital market, which includes debt securities with a maturity of more than one but less than ten years, and a long-term capital market, which includes debt securities which generally have a maturity of more than ten years. Financing activities for capital projects rarely utilize money market funds.

The capital market also encompasses the market for equity securities. These sources of funds include the issuance of both preferred stock and common stock. They have the longest time horizon since these securities are normally issued for the life of the corporation. Thus they may be thought of as sources of financing with an infinite time horizon.

This article will deal with the major sources and forms of capital market funds. In addition, leasing will be discussed. Leasing is a specific type of economic transaction that blends the acquisition of an asset (an investment decision) with a long-term loan commitment (a financing decision). Under the right conditions, leasing can represent a most attractive approach to the financing process for a firm.

### 185.1 Debt Financing

---

There are three primary sources of intermediate and long-term debt financing: term loans,

mortgage loans, and bonds.

## Term Loans

A **term loan** is simply a loan that is paid off over some number of years called the term of the loan. Term loans may or may not be secured by the assets of the firm. These loans are usually negotiated with commercial banks, insurance companies, or some other financial institution. They can generally be negotiated fairly quickly and at a low administrative cost. The amount borrowed (principal) and the interest are paid off in installments over the life of the loan. The rate of interest on the term loan can be fixed over the life of the loan, but usually it is a variable interest rate that is linked to the prime rate. Thus if the rate of interest for a term loan is set at "2% over the prime rate," the borrower may pay 8% when the prime rate is 6% in the first year, and 9% in the second year when the prime rate is 7%. Variable-rate loans of this type can include a "collar," which sets upper and lower limits on the interest rate that can be charged, or a "cap," which sets an upper limit only. The prime rate of interest is the rate charged on short-term business loans to financially sound companies. Term loans often carry restrictive provisions or constraints on the financial activities of the borrower, such as a dividend restriction.

## Mortgage Loans

Mortgage loans are essentially term loans secured by the real estate that was purchased with the cash proceeds from the loan. They infrequently extend past a time period of 20 years and may be repaid only over 5 years.

## Bonds

**Bonds** are intermediate to long-term debt agreements issued by corporations, governments, or other organizations, generally in units of \$1000 principal value per bond. A bond represents two commitments on the part of the issuing organization: the promise to pay the stated interest rate periodically and the commitment to repay the \$1000 principal when the bond matures. Most bonds pay interest semiannually at one-half the annual stated or coupon rate of interest. The term *coupon rate* refers to the fact that bonds often come with coupons that may be detached and redeemed for each interest payment.

A company that floats a bond issue may have the bonds sold directly to the public by placing the bond issue with an investment banker. A bond issue can also be privately placed with a financial institution such as a commercial bank, insurance company, corporate pension fund, or university endowment fund. The contract between the issuer and the lender is called the bond indenture. If the bond is publicly marketed, a trustee is named to ensure compliance with the bond indenture. In most cases the trustee is a commercial bank or an investment banker. In a private placement, the purchasing institution normally acts as its own trustee.

Bonds are issued in a wide variety of circumstances. A debenture is an unsecured bond that is backed by the full faith and credit of the issuer. No specific assets are pledged as collateral. In the event of default, debenture holders become general creditors of the issuer. A mortgage bond, like a

mortgage loan, is collateralized by a mortgage on some type of asset, such as land or a building.

Convertible bonds are bonds that may be converted into shares of common stock in a pre-determined conversion rate. The decision to convert the bond to common stock is at the option of the bondholder. The conversion feature is attractive to investors and therefore lowers the cost of borrowing.

Most bond issues include a call feature among the provisions. This feature gives the issuer a chance to buy back the bonds at a stated price prior to maturity. Callable bonds give the issuer flexibility in financing its activities. The call feature favors the borrower and therefore increases the effective cost of borrowing.

A sinking fund requirement establishes a procedure for the orderly retirement of a bond issue over its life. This provision requires the periodic repurchase of a stated percentage of the outstanding bonds. The issuing company can accomplish this through purchases of their bonds in the open market or through the use of the call provision previously described.

## 185.2 Equity Financing

---

Equity financing utilizes a corporation's ability to sell ownership interest in the organization through the sale of capital stock in the corporation. This stock can be issued either as preferred stock or as common stock. While a corporation can choose to issue preferred stock, it must issue common stock as a representation of the individual ownership interest of the investors.

### Preferred Stock

**Preferred stock** is legally an equity security that represents an ownership interest in a corporation. The term *preferred* arises from the fact that preferred stock has two important preferences over common stock: preference as to payment of dividends and preference as to stockholders' claims on the assets of the business in the event of bankruptcy.

Preferred stock is often characterized as a hybrid security that possesses some of the features of bonds and some of the features of common stock. Preferred dividends are fixed in amount similar to bond interest, and must be paid before common dividends are paid. Like bondholders, preferred stockholders do not participate in the growth of corporate earnings and collect only the dividends that are stated on the stock certificate. Like common stockholders, each dividend payment must be voted on and approved by the board of directors of the corporation. Most preferred stock issuances are cumulative, meaning that missed dividends accumulate as "dividends in arrears," which must be paid before any common dividends can be paid. Timely payment of preferred dividends is very important to the corporation. Corporations with preferred dividends in arrears find it very difficult to raise other forms of capital. Therefore, payment of preferred dividends is almost as important as timely payment of bond interest to the corporation. Preferred stock can also be convertible and callable, as discussed earlier in conjunction with bonds.

### Common Stock

The common stockholders are the owners of the corporation, as evidenced by the voting rights that

are normally associated with **common stock**. Each share of common stock normally has one vote in electing the corporation's board of directors. The board is the ultimate corporate authority and is the group to which the president of the corporation reports. The board selects the president and approves the appointment of all the corporate officers. The board members are responsible to the stockholders, and the stockholders have the authority to elect a new board if their performance is deemed unsatisfactory.

Common stock is often referred to as the residual equity of the corporation. Money paid to the corporation for common stock does not have to be repaid, and dividends on common stock are paid only when approved by the board of directors. Common dividends are never in arrears, and common stockholders never have a claim to any specified dividend level. If a corporation prospers and grows, the board will normally vote to increase the dividend level to reflect this growth. As dividends grow, the value of the stock increases, and the stockholders also benefit from capital gains which accrue from the increased value. Of course, all of this can also work in reverse if the corporation does not prosper. The shareholders could ultimately lose their investment if the corporation goes bankrupt.

Corporations sometimes issue two classes of common stock, one of which does not include any voting rights. Thus, voting rights are retained by one class of common stock, and the other class will not have any voting rights but will have the ability to participate in earnings growth. This will allow one group of owners to retain voting control while allowing another group to participate in earnings and dividends but with no control. This is a way to raise additional equity capital without losing control of the corporation. Large publicly held corporations rarely have multiple classes of common stock.

## 185.3 Leasing

---

A **lease** can be defined as a contractual relationship in which the owner of the property or asset (lessor) conveys to a business or person (lessee) the right to use the property or asset for a specified period of time in exchange for a series of payments. Thus in a lease contract the lessee is able to use the leased assets without assuming ownership. Leasing has become a very popular way for many businesses to acquire the necessary resources to run their operations. The leasing event is a hybrid transaction which combines the acquisition of necessary resources with a commitment to finance the acquisition. In general, two types of leases are offered in the market today: operating leases and financial or capital leases.

### Operating Lease

An **operating lease** is a short-term lease written usually for a period of time that is substantially shorter than the asset's useful life. The lessor assumes most of the risks of ownership, including maintenance, service, insurance, liability, and property taxes. The lessee can cancel an operating lease on short notice. Thus, the operating lease does not involve the long-term fixed future commitment of financial resources and is similar to renting. There are no balance sheet effects recorded with an operating lease, and the rental payments are treated as period expenses.

## Financial Lease

A **financial lease**, or capital lease, is a contract by which the lessee agrees to pay the lessor a series of payments whose sum equals or exceeds the purchase price of the asset. Typically, the total cash flows from the lease payments, the tax savings, and the residual value of the asset at the end of the lease will be sufficient to pay back the lessor's investment and provide a profit. Most financial leases are "net" leases, in that the fundamental ownership responsibilities such as maintenance, insurance, and property and sales taxes are placed upon the lessee. The lease agreement is a long-term agreement between both parties and is not cancelable. In the case of an unforeseen event, the contract may be cancelable, but the lessor will typically impose a substantial prepayment penalty. The decision to use financial leasing as a means of acquiring an asset is often viewed as an alternative to a purchase transaction using long-term debt financing to generate the necessary cash.

## Sale and Leaseback

A sale and leaseback is a fairly common arrangement where a firm sells an asset to a lender/lessor and then immediately leases back the property. The advantage to the seller/lessee is that the selling firm receives a large infusion of cash that may be used to finance other business activities. In return for this cash, the selling firm issues a long-term lease obligation to the lessor in order to make economic use of the asset during the lease period. Title to the asset is transferred to the lessor, and the lessor will realize any residual value the asset may have at the end of the lease. The lessee may realize a tax advantage in this situation if the lease involves a building on owned land. Land is not depreciable if owned outright, but the full amount of the lease payment may be tax-deductible. This allows the lessee to indirectly depreciate the cost of the land through the deductibility of the lease payment.

## Accounting Treatment of Leases

Accounting for leases has changed dramatically over time. Prior to 1977, lease financing was attractive to some firms because the lease obligation did not appear on the company's balance sheet. As a result, leasing was regarded as a form of "off-balance sheet financing." In 1977 the Financial Accounting Standards Board issued an accounting standard that clearly set out criteria distinguishing a capital lease from an operating lease. Any lease that does not meet the criteria for a capital lease must be classified as an operating lease. A lease is considered to be a capital lease if it meets any one of the following conditions:

1. The title of the asset being leased is transferred to the lessee at the end of the lease term.
2. The lease contains an option for the lessee to purchase the asset at a very low price.
3. The term of the lease is greater than or equal to 75% of the economic life of the asset.
4. The present value of the lease payments is greater than or equal to 90% of the fair value of the leased property.

A capital lease is recorded on the lessee's balance sheet as a lease asset with an associated lease liability. The amount of the asset and liability is equal to the present value of the minimum future lease payments. Thus, for capital leases, leasing no longer provides a source of off-balance sheet financing. Operating leases must be fully disclosed in the footnotes to the financial statements.

## Tax Treatment of Leases

For tax purposes, the lessee can deduct the full amount of the lease payment in a properly structured lease. The Internal Revenue Service (IRS) wants to be sure that the lease contract truly represents a lease and not an installment purchase of the asset. To assure itself that a true lease is present, the IRS will look for a meaningful residual value at the end of the lease term. This is usually construed to mean that the term of the lease cannot exceed 90% of the economic life of the asset. In addition, the lessee should not be given the option to purchase the asset for anything less than the fair market value of the leased asset at the end of the lease term. The lease payments should be reasonable in that they provide the lessor not only a return of principal, but a reasonable interest return as well. In addition, the lease term must be less than 30 years; otherwise it will be construed as an installment purchase of the asset. The IRS wants to assure itself that the transaction is not a disguised installment purchase that has more rapid payments which will lead to greater deductions than would be allowed from the depreciation deduction under an outright purchase. With leasing, the cost of any land is amortized in the lease payments. In an outright purchase of land, depreciation is not allowable. When the value of a lease includes land, lease financing can offer a tax advantage to the firm. Offsetting this tax advantage is the likely residual value of the land at the end of the lease period.

## Evaluating Lease Financing

To evaluate whether or not a lease financing proposal makes economic sense, it should be compared with financing the asset with debt. Whether leasing or borrowing is best will depend on the patterns of cash flow for each financing method and on the opportunity cost of funds. The method of analysis will require the calculation of the present values for the lease alternative and the present value of the borrowing alternative. The alternative with the lower present value of the cash outflows will be the better choice. The calculations can be rather extensive and complicated. Also, it should be recognized that the calculations do deal with uncertainty, which must be considered in making the final decision.

## Defining Terms

**Bond:** A long-term debt instrument issued by a corporation or government.

**Common stock:** The stock representing the most basic rights to ownership in the corporation.

**Financial lease:** A long-term lease that is not cancelable.

**Lease:** A contract under which one party, the owner of an asset (lessor), agrees to grant the use of the asset to another party (lessee), in exchange for periodic payments.

**Operating lease:** A short-term lease that is often cancelable.



**Preferred stock:** A type of stock that usually promises a fixed dividend, with the approval of the board of directors. It has preference over common stock in the payment of dividends and claims on assets.

**Term loan:** Debt originally scheduled for repayment in more than one year but generally less than ten years.

## References

- Levy, H. and Sarnat, M. 1990. *Capital Investment & Financial Decisions*, 4th ed. Prentice Hall International (UK) Ltd., London.
- Schall, L. D. and Haley, C. W. 1991. Leasing. In *Introduction to Financial Management*, 6th ed. p. 753–778. McGraw-Hill, Inc., New York.
- Seitz, N. E. 1990. *Capital Budgeting and Long-term Financing Decisions*. Dryden Press, Chicago.
- Van Horne, J. C. and Wachowicz, J. M. 1992. *Fundamentals of Financial Management*, 8th ed. Prentice Hall, Inc., Englewood Cliffs, NJ.

## Further Information

- The Institute of Management Accountants offers a variety of publications and educational programs on finance and accounting topics. They can be reached at 10 Paragon Drive, Montvale, NJ 07645, telephone 1-800-638-4427.
- Clay, R. H. and Holder, W. W. 1977. A practitioners' guide to accounting for leases. *J. Accountancy*. August, p. 61–68.
- Financial Accounting Standards Board. 1976. Statement of Financial Accounting Standards No. 13: Accounting for leases. FASB, Stamford, CT.

Ruegg, R. T. "Risk Assessment"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

Contribution of the National Institute of Standards and Technology. Not subject to copyright.

### 186.1 Expected Value (EV) Analysis

How to Calculate EV • Advantages and Disadvantages of the EV Technique • Expected Value and Risk Attitude

### 186.2 Mean-Variance Criterion (MVC) and Coefficient of Variation (CV)

How to Use the MVC • How to Calculate the CV • Advantages and Disadvantages of the MVC and CV Techniques

### 186.3 Risk-Adjusted Discount Rate (RADR) Technique

How to Calculate the RADR • Advantages and Disadvantages of the RADR Technique

### 186.4 Certainty Equivalent (CE) Technique

Establishing CEFs and Applying the CEF Technique • Advantages and Disadvantages of the CE Technique

### 186.5 Simulation Technique

How to Perform Simulation • Advantages and Disadvantages of Simulation

### 186.6 Decision Analysis

How to Perform Decision Analysis • Advantages and Disadvantages of Decision Analysis

## **Rosalie T. Ruegg**

*National Institute of Standards and Technology*

**Risk assessment** provides decision makers with information about the **risk exposure** inherent in a given decision, that is, the probability that the outcome will be different from the "best-guess" estimate. Risk assessment also is concerned with the **risk attitude** of the decision maker—that is, the willingness of the decision maker to take a chance on an investment of uncertain outcome. Risk assessment techniques are typically used in conjunction with conventional methods of investment analysis, such as **net present value** (NPV), benefit-to-cost ratio (BCR), and rate-of-return (ROR) analysis; the risk assessment techniques are generally not used as "stand-alone" evaluation techniques. Techniques range from simple and partial to complex and comprehensive. Although none of these techniques take the risk out of making decisions, if used correctly, they can help the decision maker to make more informed choices in the face of uncertainty.

This chapter provides an overview of the following risk assessment techniques:

- Expected value analysis
- Mean-variance criterion and coefficient of variation
- Risk-adjusted discount rate technique
- Certainty equivalent technique

- Simulation analysis
- Decision analysis

Other techniques are also used to assess risks—for example, the mathematical/analytical technique and various combinations of techniques—but those covered here are the most widely used.

## 186.1 Expected Value (EV) Analysis

Expected value analysis provides a simple way of taking into account **uncertainty** in input values, but it does not provide an explicit measure of risk in the outcome. It is helpful in explaining and illustrating risk attitudes.

### How to Calculate EV

An expected value is the sum of the products of the dollar value of alternative outcomes and their probabilities of occurrence. That is, where  $a_{1-n}$  indicates values associated with alternative outcomes of a decision and  $p_{1-n}$  indicates the probabilities of occurrence of the alternative outcomes, the expected value (EV) of the decision is calculated as follows:

$$EV = a_1p_1 + a_2p_2 + \cdots a_np_n \quad (186.1)$$

**Example 186.1 - EV Analysis.** The following simplified example illustrates the combining of expected value analysis and net present value analysis to support a purchase decision.

Assume that the postal service must decide whether to buy a given piece of labor-saving equipment. Assume the unit purchase price of the equipment is \$100 000, the yearly operating cost is \$5000 (obtained by a fixed-price contract), and both costs are known with certainty. The annual labor cost savings, on the other hand, are uncertain but can be estimated in probabilistic terms, as shown in [Table 186.1](#) in the columns headed  $a_1$ ,  $p_1$ ,  $a_2$ , and  $p_2$ . The expected value calculations and the net present value calculations are also given in [Table 186.1](#).

**Table 186.1** Expected Value and Net Present Value Calculations

Year	Equipment Purchase (\$1000)	Operating Cost (\$1000)	$a_t$ (\$1000)	Labor Savings			PV Factor*	PV (\$1000)
				$p_1$	$a_2$ (\$1000)	$p_2$		
0	−100	—	—	—	—	—	1.00	−100
1		−5	25	0.8	50	0.2	0.93	23
2		−5	30	0.8	60	0.2	0.86	27
3		−5	30	0.7	60	0.3	0.79	27
4		−5	30	0.6	60	0.4	0.74	27
5		−5	30	0.8	60	0.2	0.68	21
Expected Net Present Value								25

**Present value** (PV) calculations are based on a discount rate of 8%. Probabilities sum to 1.0 in a given year.

If the equipment decision were based only on net present value calculated with the best-guess labor savings (column  $a_1$ ), the equipment purchase would be found to be uneconomic. But if the possibility of greater labor savings is taken into account by using the expected value of labor savings rather than the best guess, the conclusion is that over repeated applications the equipment should be cost-effective. The expected net present value of the labor-saving equipment is \$25 000 per unit.

## Advantages and Disadvantages of the EV Technique

An advantage of the technique is that it provides a predicted value that tends to be closer to the actual value than a simple best-guess estimate over repeated instances of the same event—provided, of course, that the input probabilities can be predicted with some accuracy.

A disadvantage of the EV technique is that it expresses the outcome as a single-value measure, such that there is no explicit measure of risk. Another is that the estimated outcome is predicated on many replications of the event—with the expected value in effect a weighted average of the outcome over many like events. But in the case of a single instance of an event, the expected value is unlikely to occur. This is analogous to the case of a single coin toss: The outcome will be either heads or tails, not the probabilistic-based weighted average of both.

## Expected Value and Risk Attitude

Expected values are useful in explaining risk attitude. Risk attitude may be thought of as a decision maker's preference of taking a chance on an uncertain money payout of known probability versus accepting a sure money amount. Suppose, for example, a person were given a choice between accepting the outcome of a fair coin toss, where heads means winning \$10 000 and tails means losing \$5000, and accepting a certain cash amount of \$2000. Expected value analysis can be used to evaluate and compare the choices. In this case the expected value is \$2500, which is \$500 more than the certain money amount. The "risk-neutral" decision maker will prefer the coin toss because of its higher expected value. The decision maker who prefers the \$2000 certain amount is demonstrating a "risk-averse" attitude. On the other hand, if the certain amount were raised to \$3000 and the first decision maker still preferred the coin toss, he or she would be demonstrating a "risk-taking" attitude. Such trade-offs can be used to derive a *utility function* that represents a decision maker's risk attitude.

The risk attitude of a given decision maker typically is a function of the amount at risk. Many people who are risk-averse when faced with the possibility of significant loss become risk-neutral, or even risk taking, when potential losses are small. Since decision makers vary substantially in their risk attitudes, there is a need to assess not only risk exposure—that is, the degree of risk inherent in the decision—but also the risk attitude of the decision maker.

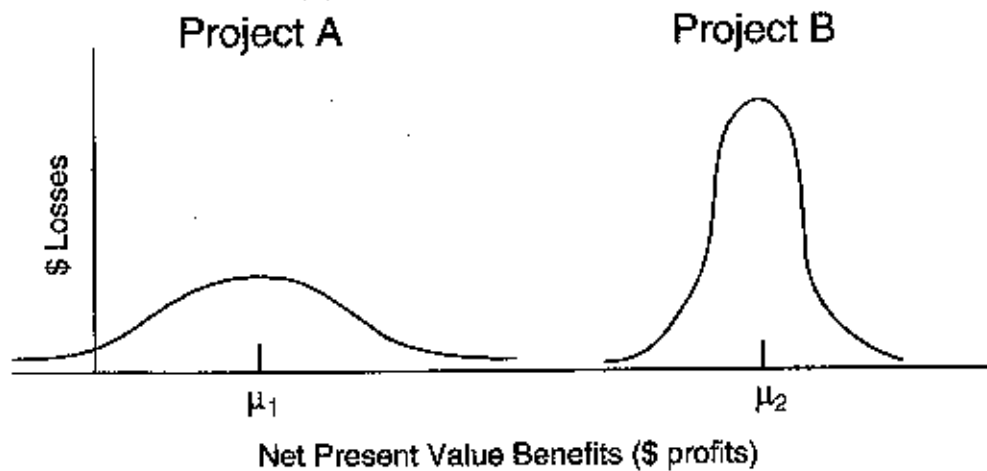
## 186.2 Mean-Variance Criterion (MVC) and Coefficient of Variation (CV)

The MVC and CV can be useful in choosing among risky alternatives when the mean outcomes and standard deviations (variation from the mean) can be calculated.

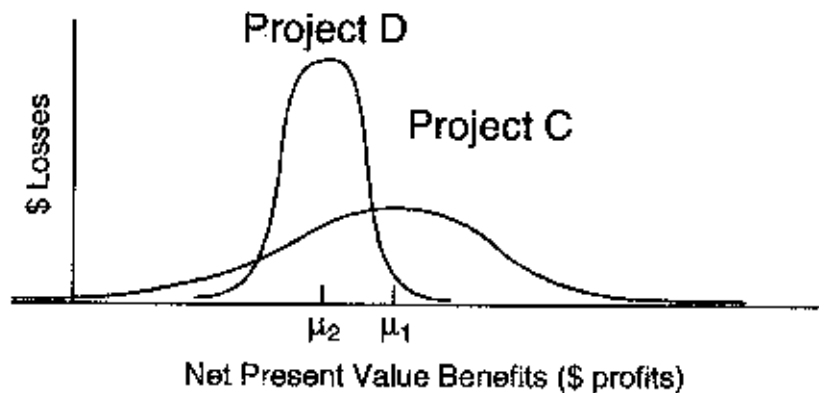
### How to Use the MVC

Consider a choice between two projects—one with higher mean net benefits and a lower standard deviation than the other. This situation is illustrated in Fig. 186.1. In this case the project whose probability distribution is labeled B can be said to have stochastic dominance over the project labeled A. Project B is preferable to project A both on grounds that its output is likely to be higher and on grounds that it entails less risk of loss. But what if the alternative with the higher mean output entails higher risk of an unfavorable outcome, as illustrated in Fig. 186.2? If this were the case, the MVC would provide inconclusive results.

**Figure 186.1** Stochastic dominance as demonstrated by mean-variance criterion.



**Figure 186.2** Inconclusive results from mean-variance criterion.



## How to Calculate the CV

It is helpful to compute the coefficient of variation (CV) to determine the relative risk of two alternative projects. The CV indicates which alternative has the lower risk per unit of project output. Risk-averse decision makers will prefer the alternative with the lower CV, other things being equal. The CV is calculated as follows:

$$CV = \sigma/\mu \quad (186.2)$$

where

CV = coefficient of variation

$\sigma$  = standard deviation

$\mu$  = mean

## Advantages and Disadvantages of the MVC and CV Techniques

The principal advantage of these techniques is that they provide quick, easy-to-calculate indications of the returns and risk exposure of one project relative to another. The principal disadvantage is that the MVC does not provide a clear indication of preference when the alternative with the higher mean output has the higher risk, or vice versa.

## 186.3 Risk-Adjusted Discount Rate (RADR) Technique

---

The RADR technique takes account of risk through the **discount rate**. If a project's benefit stream is riskier than that of the average project in the decision maker's portfolio, a higher than normal discount rate is used; if the benefit stream is less risky, a lower than normal discount rate is used. If costs are the source of the higher-than-average uncertainty, a lower than normal discount rate is used, and vice versa. The greater the variability in benefits or costs, the greater the adjustment in the discount rate.

## How to Calculate the RADR

The RADR is calculated as follows:

$$RADR = RFR + NRA + XRA \quad (186.3)$$

where

RADR = risk-adjusted discount rate

RFR = risk-free discount rate, generally set equal to the U.S. Treasury bill rate

NRA = "normal" risk adjustment to account for the average level of risk encountered in the decision maker's operations

XRA = extra risk adjustment to account for risk greater or less than normal risk

**Example 186.2<sup>3</sup> RADR Technique.** A company is considering an investment with high payoff potential and high risk. The normal discount rate used by the company, computed as the weighted average cost of capital, is 12%. The treasury bill rate, taken as the risk-free rate, is 8%. The normal risk adjustment factor is 5%. This investment is twice as risky as the company's average investment, such that the risk adjustment factor is 10%. Hence, the RADR is 18%. The projected cost and revenue streams and the discounted present values are shown in [Table 186.2](#).

**Table 186.2** RADR Example

Year	Costs (\$M)	Revenue (\$M)	PV Costs (\$M)	PV Revenue (\$M)	NPV (\$M)
0	80		80	—	−80
1	5	20	4	17	13
2	5	20	4	14	10
3	5	20	4	12	8
4	5	20	3	10	7
5	5	20	3	9	6
6	5	20	3	7	4
7	5	20	2	6	4
Total NPV					−28

Note: Costs are discounted with a discount rate of 12%, revenue with a discount rate of 18%.

## Advantages and Disadvantages of the RADR Technique

One advantage of RADR is that it provides a way to account for both risk exposure and risk attitude. Furthermore, RADR does not require any additional calculation steps once the value(s) of the RADR is established. One disadvantage is that it provides only an approximate adjustment. The value of the RADR is typically a rough estimate based on sorting investments into risk categories and adding a "fudge factor" to account for the decision maker's risk attitude. RADR is not a fine-tuned measure of the inherent risk associated with variations in cash flows. Further, it typically is biased toward investments with short payoffs, because it applies a constant RADR over the entire time period even though risk may vary over time.

## 186.4 Certainty Equivalent (CE) Technique

The CE technique adjusts investment cash flows by a factor that will convert the measure of



economic worth to a "certainty equivalent" amount—the amount a decision maker will find equally acceptable to a given investment with an uncertain outcome. Central to the technique is the derivation of the certainty equivalent factor (CEF), which is used to adjust net cash flows for uncertainty.

## Establishing CEFs and Applying the CEF Technique

Risk exposure can be built into the CEF by establishing categories of risky investments for the decision maker's organization and linking the CEF to the coefficient of variation of the returns—greater variation translating into smaller CEF values. The procedure is as follows:

1. Divide the organization's portfolio of projects into risk categories. Examples of investment risk categories for a given company might be the following: low-risk investments—expansion of current product lines and equipment replacement; moderate-risk investments—development of "spin-off" product lines and introduction of new equipment; and high-risk investments—new product development and new business acquisition.
2. Estimate the coefficients of variation (see section 186.2) for each investment-risk category—for example, on the basis of historical risk-return data.
3. Assign CEFs by year according to the coefficients of variation, with the highest risk projects being given the lowest CEFs. If the objective is to reflect only risk exposure, set the CEFs such that a risk-neutral decision maker will be indifferent between receiving the estimated certain amount and receiving the uncertain investment. If the objective is to reflect risk attitude as well as risk exposure, set the CEFs such that the risk-averse or risk-taking decision maker will be indifferent.
4. Select the conventional measure of economic worth to be used—such as the measure of net present value (i.e., net benefits).
5. Estimate the net cash flows and decide in which investment-risk category the project in question fits.
6. Multiply the yearly net cash flow amounts by the appropriate CEFs.
7. Discount the adjusted yearly net cash flow amounts with a risk-free discount rate (a risk-free discount rate is used because the risk adjustment is accomplished by the CEFs).
8. Proceed with the remainder of the analysis in the conventional way.

In summary, the certainty equivalent net present value is calculated as follows:

$$NPV_{CE} = \sum_{t=0}^N [(CEF_t(B_t - C_t)) / (1 + RFD)^t] \quad (186.4)$$

where

$NPV_{CE}$  = net present value adjusted for uncertainty by the CE technique

$B_t$  = estimated benefits in time period  $t$

$C_t$  = estimated costs in time period  $t$

RFD = risk-free discount rate

**Example 186.3 - CE Technique.** Table 186.3 illustrates the use of this technique for adjusting net present value calculations for a high-risk investment in a new plant startup. The CEF is set at 0.76 and is assumed to be constant with respect to time.

**Table 186.3** Applying the CE Technique—Investment-Risk Category: High Risk—New Plant Startup

Yearly Net Cash Flow (\$M)		CV	CEF	RFD Discount Factors*	NPV (\$M)
1	−100	0.22	0.76	0.94	−71
2	−100	0.22	0.76	0.89	−68
3	20	0.22	0.76	0.84	13
4	30	0.22	0.76	0.79	18
5	45	0.22	0.76	0.75	26
6	65	0.22	0.76	0.70	35
7	65	0.22	0.76	0.67	33
8	65	0.22	0.76	0.63	31
9	50	0.22	0.76	0.59	22
10	50	0.22	0.76	0.56	21
Total NPV					60

\*The RFD is assumed equal to 6%.

## Advantages and Disadvantages of the CE Technique

A principal advantage of the CE technique is that it can be used to account for both risk exposure and risk attitude. Another is that it separates the adjustment of risk from discounting and makes it possible to make more precise risk adjustments over time. A major disadvantage is that the estimation of CEF is only approximate.

## 186.5 Simulation Technique

Simulation entails the iterative calculation of the measure of economic worth from probability functions of the input variables. The results are expressed as a probability density function and as a cumulative distribution function. The technique thereby enables explicit measures of risk exposure to be calculated. A conventional method is used to calculate economic worth; a computer is employed to sample repeatedly—hundreds of times—and make the calculations.

## How to Perform Simulation

1. Express variable inputs as probability functions. Where there are interdependencies among input values, multiple probability density functions, tied to one another, may be needed.
2. For each input with a probability function, draw randomly an input value; for each input with only a single value, take that value for calculations.
3. Use the input values to calculate the economic measure of worth and record the results.
4. If inputs are interdependent, such that input  $X$  is a function of input  $Y$ , first draw the value of  $Y$ , then draw randomly from the  $X$  values that correspond to the value of  $Y$ .
5. Repeat the process many times until the number of results is sufficient to construct a probability density function and a cumulative distribution function.
6. Construct the probability density function and the cumulative distribution function for the economic **measure of project worth** and perform statistical analysis of the variability.

## Advantages and Disadvantages of Simulation

The strong advantage of simulation is that it expresses the results in probabilistic terms, thereby providing detailed assessment of risk exposure. A disadvantage is that it does not explicitly treat risk attitude; however, by providing a clear measure of risk exposure, it facilitates the implicit incorporation of risk attitude in the decision. The necessity of expressing inputs in probabilistic terms and the extensive calculations are also often considered disadvantages.

## 186.6 Decision Analysis

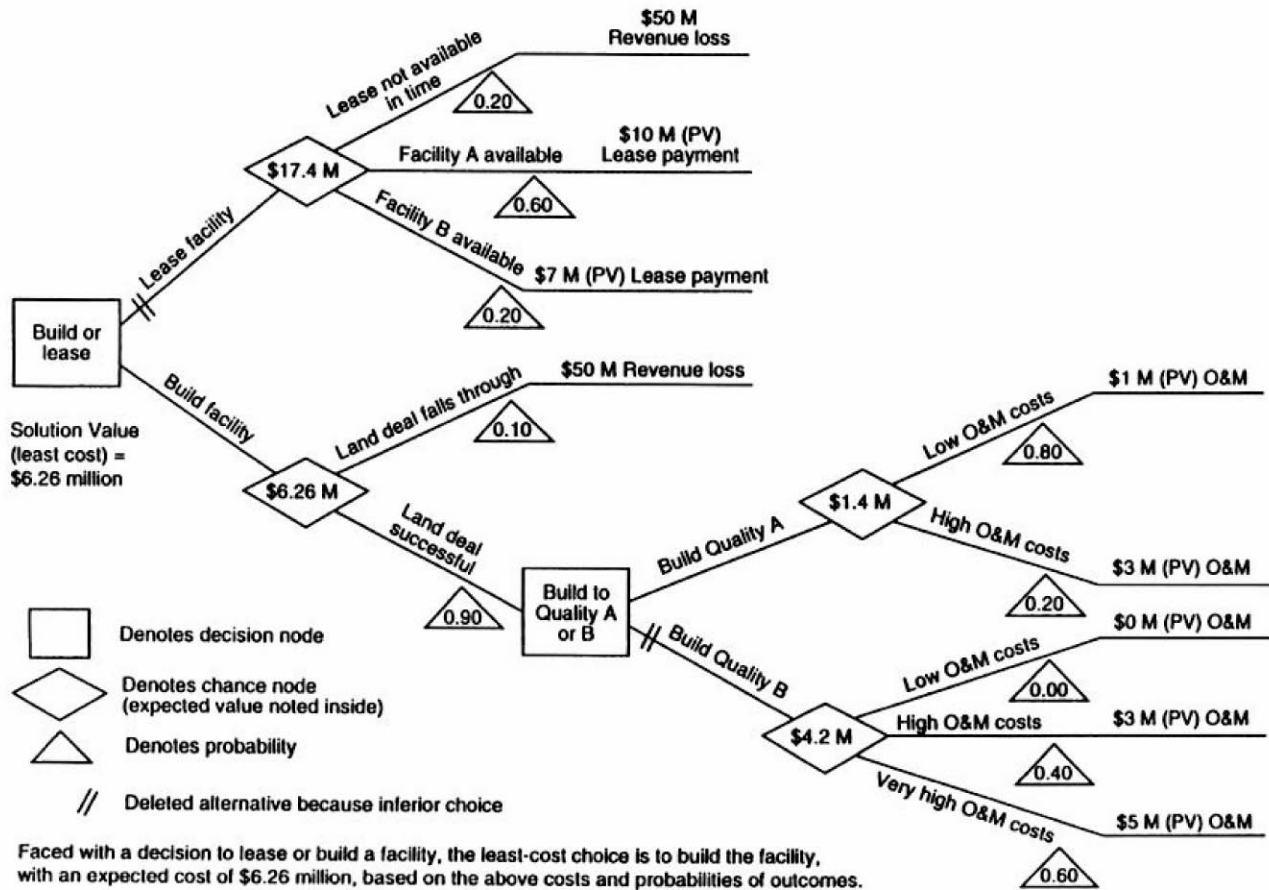
---

Decision analysis is a versatile technique that enables both risk exposure and risk attitude to be taken into account in the economic assessment. It diagrams possible choices, costs, benefits, and probabilities for a given decision problem in *decision trees*, which are useful in understanding the possible choices and outcomes.

## How to Perform Decision Analysis

Although it is not possible to capture the richness of this technique in a brief overview, a simple decision tree (shown in [Fig. 186.3](#)) is presented to give a sense of how the technique is used. The decision problem is whether to lease or build a facility. The decision must be made now, based on uncertain data. The decision tree helps to structure and analyze the problem. The tree is constructed left to right and analyzed right to left. The tree starts with a box representing a decision juncture or node—in this case, whether to lease or build a facility. The line segments branching from the box represent the two alternative paths, the upper one the lease decision and the lower one the build decision.

**Figure 186.3** Decision tree illustration.



## Advantages and Disadvantages of Decision Analysis

An advantage of this technique is that it helps decision makers to understand the problem and to compare alternative solutions. Another advantage is that, in addition to treating risk exposure, it can also accommodate risk attitude by converting benefits and costs to utility values (not addressed in this chapter). A disadvantage is that the technique as typically applied does not provide an explicit measure of the variability of the outcome.

## Defining Terms

**Discount rate:** The minimum acceptable rate of return used in converting benefits and costs occurring at different times to their equivalent values at a common time.

**Measures of project worth:** Economic methods that combine project benefits (savings) and costs in various ways to evaluate the economic value of a project. Examples are life-cycle costs, net benefits or net savings, benefit-to-cost ratio or savings-to-investment ratio, and adjusted internal rate of return.

**Net present value:** Present value benefits minus present value costs.

**Present value:** Time-equivalent value at the present of past, present, and future cash flows.

**Risk assessment:** As applied to economic decisions, the body of theory and practice that helps decision makers assess their risk exposures and risk attitudes in order to increase the probability that they will make economic choices that are best for them.

**Risk attitude:** The willingness of decision makers to take chances on investments of uncertain, but predictable, outcome. Risk attitudes may be classified as risk-averse, risk-neutral, and risk taking.

**Risk exposure:** The probability of investing in a project whose economic outcome is less favorable than what is economically acceptable.

**Uncertainty:** As used in the context of this chapter, a state of not being certain about the values of variable inputs to an economic analysis.

## References

- Greenberg, J. 1982. *Investment Decisions; The Influence of Risk and Other Factors*. American Management Association, New York.
- Park, C. S. 1992. Probabilistic benefit-cost analysis. *Eng. Economist*. 29(2):83–100.
- Ruegg, R. T. and Marshall, H. E. 1990. *Building Economics: Theory and Practice*. Chapman and Hall, New York.

## Further Information

There are a number of computer programs on the market for applying risk analysis techniques. An overview of a selection of software is provided by [Ruegg and Marshall, 1990](#), Appendix D; however, it should be recognized that any assessment of software is quickly dated. For a video tutorial on risk analysis techniques, see the video training tape, *Uncertainty and Risk*, part II in a series on least-cost energy decisions for buildings, National Institute of Standards and Technology, 1992, 5709-B Arundel Avenue, Rockville, MD 20852. Phone: (301)881-0270.

Marshall, H. E. "Sensitivity Analysis"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Sensitivity Analysis

Contribution of the National Institute of Standards and Technology. Not subject to copyright.

### 187.1 Sensitivity Analysis Applications

Sensitivity Table for Programmable Control System • Sensitivity Graph for Gas Heating Systems • Spider Diagram for a Commercial Building Investment

### 187.2 Advantages and Disadvantages

#### Harold E. Marshall

*National Institute of Standards and Technology*

**Sensitivity analysis** measures the impact on project outcomes of changing one or more key input values about which there is uncertainty. For example, a pessimistic, expected, and optimistic value might be chosen for an uncertain variable. Then an analysis could be performed to see how the outcome changes as each of the three chosen values is considered in turn, with other things held the same.

In engineering economics, sensitivity analysis measures the economic impact resulting from alternative values of uncertain variables that affect the economics of the project. When computing **measures of project worth**, for example, sensitivity analysis shows just how sensitive the economic payoff is to uncertain values of a critical input, such as the **discount rate** or project maintenance costs expected to be incurred over the project's **study period**. Sensitivity analysis reveals how profitable or unprofitable the project might be if input values to the analysis turn out to be different from what is assumed in a single-answer approach to measuring project worth.

Sensitivity analysis can also be performed on different combinations of input values. That is, several variables are altered at once and then a measure of worth is computed. For example, one scenario might include a combination of all pessimistic values, another all expected values, and a third all optimistic values. Note, however, that sensitivity analysis can in fact be misleading [Hillier, 1969] if all pessimistic assumptions or all optimistic assumptions are combined in calculating economic measures. Such combinations of inputs would be unlikely in the real world.

Sensitivity analysis can be performed for any measure of worth. And since it is easy to use and understand, it is widely used in the economic evaluation of government and private-sector projects. Office of Management and Budget [1992] Circular A-94 recommends sensitivity analysis to federal agencies as one technique for treating uncertainty in input variables. And the American Society for Testing and Materials (ASTM) [1994], in its *Standard Guide for Selecting Techniques for Treating Uncertainty and Risk in the Economic Evaluation of Buildings and Buildings Systems*, describes sensitivity analysis for use in government and private-sector applications.

## 187.1 Sensitivity Analysis Applications

How to use sensitivity analysis in engineering economics is best illustrated with examples of applications. Three applications are discussed. The first two focus on changes in project worth as a function of the change in one variable only. The third allows for changes in more than one uncertain variable.

The results of sensitivity analysis can be presented in text, tables, or graphs. The following illustration of sensitivity analysis applied to a programmable control system uses text and a simple table. Subsequent illustrations use graphs. The advantage of using a graph comes from being able to show in one picture the outcome possibilities over a range of input variations for one or several input factors.

### Sensitivity Table for Programmable Control System

Consider a decision on whether or not to install a programmable time clock to control heating, ventilating, and air conditioning (HVAC) equipment in a commercial building. The time clock would reduce electricity consumption by turning off that part of the HVAC equipment that is not needed during hours when the building is unoccupied.

Using **net savings** (NS) as the measure of project worth, the time clock is acceptable on economic grounds if its NS is positive—that is, if its **present value** savings exceed present value costs. The control system purchase and maintenance costs are felt to be relatively certain. The savings from energy reductions resulting from the time clock, however, are not certain. They are a function of three factors: the initial price of energy, the rate of change in energy prices over the life cycle of the time clock, and the number of kilowatt hours (kWh) saved. Two of these, the initial price of energy and the number of kWh saved, are relatively certain. But future energy prices are not.

To test the sensitivity of NS to possible energy price changes, three values of energy price change are considered: a low rate of energy price escalation (slowly increasing benefits from energy savings), a moderate rate of escalation (moderately increasing benefits), and a high rate of escalation (rapidly increasing benefits).

Table 187.1 shows three NS estimates that result from repeating the NS computation for each of the three energy price escalation rates.

**Table 187.1** Energy Price Escalation Rates

Energy Price Escalation Rate	Net Savings
Low	\$-15 000
Moderate	20 000
High	50 000

To appreciate the significance of these findings, it is helpful to consider what extra information is gained over the conventional single-answer approach, where, say, a single NS estimate of \$20 000 was computed. Table 187.1 shows that the project could return up to \$50 000 in NS if future



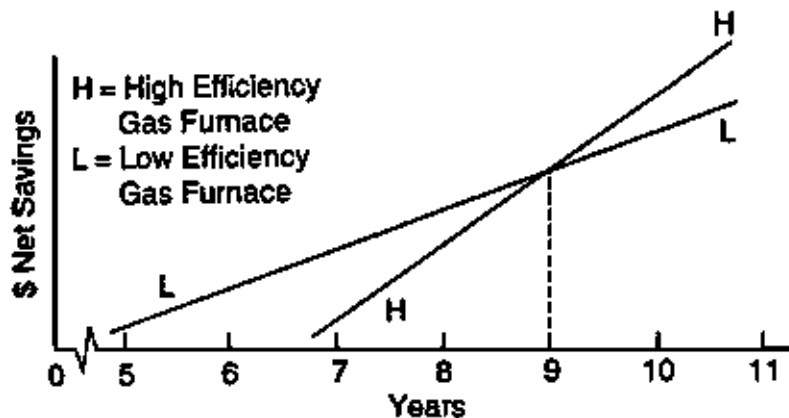
energy prices escalated at a high rate. On the other hand, it is evident that the project could lose as much as \$15 000. This is considerably less than **breakeven**, where the project would at least pay for itself. It is also \$35 000 less than what was calculated with the single-answer approach. Thus, sensitivity analysis reveals that accepting the time clock could lead to an uneconomic outcome.

There is no explicit measure of the likelihood that any one of the NS outcomes will happen. The analysis simply shows what the outcomes will be under alternative conditions. However, if there is reason to expect energy prices to rise, at least at a moderate rate, then the project very likely will make money, other factors remaining the same. This adds helpful information over the traditional single-answer approach to measures of project worth.

## Sensitivity Graph for Gas Heating Systems

Figure 187.1 shows how sensitive NS is to the time over which two competing gas heating systems might be used in a building. The sensitivity graph helps you decide which system to choose on economic grounds.

**Figure 187.1** Sensitivity of net savings to holding period.



Assume that you have an old electric heating system that you are considering replacing with a gas furnace. You have a choice between a high-efficiency or low-efficiency gas furnace. You expect either to last at least 15 to 20 years. And you do not expect any significant difference in building resale value or salvage value from selecting one system over the other. So you compute the NS of each gas furnace as compared to the old electric system. You will not be able to say which system is more economical until you decide how long you will hold the building before selling it. This is where the sensitivity graph is particularly helpful.

Net savings are measured on the vertical axis, and time on the horizontal axis. The longer you hold the building, the greater will be the present value of NS from installing either of the new systems, up to the estimated life of the systems. But note what happens in the ninth year. One line crosses over another. This means that the low-efficiency system is more **cost-effective** than the high-efficiency system for any holding period up to 9 years. To the left of the crossover point, NS

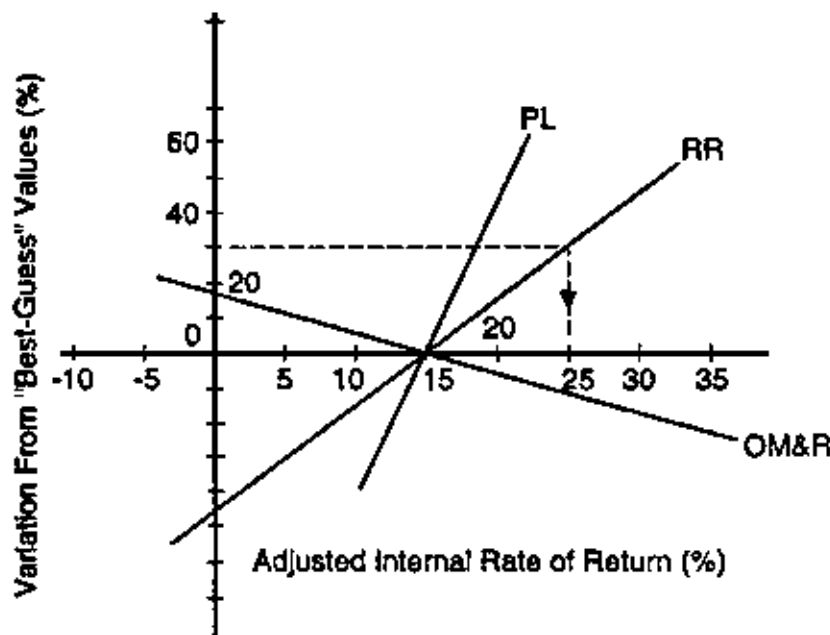
values are higher for the low-efficiency system than for the high-efficiency system. But for longer holding periods, the high-efficiency system is more cost-effective than the low-efficiency system. This is shown to the right of the crossover point.

How does the sensitivity graph help you decide which system to install? First, it shows that neither system is more cost-effective than the other for all holding periods. Second, it shows that the economic choice between systems is sensitive to the uncertainty of how long you hold the building. You would be economically indifferent between the two systems only if you plan to hold the building 9 years. If you plan to hold the building longer than 9 years, for example, then install the high-efficiency unit. But if you plan to hold it less than 9 years, then the low-efficiency unit is the better economic choice.

## Spider Diagram for a Commercial Building Investment

Another useful graph for sensitivity analysis is the **spider diagram**. It presents a snapshot of the potential impact of uncertain input variables on project outcomes. [Figure 187.2](#) shows—for a prospective commercial building investment—the sensitivity of the **adjusted internal rate of return (AIRR)** to three uncertain variables: project life (PL); the reinvestment rate (RR); and operation, maintenance, and replacement costs (OM&R). The spider diagram helps the investor decide if the building is likely to be a profitable investment.

**Figure 187.2** Spider diagram showing sensitivity of the adjusted internal rate of return to variations in uncertain variables. PL = project life; RR = reinvestment rate; and OM&R = operation, maintenance, and replacement costs.



Each of the three uncertain variables is represented by a labeled function that shows what AIRR

value results from various values of the uncertain variable. (Although these functions are not necessarily linear, they are depicted as linear here to simplify exposition.) For example, the downward-sloping OM&R function indicates that the AIRR is inversely proportional to OM&R costs. By design, the OM&R function (as well as the other two functions) passes through the horizontal axis at the "best-guess" estimate of the AIRR (15% in this case), based on the best-guess estimates of the three uncertain variables. Other variables (e.g., occupancy rate) will impact the AIRR, but these are assumed to be known for the purpose of this analysis. Since each of the variables is measured by different units (years, percent, and money), the vertical axis is denominated in positive and negative percent changes from the best-guess values fixed at the horizontal axis. The AIRR value corresponding to any given percent variation indicated by a point on the function is found by extending a line perpendicular to the horizontal axis and directly reading the AIRR value. Thus a 30% increase in the best-guess reinvestment rate would yield a 25% AIRR, assuming that other values remain unchanged. Note that if the measure of AIRR were also given in percent differences, then the best-guess AIRR would be at the origin.

The spider diagram's contribution to decision making is its instant picture of the relative importance of several uncertain variables. In this case, the lesser the slope of a function is, the more sensitive is the AIRR to that variable. For example, any given percent change in OM&R will have a greater impact on the AIRR than will an equal percent change in RR or PL, and a percentage change in RR will have a greater impact than an equal percentage change in PL. Thus an investor will want to know as much as possible about likely OM&R costs for this project, because a relatively small variation in estimated costs could make the project a loser.

## 187.2 Advantages and Disadvantages

---

There are several advantages of using sensitivity analysis in engineering economics. First, it shows how significant any given input variable is in determining a project's economic worth. It does this by displaying the range of possible project outcomes for a range of input values, which shows decision makers the input values that would make the project a loser or a winner. Sensitivity analysis also helps identify critical inputs in order to facilitate choosing where to spend extra resources in data collection and in improving data estimates.

Second, sensitivity analysis is an excellent technique to help in anticipating and preparing for the "what if" questions that are asked in presenting and defending a project. For instance, when one is asked what the outcome will be if operating costs are 50% more expensive than expected, one will be ready with an answer. Generating answers to "what if" questions will help assess how well a proposal will stand up to scrutiny.

Third, sensitivity analysis does not require the use of probabilities, as do many techniques for treating uncertainty.

Fourth, sensitivity analysis can be used on any measure of project worth.

And, finally, sensitivity analysis can be used when there is little information, resources, and time for more sophisticated techniques.

The major disadvantage of sensitivity analysis is that there is no explicit probabilistic measure of **risk exposure**. That is, although one might be sure that one of several outcomes might happen, the analysis contains no explicit measure of their respective likelihoods.

## Defining Terms

**Adjusted internal rate of return (AIRR):** The annual percentage yield from a project over the study period, taking into account the returns from reinvested receipts.

**Breakeven:** A combination of benefits (savings or revenues) that just offset costs, such that a project generates neither profits nor losses.

**Cost-effective:** The condition whereby the present value benefits (savings) of an investment alternative exceed its present value costs.

**Discount rate:** The minimum acceptable rate of return used in converting benefits and costs occurring at different times to their equivalent values at a common time. Discount rates reflect the investor's time value of money (or opportunity cost). "Real" discount rates reflect time value apart from changes in the purchasing power of the dollar (i.e., exclude inflation or deflation) and are used to discount constant dollar cash flows. "Nominal" or "market" discount rates include changes in the purchasing power of the dollar (i.e., include inflation or deflation) and are used to discount current dollar cash flows.

**Measures of project worth:** Economic methods that combine project benefits (savings) and costs in various ways to evaluate the economic value of a project. Examples are life-cycle costs, net benefits or net savings, benefit-to-cost ratio or savings-to-investment ratio, and adjusted internal rate of return.

**Net savings:** The difference between savings and costs, where both are discounted to present or annual values. The net savings method is used to measure project worth.

**Present value:** The time-equivalent value at a specified base time (the present) of past, present, and future cash flows.

**Risk exposure:** The probability that a project's economic outcome is different from what is desired (the target) or what is acceptable.

**Sensitivity analysis:** A technique for measuring the impact on project outcomes of changing one or more key input values about which there is uncertainty.

**Spider diagram:** A graph that compares the potential impact, taking one input at a time, of several uncertain input variables on project outcomes.

**Study period:** The length of time over which an investment is evaluated.

## References

- ASTM. 1994. Standard guide for selecting techniques for treating uncertainty and risk in the economic evaluation of buildings and building systems. E1369-93. *ASTM Standards on Buildings Economics*, 3rd ed. American Society for Testing and Materials. Philadelphia, PA.
- Hillier, F. 1963. The derivation of probabilistic information for the evaluation of risky investments. *Manage. Sci.* p. 444. April.
- Office of Management and Budget. 1992. *Guidelines and Discount Rates for Benefit-Cost Analysis of Federal Programs*, p. 12–13. Circular A-94, 29 October. Washington, DC.

## Further Information

- Uncertainty and Risk*, part II in a series on least-cost energy decisions for buildings. National Institute of Standards and Technology, 1992. VHS tape and companion workbook are available from Video Transfer, Inc., 5709-B Arundel Avenue, Rockville, MD 20852. Phone: (301)881-0270.
- Marshall, H. E. 1988. *Techniques for Treating Uncertainty and Risk in the Economic Evaluation of Building Investments*. Special Publication 757. National Institute of Standards and Technology, Gaithersburg, MD.
- Ruegg, R. T. and Marshall, H. E. 1990. *Building Economics: Theory and Practice*. Chapman and Hall, New York.

Fabrycky, W. J., Blanchard, B. S. "Life-Cycle Costing"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Life-Cycle Costing

---

Material presented in this chapter is adapted from chapter 6 in W. J. Fabrycky and B. S. Blanchard, *Life-Cycle Cost and Economic Analysis*, Prentice Hall, 1991.

### 188.1 The Life-Cycle Costing Situation

### 188.2 Cost Generated over the Life Cycle

Conceptual System Design • Preliminary System Design • Detail Design and Development • Production, Utilization, and Support

### 188.3 The Cost Breakdown Structure

### 188.4 Life-Cycle Cost Analysis

Cost Analysis Goals • Analysis Guidelines and Constraints • Identification of Alternatives • Applying the Cost Breakdown Structure

### 188.5 Cost Treatment over the Life Cycle

### 188.6 Summary

#### **Wolter J. Fabrycky**

*Virginia Polytechnic Institute & State University*

#### **Benjamin S. Blanchard**

*Virginia Polytechnic Institute & State University*

A major portion of the projected **life-cycle cost (LCC)** for a given product, system, or structure is traceable to decisions made during conceptual and preliminary design. These decisions pertain to operational requirements, performance and effectiveness factors, the design configuration, the maintenance concept, production quantity, utilization factors, logistic support, and disposal. Such decisions guide subsequent design and production activities, product distribution functions, and aspects of sustaining system support. Accordingly, if the final life-cycle cost is to be minimized, it is essential that a high degree of cost emphasis be applied during the early stages of system design and development.

### 188.1 The Life-Cycle Costing Situation

---

The combination of rising inflation, cost growth, reduction in purchasing power, budget limitations, increased competition, and so on has created an awareness and interest in the total cost of products, systems, and structures. Not only are the acquisition costs associated with new systems rising, but the costs of operating and maintaining systems already in use are also

increasing rapidly. This is due primarily to a combination of inflation and cost growth factors traceable to the following:

1. Poor quality of products, systems, and structures in use
2. Engineering changes during design and development
3. Changing suppliers in the procurement of system components
4. System production and/or construction changes
5. Changes in logistic support capability
6. Estimating and forecasting inaccuracies
7. Unforeseen events and problems

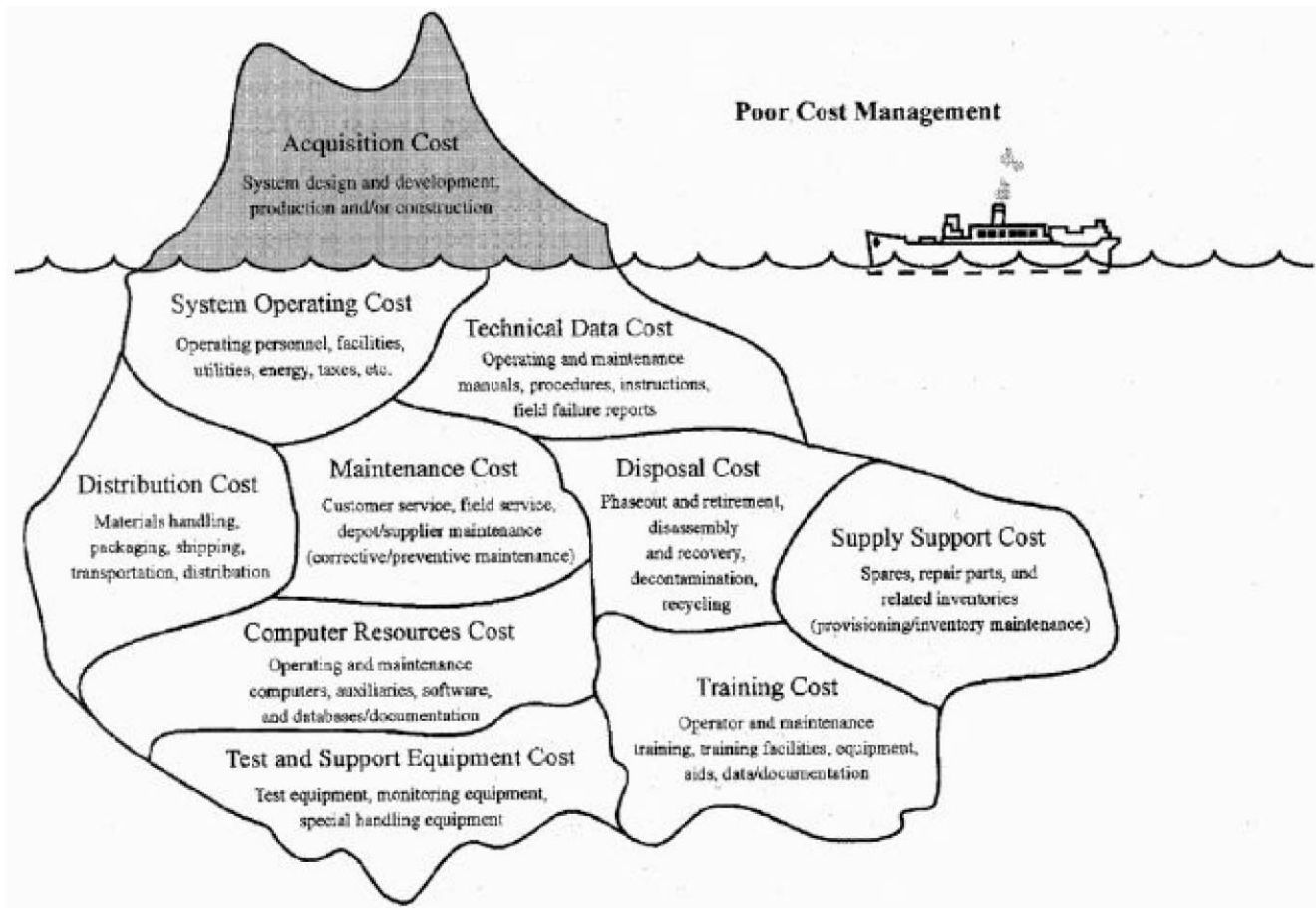
Experience indicates that cost growth due to various causes has ranged from 5 to 10 times the rate of inflation over the past several decades. At the same time, budget allocations for many programs are decreasing from year to year. The result is that fewer resources are available for acquiring and operating new systems or products and for maintaining and supporting existing systems. Available funds for projects, when inflation and cost growth are considered, are decreasing rapidly.

The current economic situation is further complicated by some additional problems related to the actual determination of system and/or product cost. Some of these are:

1. Total system cost is often not visible, particularly those costs associated with system operation and support. The cost visibility problem is due to an "iceberg" effect, as is illustrated in [Fig. 188.1](#).
2. Individual cost factors are often improperly applied. Costs are identified and often included in the wrong category; variable costs are treated as fixed (and vice versa); indirect costs are treated as direct costs; and so on.
3. Existing accounting procedures do not always permit a realistic and timely assessment of total cost. In addition, it is often difficult (if not impossible) to determine costs on a functional basis.
4. Budgeting practices are often inflexible regarding the shift in funds from one category to another, or from year to year, to facilitate cost improvements in system acquisition and utilization.



**Figure 188.1** The problem of total cost visibility.



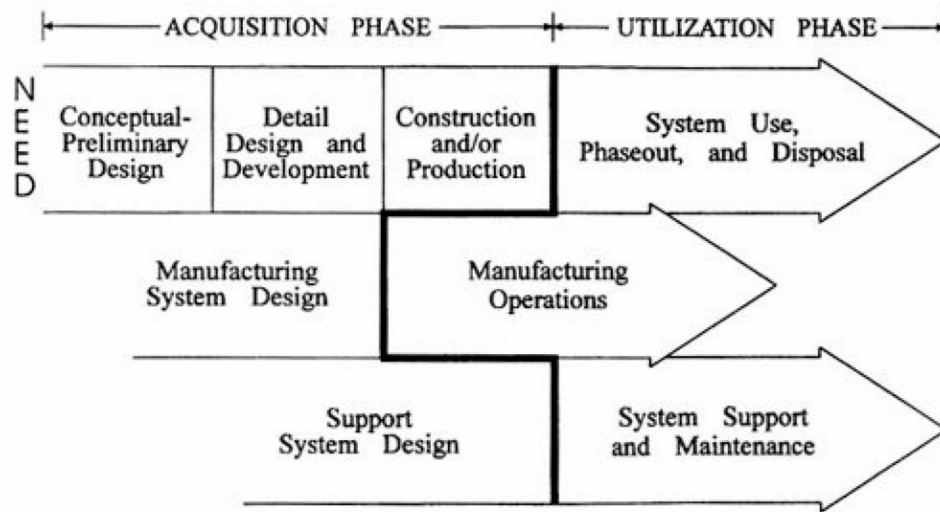
The current trends of inflation and cost growth, combined with these additional problems, have led to inefficiencies in the utilization of valuable resources. Systems and products have been developed that are not cost-effective. It is anticipated that these conditions will become worse unless an increased degree of cost consciousness is assumed by engineers. Economic feasibility studies must address all aspects of life-cycle cost, not just portions thereof.

Life-cycle cost is determined by identifying the applicable functions in each phase of the life cycle, costing these functions, applying the appropriate costs by function on a year-to-year basis, and then accumulating the costs over the entire span of the life cycle. Life-cycle cost must include all producer and consumer costs to be complete.

## 188.2 Cost Generated over the Life Cycle

Life-cycle cost includes all costs associated with the product, system, or structure as applied over the defined life cycle. The life cycle and the major functions associated with each phase are illustrated in [Fig. 188.2](#). Life-cycle costing is employed in the evaluation of alternative system design configurations, alternative production schemes, alternative logistic support policies, alternative disposal concepts, and so on. The life cycle, tailored to the specific system being addressed, forms the basis for life-cycle costing.

**Figure 188.2** Product, process, and support life cycles.



There are many technical and nontechnical decisions and actions required throughout the product or system life cycle. Most actions, particularly those in the earlier phases, have life-cycle implications and greatly affect life-cycle cost. The analysis constitutes a step-by-step approach employing life-cycle cost figures of merit as criteria to arrive at a cost-effective solution. This analysis process is iterative in nature and can be applied to any phase of the life cycle of the product, system, or structure. Cost emphasis throughout the system/product life cycle is summarized in the following sections.

## Conceptual System Design

In the early stages of system planning and conceptual design, when requirements are being defined, quantitative cost figures of merit should be established to which the system or product is to be designed, tested, produced (or constructed), and supported. A **design-to-cost (DTC)** goal may be adopted to establish cost as a system or product design constraint, along with performance, effectiveness, capacity, accuracy, size, weight, reliability, maintainability, supportability, and so on. Cost must be an active rather than a resultant factor throughout the system design process.

## Preliminary System Design

With quantitative cost requirements established, the next step includes an iterative process of synthesis, trade-off and optimization, and system/product definition. The criteria defined in the conceptual system design are initially allocated, or apportioned, to various segments of the system to establish guidelines for the design and/or the procurement of needed element(s). Allocation is accomplished from the system level down to the level necessary to provide an input to design and also to ensure adequate control. The factors projected reflect the target cost per individual unit (i.e., a single equipment unit or product in a deployed population) and are based on system operational requirements, the system maintenance concept, and the disposal concept.

As system development evolves, various approaches are considered that may lead to a preferred configuration. Life-cycle cost analyses are accomplished in (1) evaluating each possible candidate, with the objective of ensuring that the candidate selected is compatible with the established cost targets, and (2) determining which of the various candidates being considered is preferred from an overall cost-effectiveness standpoint. Numerous trade-off studies are accomplished, using life-cycle cost analysis as an evaluation tool, until a preferred design configuration is chosen. Areas of compliance are justified, and noncompliant approaches are discarded. This is an iterative process with an active-feedback and corrective-action loop.

## Detail Design and Development

As the system or product design is further refined and design data become available, the life-cycle cost analysis process involves the evaluation of specific design characteristics (as reflected by design documentation and engineering or prototype models), the prediction of cost-generating sources, the estimation of costs, and the projection of life-cycle cost as a **life-cycle cost profile (LCCP)**. The results are compared with the initial requirement and corrective action is taken as necessary. Again, this is an iterative process, but at a lower level than what is accomplished during preliminary system design.

## Production, Utilization, and Support

Cost concerns in the production, utilization, support, and disposal stages of the system or product life cycle are addressed through data collection, analysis, and an assessment function. High-cost contributors are identified, cause-and-effect relationships are defined, and valuable information is gained and utilized for the purposes of product improvement through redesign or reengineering.

## 188.3 The Cost Breakdown Structure

---

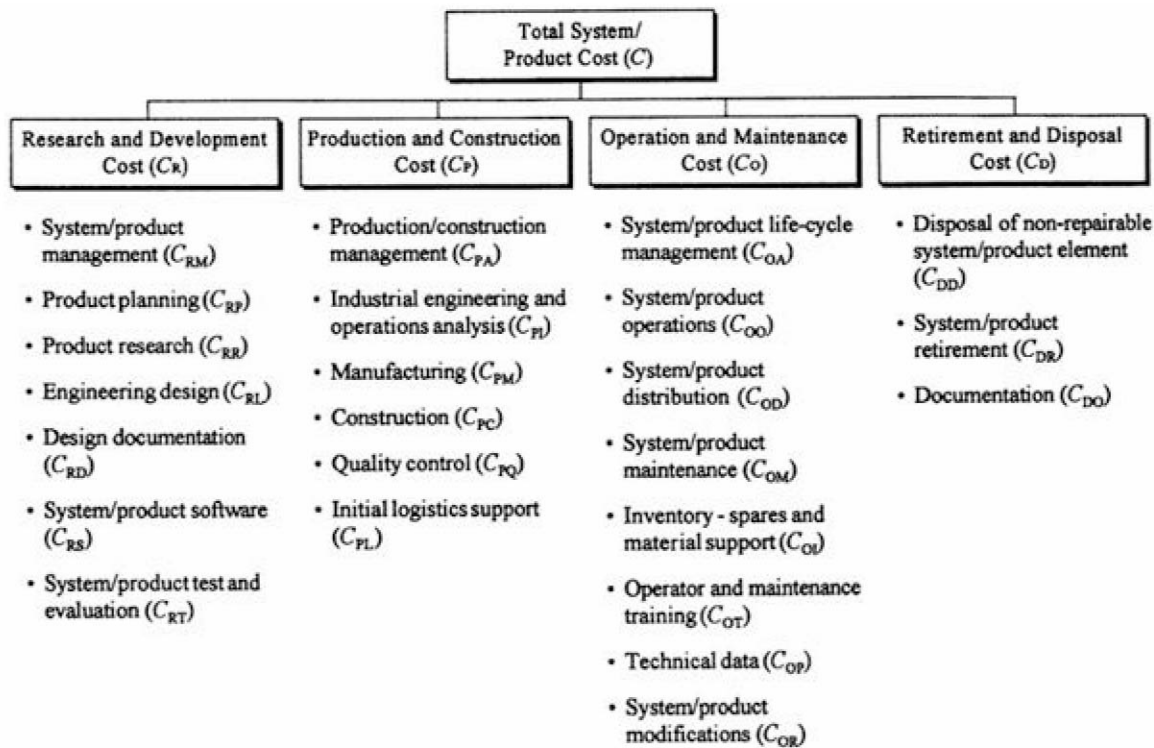
In general, costs over the life-cycle fall into categories based on organizational activity needed to bring a system into being. These categories and their constituent elements constitute a **cost breakdown structure (CBS)**, as illustrated in [Fig. 188.3](#). The main CBS categories are as follows:

1. *Research and development cost.* Initial planning, market analysis, feasibility studies, product research, requirements analysis, engineering design, design data and documentation, software, test and evaluation of engineering models, and associated management functions.
2. *Production and construction cost.* Industrial engineering and operations analysis, manufacturing (fabrication, assembly, and test), facility construction, process development, production operations, quality control, and initial logistic support requirements (e.g., initial consumer support, the manufacture of spare parts, the production of test and support equipment, etc.).
3. *Operation and support cost.* Consumer or user operations of the system or product in the field, product distribution (marketing and sales, transportation, and traffic management), and sustaining maintenance and logistic support throughout the system or product life cycle (e.g.,

customer service, maintenance activities, supply support, test and support equipment, transportation and handling, technical data, facilities, system modifications, etc.).

4. *Retirement and disposal cost.* Disposal of nonrepairable items throughout the life cycle, system/product retirement, material recycling, and applicable logistic support requirements.

**Figure 188.3** A general cost breakdown structure.



The cost breakdown structure links objectives and activities with organizational resource requirements. It constitutes a logical subdivision of cost by functional activity area, major system elements, and/or one or more discrete classes of common or like items. The CBS provides a means for initial resource allocation, cost monitoring, and cost control.

## 188.4 Life-Cycle Cost Analysis

The application of life-cycle costing methods during product and system design and development is realized through the accomplishment of **life-cycle cost analyses (LCCA)**. A life-cycle cost analysis may be defined as a systematic analytical process of evaluating various designs or alternative courses of action with the objective of choosing the best way to employ scarce resources.

Where feasible alternative solutions exist for a specific problem and a decision is required for the

selection of a preferred approach, there is a formal analysis process that should be followed. Specifically, the analyst should define the need for analysis, establish the analysis approach, select a model to facilitate the evaluation process, generate the appropriate information for each alternative being considered, evaluate each alternative, and recommend a proposed solution that is responsive to the problem.

## **Cost Analysis Goals**

There are many questions that the decision maker might wish to address. There may be a single overall analysis goal (e.g., design to minimum life-cycle cost) and any number of subgoals. The primary question should be as follows: What is the purpose of the analysis, and what is to be learned through the analysis effort?

In many cases the nature of the problem appears to be obvious, but its precise definition may be the most difficult part of the entire process. The design problem must be defined clearly and precisely and presented in such a manner as to be easily understood by all concerned. Otherwise, it is doubtful whether an analysis of any type will be meaningful. The analyst must be careful to ensure that realistic goals are established at the start of the analysis process and that these goals remain in sight as the process unfolds.

## **Analysis Guidelines and Constraints**

Subsequent to definition of the problem and the goals, the cost analyst must define the guidelines and constraints (or bounds) within which the analysis is to be accomplished. Guidelines are composed of information concerning such factors as the resources available for conducting the analysis (e.g., necessary labor skills, availability of appropriate software, etc.), the time schedule allowed for completion of the analysis, and/or related management policy or direction that may affect the analysis.

In some instances a decision maker or manager may not completely understand the problem or the analysis process and may direct that certain tasks be accomplished in a prescribed manner or time frame that may not be compatible with the analysis objectives. On other occasions a manager may have a preconceived idea as to a given decision outcome and direct that the analysis support the decision. Also, there could be external inhibiting factors that may affect the validity of the analysis effort. In such cases the cost analyst should make every effort to alleviate the problem by educating the manager. Should any unresolved problems exist, the cost analyst should document them and relate their effects to the analysis results.

Relative to the technical characteristics of a system or product, the analysis output may be constrained by bounds (or limits) that are established through the definition of system performance factors, operational requirements, the maintenance concept, and/or through advanced program planning. For example, there may be a maximum weight requirement for a given product, a minimum reliability requirement, a maximum allowable first cost per unit, a minimum rated capacity, and so on. These various bounds, or constraints, should provide for trade-offs in the evaluation of alternatives. Candidates that fall outside these bounds are not allowable.

## Identification of Alternatives

Within the established bounds and constraints, there may be any number of approaches leading to a possible solution. All possible alternatives should be considered, with the most likely candidates selected for further evaluation. Alternatives are frequently proposed for analysis even though there seems to be little likelihood that they will prove feasible. This is done with the thought that it is better to consider many alternatives than to overlook one that may be very good. Alternatives not considered cannot be adopted, no matter how desirable they may actually prove to be.

## Applying the Cost Breakdown Structure

Applying the cost breakdown structure is one of the most significant steps in life-cycle costing. The CBS constitutes the framework for defining life-cycle cost categories and provides the communications link for cost reporting, analysis, and ultimate cost control.

In developing the CBS one needs to proceed to the depth required to provide the necessary information for a true and valid assessment of the system or product life-cycle cost, identify high-cost contributors and enable determination of the cause-and-effect relationships, and illustrate the various cost parameters and their application in the analysis. Traceability is required from the system-level LCC figure of merit to the specific input factor.

## 188.5 Cost Treatment over the Life Cycle

---

With the system/product cost breakdown structure defined and cost-estimating approaches established, it is appropriate to apply the resultant data to the system life cycle. To accomplish this, the cost analyst needs to understand the steps required in developing cost profiles that include aspects of inflation, the effects of learning curves, the time value of money, and so on.

In developing a cost profile, there are different procedures that may be used. The following steps are suggested:

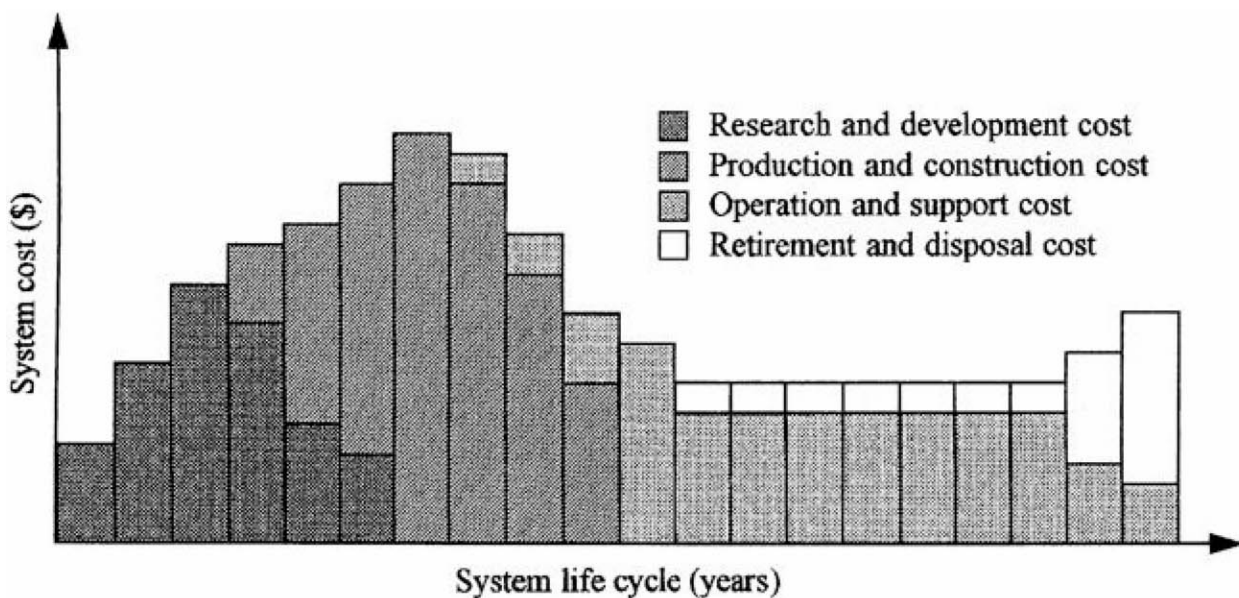
1. Identify all activities throughout the life cycle that will generate costs of one type or another. This includes functions associated with planning, research and development, test and evaluation, production/construction, product distribution, system/product operational use, maintenance and logistic support, and so on.
2. Relate each activity identified in step 1 to a specific cost category in the cost breakdown structure. All program activities should fall into one or more of the CBS categories.
3. Establish the appropriate cost factors in constant dollars for each activity in the CBS, where constant dollars reflect the general purchasing power of the dollar at the time of decision (i.e., today). Relating costs in terms of constant dollars will allow for a direct comparison of activity levels from year to year prior to the introduction of inflationary cost factors, changes in price levels, economic effects of contractual agreements with suppliers, and so on, which can often cause some confusion in the evaluation of alternatives.
4. Within each cost category in the CBS, the individual cost elements are projected into the future on a year-to-year basis over the life cycle as applicable. The result should be a cost



- stream in constant dollars for the activities that are included.
- For each cost category in the CBS and for each applicable year in the life cycle, introduce the appropriate inflationary factors, economic effects of learning curves, changes in price levels, and so on. The modified values constitute a new cost stream and reflect realistic costs as they are anticipated for each year of the life cycle (i.e., expected 1996 costs in 1996, 1997 costs in 1997, etc.). These costs may be used directly in the preparation of future budget requests, since they reflect the actual dollar needs anticipated for each year in the life cycle.
  - Summarize the individual cost streams by major categories in the CBS and develop a top-level cost profile.

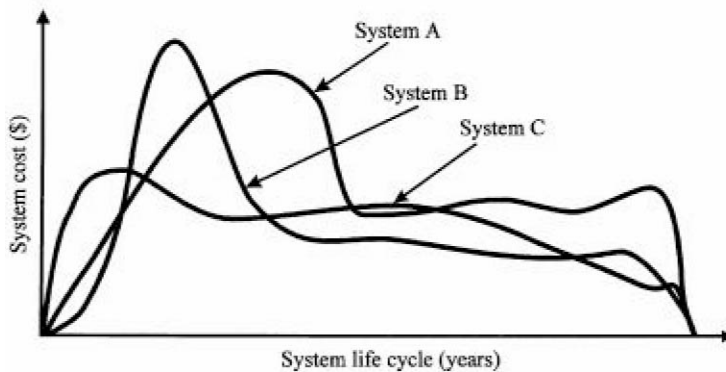
Results from the foregoing sequence of steps are presented in Fig. 188.4. First, it is possible and often beneficial to evaluate the cost stream for individual activities of the life cycle such as research and development, production, operation and support, and so on. Second, these individual cost streams may be shown in the context of the total cost spectrum. Finally, the total cost profile may be viewed from the standpoint of the logical flow of activities and the proper level and timely expenditure of dollars. The profile in Fig. 188.4 represents a budgetary estimate of future resource needs.

**Figure 188.4** Development of life-cycle cost profiles.



When dealing with two or more alternative system configurations, each will include different levels of activity, different design approaches, different logistic support requirements, and so on. No two systems alternatives will be identical. Thus, individual profiles will be developed for each alternative and ultimately compared on an equivalent basis utilizing the economic analysis techniques found in earlier chapters. Figure 188.5 illustrates life-cycle cost profiles for several alternatives.

**Figure 188.5** Life-cycle cost profiles of alternatives.



## 188.6 Summary

---

Life-cycle costing is applicable in all phases of system design, development, production, construction, operational use, and logistic support. Cost emphasis is created early in the life cycle by establishing quantitative cost factors as "design to" requirements. As the life cycle progresses, cost is employed as a major parameter in the evaluation of alternative design configurations and in the selection of a preferred approach. Subsequently, cost data are generated based on established design and production characteristics and used in the development of life-cycle cost projections. These projections, in turn, are compared with the initial requirements to determine the degree of compliance and the necessity for corrective action. In essence, life-cycle cost evolves from a series of rough estimates to a relatively refined methodology and is employed as a management tool for decision-making purposes.

### Defining Terms

**Cost breakdown structure (CBS):** A framework for defining life-cycle costs, and it provides the communications link for cost reporting, analysis, and ultimate cost control.

**Design-to-cost (DTC):** A concept that may be adopted to establish cost as a system or product design constraint, along with performance, effectiveness, capacity, accuracy, size, weight, reliability, maintainability, supportability, and others.

**Life-cycle cost (LCC):** All costs associated with the product or system as anticipated over the defined life cycle.

**Life-cycle cost analysis (LCCA):** A systematic analytical process for evaluating various alternative courses of action with the objective of choosing the best way to employ scarce resources.

**Life-cycle cost profile (LCCP):** A budgetary estimate of future resource needs over the life cycle.

### Reference

Fabrycky, W. J. and Blanchard, B. S. 1991. *Life-Cycle Cost and Economic Analysis*. Prentice Hall,



Englewood Cliffs, NJ.

## **Further Information**

The reference above should be studied by readers who want a complete view of life-cycle cost and economic analysis. Further information may be obtained from the following:

Blanchard, B. S. and Fabrycky, W. J. 1990. *Systems Engineering and Analysis*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.

Canada, J. R., and Sullivan, W. G. 1989. *Economic and Multiattribute Evaluation of Advanced Manufacturing Systems*. Prentice Hall, Englewood Cliffs, NJ.

Dhillon, B. S. 1989. *Life Cycle Costing: Techniques, Models and Applications*. Gordon and Breach Science Publishers, NY.

Ostwald, P. F. 1992. *Engineering Cost Estimating*, 3rd ed. Prentice Hall, Englewood Cliffs, NJ.

Thuesen, G. J. and Fabrycky, W. J. 1993. *Engineering Economy*, 8th ed. Prentice Hall, Englewood Cliffs, NJ.

Thamhain, H. J. "Project Evaluation and Selection"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Project Evaluation and Selection

---

## 189.1 Quantitative Approaches to Project Evaluation and Selection

Net Present Value (NPV) Comparison • Return-on-Investment Comparison • Pay-Back Period (PBP) Comparison • Pacifico and Sobelman Project Ratings • Limitations of Quantitative Methods

## 189.2 Qualitative Approaches to Project Evaluation and Selection

Collective Multifunctional Evaluations

## 189.3 Recommendations for Effective Project Evaluation and Selection

A Final Note

### Hans J. Thamhain

*Bentley College*

For most organizations, resources are limited. The ability to select and fund the best projects with the highest probability of success is crucial to an organization's ability to survive and prosper in today's highly competitive environment. Project selections, necessary in virtually every business area, cover activities ranging from product developments to organizational improvements, customer contracts, R&D activities, and bid proposals. Evaluation and selection methods support two principal types of decisions:

1. Judging the chances of success for one proposed project
2. Choosing the best project among available alternatives

Although most decision processes evaluate projects in terms of cost, time, risks and benefits, such as shown in [Table 189.1](#), it is often extremely difficult, if not impossible, to define a meaningful aggregate measure for ranking projects regarding business success, technical risks, or profit. Managers can use traditional, purely rational selection processes toward "right," "successful," and "best" only for a limited number of business situations. Many of today's complex project evaluations require the integration of both analytical and judgmental techniques to be meaningful.

**Table 189.1** Typical Criteria for Project Evaluation and Selection

---

The criteria relevant to the evaluation and selection of a particular project depend on the specific project type and business situation such as project development, custom project, process development, industry and market. Typically, evaluation procedures include the following criteria:

- Development cost
- Development time
- Technical complexity

- Risk
- Return on investment
- Cost benefit
- Product life cycle
- Sales volume
- Market share
- Project business follow-on
- Organizational readiness and strength
- Consistency with business plan
- Resource availability
- Cash flow, revenue, and profit
- Impact on other business activities

Each criterion is based on a complex set of parameters and variables.

---

Although the literature offers a great variety of project selection procedures, each organization has its own special methods. Approaches fall into one of three principal classes:

1. Primarily quantitative and rational approaches
2. Primarily qualitative and intuitive approaches
3. Mixed approaches, combining both quantitative and qualitative methods

## 189.1 Quantitative Approaches to Project Evaluation and Selection

---

Quantitative approaches are often favored to support project evaluation and selections if the decisions require economic justification. Supported by numeric measures for simple and effective comparison, ranking, and selection, they help to establish quantifiable norms and standards and lead to repeatable processes. However, the ultimate usefulness of these methods depends on the assumption that the decision parameters—such as cash flow, risks, and the underlying economic, social, political, and market factors—can actually be quantified. Typically, these quantitative techniques are effective decision support tools if meaningful estimates of capital expenditures and future revenues can be obtained and converted into net present values for comparison. [Table 189.2](#) shows the cash flow of four project options to be used for illustrating the quantitative methods described in this chapter.

**Table 189.2** Cash Flow of Four Project Options or Proposals\*

End of Year	Do-Nothing Option P1	Project Option P2	Project Option P3	Project Option P4
0	0	–1000	–2000	–5000
1	0	200	1500	1000
2	0	200	1000	1500
3	0	200	800	2000
4	0	200	900	3000
5	0	200	1200	4000

Net cash flow	0	0	+3400	+7 500
NPV <sub>IN</sub> = 5	0	-242	+2153	+3 192
NPV <sub>IN</sub> = ∞	0	+1000	+9904	+28 030
ROI <sub>IN</sub> = 5	0	20%	54%	46%
CB = ROI <sub>NPV</sub>   $n = 5$	0	67%	108%	164%
$N_{PBP}   i = 0$	0	5	1.5	3.3
$N_{NPV}   i$	0	7.3	5	3.8

\*Assuming an MARR of  $i = 10\%$

*Note:* The first line of negative numbers represents the initial investment at the beginning of the life cycle.

## Net Present Value (NPV) Comparison

This method uses discounted cash flow as the basis for comparing the relative merit of alternative project opportunities. It assumes that all investment costs and revenues are known and that economic analysis is a valid singular basis for project selection.

We can determine the *net present value* (NPV) of a single revenue, stream of revenues, and/or costs expected in the future.

*Present worth* of a single revenue or cost (often called annuity,  $A$ ) occurring at the end of period  $n$  and subject to an effective interest rate  $i$  (sometimes referred to as discount rate or **minimum attractive rate of return, MARR**) can be calculated as:

$$PW(A | i, n) = A \frac{1}{(1 + i)^n} = PW_n$$

*Net present value* of a series of revenues or costs,  $A_n$ , over  $N$  periods of time is as follows:

$$NPV(A_n | i, N) = \sum_{n=1}^N A_n \frac{1}{(1 + i)^n} = \sum_{n=1}^N PW_n$$

[Table 189.2](#) applies these formulas to four project alternatives, showing the most favorable 5-year net present value of \$3192 for project option P4. (There are three special cases of net present value: (1) for a uniform series of revenues or costs over  $N$  periods,  $NPV(A_n | i, N) = A[(1+i)^N - 1]/i(1+i)^N$ ; (2) for an annuity or interest rate  $i$  approaching zero,  $NPV = A \times N$ ; and (3) for the revenue or cost series to continue forever,  $NPV = A/i$ .)

## Return-on-Investment Comparison

Perhaps one of the most popular measures for project evaluation is the *return on investment* (ROI):

$$ROI = \frac{\text{Revenue } (R) - \text{Cost } (C)}{\text{Investment } (I)}$$

It calculates the ratio of net revenue over investment. In its simplest form it entails the revenue on a year-by-year basis relative to the initial investment (for example, project option 2 in [Table 189.2](#) would produce a 20% ROI). Although this is a popular measure, it does not permit a comparative evaluation of alternative projects with fluctuating costs and revenues. In a more sophisticated way we can calculate the average ROI per year

$$\overline{\text{ROI}}(A_n, I_n \mid N) = \left[ \sum_{n=1}^N \frac{A_n}{I_n} \right] / N$$

and compare it to the minimum attractive rate of return. All three project options, P2, P3 and P4, compare favorably to the MARR of 10%, with project P3 yielding the highest average return on investment of 54%. Or we can calculate the net present value of the total ROI over the project lifecycle, also known as *cost-benefit* (CB). This is an effective measure of comparison, especially for fluctuating cash flows. ([Table 189.2](#) shows project 3 with the highest 5-year  $\text{ROI}_{\text{NPV}}$  of 108%.)

$$\text{ROI}_{\text{NPV}}(A_n, I_n \mid i, N) = \left[ \sum_{n=1}^N \text{NPV}(A_n \mid i, N) \right] / \left[ \sum_{n=1}^N \text{NPV}(I_n \mid i, N) \right]$$

## Pay-Back Period (PBP) Comparison

Another popular figure of merit for comparing project alternatives is the *payback period* (PBP). It indicates the time period of net revenues required to return the capital investment made on the project. For simplicity, *undiscounted* cash flows are often used to calculate a quick figure for comparison, which is quite meaningful if we deal with an initial investment and a steady stream of net revenue. However, for fluctuating revenue and/or cost streams, the net present value must be calculated for each period individually and cumulatively added up to the "break-even point" in time,  $N_{\text{PBP}}$ , when the net present value of revenue equals the investment.

$$\sum_{n=1}^N \text{NPV}(A_n \mid i) \geq \sum_{n=1}^N \text{NPV}(I_n \mid i)$$

## Pacifico and Sobelman Project Ratings

The previously discussed methods of evaluating projects rely heavily on the assumption that technical and commercial success is ensured and all costs and revenues are predicable. Because of these limitations, many companies have developed their own special procedures for comparing project alternatives. Examples are the *project rating factor* (PR), developed by Carl Pacifico for

assessing chemical products, and the *project value factor* ( $z$ ), developed by Sidney Sobelman for new product selections:

$$PR = \frac{pT \times pC \times R}{TC} \qquad z = (P \times T_{LC}) - (C \times T_D)$$

Pacifico's formula is in essence an ROI calculation adjusted for risk. It includes probability of technical success [ $.1 < pT < 1.0$ ], probability of commercial success [ $.1 < pC < 1.0$ ], total net revenue over project life cycle [ $R$ ], and total capital investment for product development, manufacturing setup, marketing, and related overheads [ $TC$ ]. The Sobelman formula is a modified cost-benefit measure. It takes into account both the development time and the commercial life cycle of the product. It includes average profit per year [ $P$ ], estimated product life cycle [ $T_{LC}$ ], average development cost per year [ $C$ ], and years of development [ $T_D$ ].

## Limitations of Quantitative Methods

Although quantitative methods of project evaluation have the benefit of producing relatively quickly a measure of merit for simple comparison and ranking, they also have many limitations, as summarized in [Table 189.3](#).

**Table 189.3** Comparison of Quantitative and Qualitative Approaches to Project Evaluation

Quantitative Methods	Qualitative Methods
Benefits:	Benefits:
Simple comparison, ranking, selection	Search for meaningful evaluation metrics
Repeatable process	Broad-based organizational involvement
Encourages data gathering and measurability	Understanding of problems, benefits, opportunities
Benchmarking opportunities	Problem solving as part of selection process
Programmable	Broad knowledge base
Input to sensitivity analysis and simulation	Multiple solutions and alternatives
	Multifunctional involvement leads to buy-in
	Risk sharing
Limitations:	Limitations:
Many success factors are nonquantifiable	Complex, time-consuming process
Probabilities and weights change	Biases via power and politics
True measures do not exist	Difficult to proceduralize or repeat
Analysis and conclusions are often misleading	Conflict- and energy-intensive
Methods mask unique problems and opportunities	Do not fit conventional decision processes
Stifle innovative decision making	Intuition and emotion dominates over facts
Lack people involvement, buy-in, commitment	Justify wants over needs
Do not deal well with multifunctional issues and dynamic situations	Lead to more fact finding than decision making
Pressure to act prematurely	

## 189.2 Qualitative Approaches to Project Evaluation and Selection

---

Especially for project evaluations involving complex sets of business criteria, the narrowly focused quantitative methods must often be supplemented by broad-scanning; intuitive processes; and collective, multifunctional decision making, such as Delphi, nominal group technology, brainstorming, focus groups, sensitivity analysis, and benchmarking. Each of these techniques can either be used by itself to determine the best, most successful, or most valuable option, or be integrated into an analytical framework for *collective multifunctional decision making*, which is discussed in the next section.

### Collective Multifunctional Evaluations

Collective multifunctional evaluations rely on subject experts from various functional areas for collectively defining and evaluating broad **project success** criteria, employing both quantitative and qualitative methods. The first step is to define the specific organizational areas critical to project success and to assign expert evaluators. For a typical product development project, these organizations may include R&D, engineering, testing, manufacturing, marketing, product assurance, and customer services. These function experts should be given the time necessary for the evaluation. They also should have the commitment from senior management for full organizational support. Ideally, these evaluators should have the responsibility for ultimate project implementation, should the project be selected.

The next step is for the evaluation team to define the factors that appear critical to the ultimate success of the projects under evaluation and arrange them into a concise list that includes both quantitative and qualitative factors. A mutually acceptable scale must be worked out for scoring the evaluation criteria. Studies of collective multifunctional assessment practices show that simplicity of scales is crucial to a workable team solution. Three types of scale have produced most favorable results in field studies: (1) 10-point scale, ranging from +5 = most favorable to -5 = most unfavorable; (2) 3-point scale, +1 = favorable, 0 = neutral or can't judge, -1 = unfavorable; and (3) 5-point scale, A = highly favorable, B = favorable, C = marginally favorable, D = most likely unfavorable, F = definitely unfavorable. **Weighing of criteria** is not recommended for most applications, since it complicates and often distorts the collective evaluation.

Evaluators score first individually all of the factors that they feel qualified to make an expert judgment on. Collective discussions follow. Initial discussions of project alternatives, their markets, business opportunities, and technologies involved are usually beneficial but not necessary for the first round of the evaluation process. The objective of this first round of expert judgments is to get calibrated on the opportunities and challenges presented. Further, each evaluator has the opportunity to recommend (1) actions needed for better assessment of project, (2) additional data needed, and (3) suggestions that would enhance project success and the evaluation score. Before meeting at the next group session, agreed-on action items and activities for improving the decision process should be completed. With each iteration the function-expert meetings are enhanced with more refined project data. Typically, between three and five iterations are required before a project



selection can be finalized.

## 189.3 Recommendations for Effective Project Evaluation and Selection

---

Effective evaluation and selection of project opportunities is critical to overall project success. With increasing complexities and dynamics of the business environment, most situations are too complex to use simple economic models as the sole basis for decision making. To be effective, project evaluation procedures should include a broad spectrum of variables for defining the project value to the organization.

Structure, discipline, and manageability can be designed into the selection process by grouping the evaluation variables into four categories: (1) consistency and strength of the project with the business mission, strategy, and plan; (2) multifunctional ability to produce the project results, including technical, cost, and time factors; (3) success in the customer environment; and (4) economics, including profitability. Modern **phase management** and **stage-gate processes** provide managers with the tools for organizing and conducting project evaluations effectively. [Table 189.4](#) summarizes suggestions that may help managers to effectively evaluate projects for successful implementation.

**Table 189.4** Suggestions for Effective Project Evaluation and Selection

---

1. **Seek out relevant information.** Meaningful project evaluations require relevant quality information. The four categories of variables can provide a framework for establishing the proper metrics and detailed data gathering.
2. **Take top-down look; detail comes later.** Detail is less important than information relevancy and evaluator expertise. Don't get hung up on lack of data during the early phases of the project evaluation. Evaluation processes should iterate. It does not make sense to spend a lot of time and resources gathering perfect data to justify a "no-go" decision.
3. **Select and match the right people.** Whether the project evaluation consists of a simple economic analysis or a complex multifunctional assessment, competent people from those functions critical to the overall success of the project(s) should be involved.
4. **Success criteria must be defined.** Deciding on a single project or choosing among alternatives, evaluation criteria must be defined. They can be quantitative, such as ROI, or qualitative, such as the probability of winning a contract. In either case, these evaluation criteria should cover the true spectrum of factors affecting success and failure of the project(s). Only functional experts, discussed in point 3, are qualified to identify these success criteria. Often, people from outside the company, such as vendors, subcontractors, or customers, must be included in this expert group.
5. **Strictly quantitative criteria can be misleading.** Be aware of evaluation procedures based only on quantitative criteria (ROI, cost, market share, MARR, etc). The input data used to calculate these criteria are likely based on rough estimates and are often unreliable.

Evaluations based on predominately quantitative criteria should at least be augmented with some expert judgment as a "sanity check."

6. **Condense criteria list.** Combine evaluation criteria, especially among the judgmental categories, to keep the list manageable. As a goal, try to stay within 12 criteria.
  7. **Communicate.** Facilitate communications among evaluators and functional support groups. Define the process for organizing the team and conducting the evaluation and selection process.
  8. **Ensure cross-functional cooperation.** People on the evaluation team must share a strategic vision across organizational lines. They also must sense the desire of their host organizations to support the project if selected for implantation. The purpose, goals, and objectives of the project should be clear, along with the relationship to the business mission.
  9. **Don't lose the big picture.** As discussions go into detail during the evaluation, the team should maintain a broad perspective. Two global judgment factors can help to focus on the big picture of project success: (1) overall benefit-to-cost perception and (2) overall risk-of-failure perception. These factors can be recorded on a ten-point scale: -5 to +5. This also leads to an effective two-dimensional graphic display of competing project proposals.
  10. **Do your homework between iterations.** As project evaluations are most likely conducted progressively, action items for more information, clarification, and further analysis surface. These action items should be properly assigned and followed up, thereby enhancing the quality of the evaluation with each consecutive iteration.
  11. **Stimulate innovation.** Senior management should foster an innovative ambience for the evaluation team. Evaluating complex project situations for potential success or failure involves intricate sets of variables, linked among organization, technology, and business environment. It also involves dealing with risks and uncertainty. Innovative approaches are required to evaluate the true potential of success for these projects. Risk sharing by senior management, recognition, visibility, and a favorable image in terms of high priority, interesting work, and importance of the project to the organization have been found strong drivers toward attracting and holding quality people to the evaluation team and gaining their active and innovative participation in the process.
  12. **Manage and lead.** The evaluation team should be chaired by someone who has the trust, respect, and leadership credibility with the team members. Further, management can positively influence the work environment and the process by providing some procedural guidelines, charters, visibility, resources, and active support to the project evaluation team.
- 

## A Final Note

Effective project evaluation and selection requires a broad-scanning process that can deal with the risks, uncertainties, ambiguities, and imperfections of data available at the beginning of a project cycle. It also requires managerial leadership and skills in planning, organizing, and

communicating. Above all, evaluation team leaders must be social architects in unifying the multifunctional process and its people. They must share risks and foster an environment that is professionally stimulating and strongly linked with the support organizations eventually needed for project implementation. This is an environment that is conducive to **cross-functional** communication, cooperation, and integration of the intricate variables needed for effective project evaluation and selection.

## Defining Terms

**Cross-functional:** Actions that span organizational boundaries.

**Minimum attractive rate of return (MARR):** The annual net revenue produced on average by projects in an organization as a percentage of their investments. Sometimes MARR is calculated as company earnings over assets.

**Net worth:** Discounted present value of a future revenue or cost.

**Phase management:** Projects are broken into natural implementation phases, such as development, production, and marketing, as a basis for project planning, integration, and control. Phase management also provides the framework for *concurrent engineering* and *stage-gate processes*.

**Project success:** A comprehensive measure, defined in both quantitative and qualitative terms, that includes economic, market, and strategic objectives.

**Stage-gate process:** Framework for executing projects within predefined stages (see also **phase management**) with measurable deliverables (*gate*) at the end of each stage. The gates provide the review metrics for ensuring successful transition and integration of the project into the next stage.

**Weighing of criteria:** A multiplier associated with specific evaluation criteria.

## References

- Brenner, M. 1994. Practical R&D project prioritization. *Res. Technol. Manage.* 37(5):38–42.
- Bulick, W. J. 1993. Project evaluation procedures. *Cost Eng.* 35(10):27–32.
- Menke, M. M. 1994. Improving R&D decisions and execution. *Res. Technol. Manage.* 37(5):25–32.
- Obradovitch, M. M. and Stephanou, S. E. 1990. *Project Management: Risk and Productivity*. Daniel Spencer, Bend, OR.
- Remer, D. S., Stokdyk, S. B., and Van Driel, M. 1993. Survey of project evaluation techniques currently used in industry. *Int. J. Prod. Econ.* 32(1):103–115.
- Schmidt, R. L. 1993. A model for R&D project selection. *IEEE Trans. EM.* 40(4):403–410.
- Shtub, A., Bard, J. F., and Globerson, S. 1994. *Project Management: Engineering, Technology, and Implementation*. Prentice Hall, Englewood Cliffs, NJ.
- Skelton, M. T. and Thamhain, H. J. 1993. Concurrent project management: a tool for technology transfer. *Proj. Manage. J.* 26(4):41–48.
- Ward, T. J. 1994. Which product is BE\$T? *Chemical Eng.* 101(1):102–107.

## Further Information

The following journals are good sources of further information: *Engineering Management Journal* (ASEM), *Engineering Management Review* (IEEE), *Industrial Management* (IIE), *Journal of Engineering and Technology Management*, *Project Management Journal* (PMI), and *Transactions on Engineering Management* (IEEE).

The following professional societies present annual conferences and specialty publications that include discussions on project evaluation and selection: American Society for Engineering Management (ASEM), Rolla, MO 65401, (314) 341-2101; Institute of Electrical and Electronic Engineers (IEEE), East 47 St., New York, NY 10017-2394; and Project Management Institute (PMI), Upper Darby, PA 19082, (610)734-3330.

Richards, J. L. "Critical Path Method"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

### 190.1 Planning the Project

Project Performance Measures • Activity Time-Cost Trade-off • Activity Interrelationships • Work Breakdown Structure

### 190.2 Scheduling the Project

CPM Network Models • CPM Network Calculations

### 190.3 Controlling the Project

Managing Time and Money • Hierarchical Management • Managing the Schedule

### 190.4 Modifying the Project Schedule

Cost Duration Analysis • Critical Resource Analysis • Combined CDA and CRA

### 190.5 Project Management Using CPM

## John L. Richards

*University of Pittsburgh*

The purpose of this chapter is to describe the three-step, iterative decision-making process of planning, scheduling, and controlling with the critical path method (CPM). CPM is a network-based analytical tool that models a project's activities and their predecessor/successor interrelationships. **Planning** is the development of a **work breakdown structure (WBS)** of the project's activities. **Scheduling** is the calculation of **activity parameters** by doing a forward and a backward pass through the network. **Controlling** is the monitoring of the schedule during project execution by **updating** and **upgrading**, as well as the modifying of the schedule to achieve feasibility and optimality using cost duration analysis and critical resource analysis.

## 190.1 Planning the Project

---

Project planning requires the development of a work breakdown structure, which then becomes the basis for a network model of the project. This model can then be used to evaluate the project by comparing regular measures of performance.

### Project Performance Measures

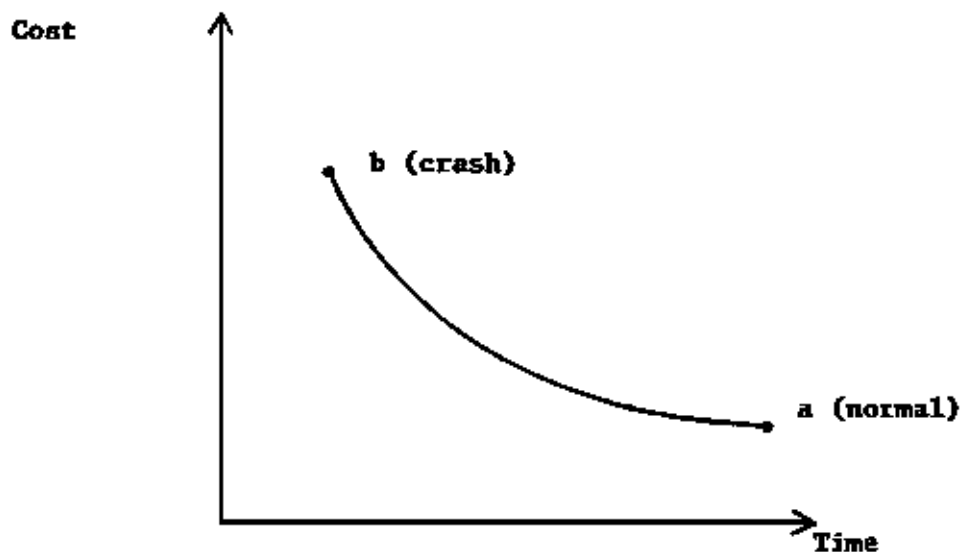
The three common performance measures in project management are *time*, *cost*, and *quality*. The overall objective is to accomplish the project in the least time, at the least cost, with the highest quality. Individually, these objectives conflict with each other. Thus, the manager must seek an overall solution by trading off among them. Further, since the overall project is defined by the

activities that must be done, the overall project duration, cost, and quality will be determined by the individual activity times, cost, and quality levels.

## Activity Time-Cost Trade-off

For a specified quality level for a given activity, the manager initially selects that combination of resources (labor, equipment, and material) that accomplishes that particular activity at the least cost. This is the *normal* duration on an *activity time-cost trade-off curve* (Fig. 190.1). Thus, since each activity is to be done at its least cost, the overall project will be done at the least total cost. However, in order to reduce an activity's duration, the activity cost must increase. For example, one can work overtime at premium rates or use more expensive equipment, which increases cost, in order to reduce an activity's duration. *Crash* is the shortest possible activity duration, no matter how high the cost. The inverse relationship between time and cost yields curves with negative slopes.

**Figure 190.1** Activity time-cost trade-off curve. The normal time (point a) is the least cost/longest activity duration. The crash time (point b) is the least activity duration/highest cost.



## Activity Interrelationships

There are two possible relationships between a pair of activities in a project network: (1) one must immediately precede the other (*predecessor*), or (2) one must immediately follow the other (*successor*). If there is no predecessor/successor relationship, the activities may be done simultaneously. These predecessor/successor relationships are derived from absolute constraints such as physical/technological, safety, and legal factors; or imposed constraints such as the selection of resources, methods, and financing. The manager should initially incorporate

relationships derived only from absolute constraints. Relationships derived from imposed constraints should be added only as necessary to achieve feasibility. This approach to predecessor/successor relationships yields the least constrained project network initially.

The basic predecessor/successor relationship is finish to start with no lead or lag. However, more sophisticated models allow three other types: start to finish, finish to finish, and start to start. In addition, each of the four could have a lead time or a lag time. Thus there are twelve possible ways to describe a particular predecessor/successor relationship.

## Work Breakdown Structure

The work breakdown structure (WBS) of a project is the listing of all the individual activities that make up the project, their durations, and their predecessor/successor relationships. It should be the least costly and least constrained method of executing the project, that is, normal durations and absolute constraints only. Therefore, if the schedule resulting from this initial WBS is feasible, then it is also optimal. If, however, the schedule is infeasible because of time and/or resource considerations, then the manager would want to achieve feasibility with the least additional cost. (Scheduling and feasibility/optimality are discussed in later sections).

There are three approaches to developing a WBS: (1) by physical components, (2) by process components, and (3) by spatial components. *Physical components* model a constructed product (e.g., build wall). *Process components* model a construction process (e.g., mix concrete). *Spatial components* model a use of time or space (e.g., order steel or cure concrete). No matter which of the three approaches is used to define a particular activity in a WBS, each should be described by an action verb to distinguish an activity from an event. (Note: There can be special activities in a WBS that involve time only and no cost, such as curing concrete. There can also be dummy activities for logic only that have no time or cost.)

A project's WBS must be developed to an appropriate level of detail. This means that activities must be broken down sufficiently to model interrelationships among them. Also, a standard time period (hour, shift, day, week, etc.) must be chosen for all activities. An appropriate WBS will have a reasonable number of activities and reasonable activity durations.

**Example 190.1¾ Work Breakdown Structure.** Table 190.1 shows a WBS for constructing a backyard in-ground swimming pool with vinyl liner. Note that a time period of days was selected for all activities.

**Table 190.1** WBS of Swimming Pool Construction

Activity ID	Duration	Description	Immediate Predecessors
A101	10	Order and deliver filtration equipment	—
A202	5	Order and deliver liner/piping	—
B301	4	Excavate for pool	—
B202	3	Install liner/piping	A202, B301
B102	2	Install filtration	A101



		equipment	
C301	2	Fill pool	B202
B401	5	Construct deck	B202
C302	2	Connect and test system	C301, B102
B501	3	Landscape area	B401

---

## 190.2 Scheduling the Project

---

The critical path method of project scheduling is based upon a network model that can be analyzed by performing a forward and a backward pass to calculate activity parameters.

### CPM Network Models

There are two types of network models: activity oriented and event oriented. Both types have nodes connected by arrows that model events (points in time) and activities (processes over time).

#### Activity-Oriented Diagram

The activity-oriented diagram is also called an *arrow diagram* (ADM) or *activity on arrow* (AOA). The activities are the arrows, and the events are the nodes. Dummy activities (depicted as dashed arrows) may be required to correctly model the project for logic only. Activity identification is by node pairs (node  $i$  to node  $j$ ). This was the original diagramming method. It is easy to visualize but difficult to draw.

#### Event-Oriented Diagram

The event-oriented diagram is also called a *precedence diagram* (PDM), *activity on node* (AON), or *circle network*. The activities are the nodes, and the events are the ends of arrows. There are no dummies. All arrows are for logic only, which also allows for modeling the twelve types of activity interrelationships discussed earlier. This diagramming method is easier to draw and well suited to computer use. Although developed after AOA, the AON has become the preferred method.

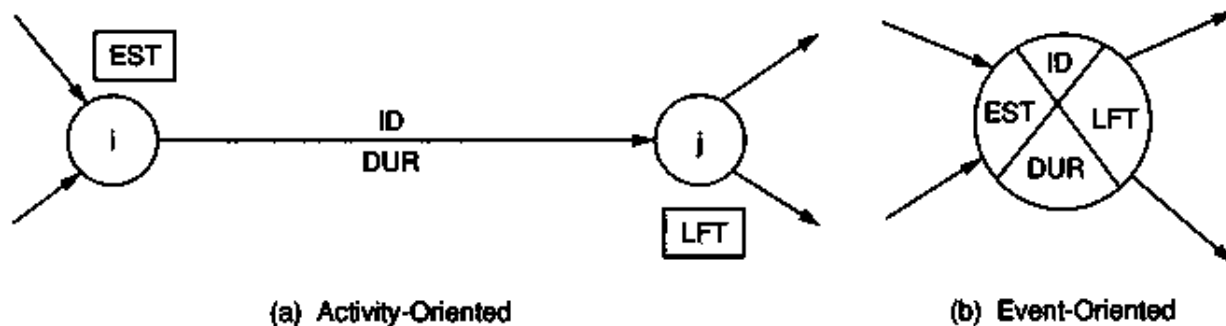
### CPM Network Calculations

There are three steps in the manual analysis of a CPM network that determine the activity parameters. The forward pass determines the *earliest start time* (EST) of each activity. The backward pass determines the *latest finish time* (LFT) of each activity. The other times, the *earliest finish time* (EFT) and the *latest start time* (LST), and the floats, *total float* (TF) and *free float* (FF), are then determined from a table. The calculation process is the same for either type of network (ADM or PDM). Before beginning the process, one needs to establish a time convention—beginning of time period or end of time period. The example in this chapter uses beginning of time period, thus the first day is day one. (The end-of-time-period convention would begin with day zero.)

## Forward Pass

To determine the EST of an activity, one compares all the incoming arrows—that is, the *heads* of arrows—choosing the *largest* earliest event time. The comparison is made among all the immediately preceding activities by adding their ESTs and respective durations. The process begins at the start node (the EST being zero or one) and proceeds to the end node, taking care that all incoming arrows get evaluated. At the completion of the forward pass, one has determined the overall project duration. The ESTs can be placed on diagrams as shown in [Fig. 190.2](#).

**Figure 190.2** Activity graphical models. The activity identification (ID) and duration (DUR) are from the WBS. The EST and LFT are calculated from the forward and backward passes, respectively.



## Backward Pass

The backward pass begins at the end node with the overall project duration (which is the LFT) and proceeds to the start node. This process determines the LFT for each activity by comparing the outgoing arrows—that is, the *tails* of arrows—choosing the *smallest* latest event time. The comparison is made among all the immediately succeeding activities by subtracting their durations from their respective LFTs. At the completion of the backward pass one should calculate the original project start time. The LFTs can be placed on diagrams as shown in [Fig. 190.2](#).

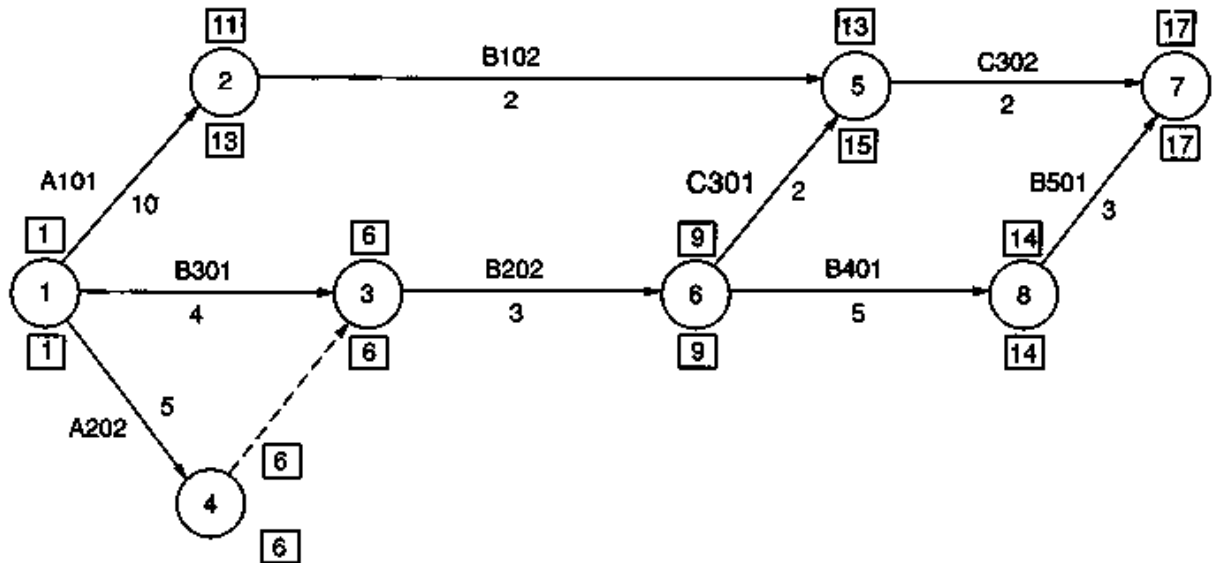
## Floats

The other two times (EFT and LST) and floats (TF and FF) are determined in a tabular format using the following relationships: (1)  $EFT = EST + \text{duration}$ , (2)  $LST = LFT - \text{duration}$ , (3)  $TF = LFT - EFT$  or  $TF = LST - EST$ , and (4)  $FF = EST \text{ (of following activities)} - EFT$ . Activities with a  $TF = 0$  are on a **critical path**. There may be more than one critical path in a network. If the duration of any critical path activity is increased, the overall project duration will increase. Activities with  $TF > 0$  may be increased without affecting the overall project duration. On the other hand, free float is that amount of total float that can be used by the activity without affecting any other activities. If  $TF$  equals 0, then  $FF$  equals 0, necessarily. Free float may be some, all, or no portion of total float.

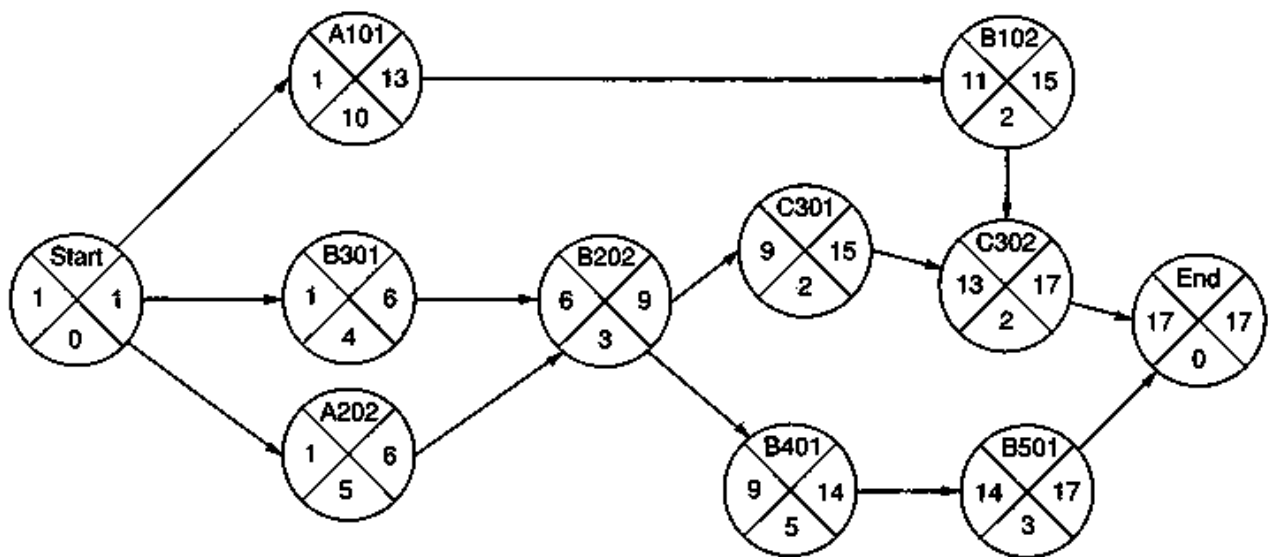
**Example 190.2 CPM Scheduling.** This example continues with the project introduced in [Table](#)

190.1. The CPM calculations for the forward and backward passes are shown in Fig. 190.3 for both network types. Table 190.2 lists all the times and floats.

**Figure 190.3** CPM network calculations (forward and backward passes). The overall project duration was determined to be 16 days ( $17 - 1 = 16$ ).



(a) Activity-Oriented Network (ADM or AOA)  
Note that activity 4-3 is a dummy and is also on the critical path.



(b) Event-Oriented Network (PDM or AON)  
Note that implied start and end nodes are shown with zero durations.

**Table 190.2** Activity Times and Floats

Activity ID	Duration	EST	EFT	LST	LFT	TF	FF
A101	10	1	11	3	13	2	0
A202*	5	1	6	1	6	0	0
B102	2	11	13	13	15	2	0
B202*	3	6	9	6	9	0	0
B301	4	1	5	2	6	1	1
B401*	5	9	14	9	14	0	0
B501*	3	14	17	14	17	0	0
C301	2	9	11	13	15	4	2
C302	2	13	15	15	17	2	2

\* Activities on critical path.

## 190.3 Controlling the Project

CPM-based project management provides the tools to control time and money in a dynamic and hierarchical project environment during project execution.

### Managing Time and Money

Computerized project management systems can provide a variety of informational outputs for use in managing a project. These include network diagrams showing activity interrelationships, bar charts showing activity durations, tabular listings showing activity parameters, and profiles showing cash flows or resource utilization during the project.

### Hierarchical Management

Project management generally occurs in a multiproject environment with multiple time parameters and multiple managerial levels. Multiple project models integrate cash flow and resource profiles of several projects. Multiple calendar models allow activities to be done at different time frames (for example, some may be done five days per week, others seven days per week). Project management information systems can provide summary information for upper-level management and detailed information for workers in the field.

### Managing the Schedule

The project schedule is a dynamic managerial tool that changes during project execution.

#### Updating the Schedule

As the project proceeds the schedule should be **updated** at periodic intervals to reflect actual activity progress. Such updates can incorporate percent complete, remaining durations, and actual start and end dates for each activity. The updates can be the basis for evaluating overall project objectives (time and money) and for making progress payments. After any update, a new schedule calculation must be done to determine new times and floats.

## Upgrading the Schedule

Any change to an existing schedule either by changing a planned duration or an activity relationship is a schedule **upgrade**. A schedule upgrade can occur either prior to start of the project, or any time during the project, based on new information. After any upgrade a new schedule calculation must be done to determine new times and floats.

## Managing the Floats

A negative total float indicates that the project will overrun the stated completion date. A positive ( $> 0$ ) total float indicates that the project will be completed earlier than the stated completion date. A particular network will generally have many different total float paths, including negative ones if it is behind schedule, and no zero ones if it is ahead of schedule.

Free float indicates that amount of time an activity can be manipulated without affecting any other activity (and therefore the project as a whole). When managing the floats, one would want to use free float before total float. Once total float is used, the activity becomes part of a new critical path.

The use of the floats can be a difficult contractual issue among the parties. It is a good idea to decide beforehand how the floats can be used. Otherwise the issue may be part of a delay claim at the completion of the project.

**Example 190.3%4Updating and Upgrading.** This example continues with the project introduced in [Table 190.1](#) and originally scheduled in [Table 190.2](#).

1. *Updating.* Assume that it is the beginning of day 6 and that activity B301 is done, and A101 has 9 days left and A202 has 2 days left. These updated durations are used in [Fig. 190.3](#) to recalculate activity times and floats as shown in [Table 190.3](#). Notice that a second critical path has developed and that the overall project duration has been extended two days.
2. *Upgrading.* Now assume that immediately after updating, the duration for activity C302 gets changed to three days and B401 must precede it. These upgrades could be incorporated into a revised [Fig. 190.3](#) (not shown) to recalculate another set of activity times and floats as shown in [Table 190.4](#). Notice that there is now only one critical path and the overall project duration has been extended another day to day 20.

**Table 190.3** Activity Times and Floats (Update)

Activity ID	Duration	EST	EFT	LST	LFT	TF	FF
A101*	9	6	15	6	15	0	0
A202*	2	6	8	6	8	0	0
B102*	2	15	17	15	17	0	0
B202*	3	8	11	8	11	0	0
B301	0	—	—	—	—	—	—
B401*	5	11	16	11	16	0	0

B501*	3	16	19	16	19	0	0
C301	2	11	13	15	17	4	4
C302*	2	17	19	17	19	0	0

\* Activities on critical path.

**Table 190.4** Activity Times and Floats (Upgrade)

Activity ID	Duration	EST	EFT	LST	LFT	TF	FF
A101*	9	6	15	6	15	0	0
A202	2	6	8	7	9	1	0
B102*	2	15	17	15	17	0	0
B202	3	8	11	9	12	1	0
B301	0	—	—	—	—	—	—
B401	5	11	16	12	17	1	0
B501	3	16	19	17	20	1	1
C301	2	11	13	15	17	4	4
C302*	3	17	20	17	20	0	0

\* Activities on critical path.

## 190.4 Modifying the Project Schedule

Project planning, scheduling, and controlling is an iterative decision-making process. It is highly unlikely for an initial schedule to be both feasible and optimal in the first iteration. Likewise, it is highly unlikely that the actual project execution will match the original project plan exactly. Therefore, one must know how to modify the project schedule in order to achieve feasibility and optimality. The modification process involves either changing activity duration, changing activity relationships, or both.

### Cost Duration Analysis

**Cost duration analysis (CDA)** utilizes activity time/cost trade-off curves (discussed earlier) in order to compress the overall project schedule. The objective is to buy back each time unit in the cheapest possible manner until the desired completion date is reached (feasibility). Only activities on a critical path need be reduced. The others with positive float simply have the float reduced. The problem can become very complex in a large network with multiple critical paths where the incremental additional costs for the activities are different.

### Critical Resource Analysis

The approach to **critical resource analysis (CRA)** is different from CDA in that it seeks to extend the overall project duration the least amount in order to resolve resource conflicts (i.e., achieve feasibility). CRA can be viewed from one of two perspectives: (1) constrained resources—staying

below a specified limit, or (2) resource leveling—selecting a constant limit. The solution approach for either is the same. The problem is one of ordering (predecessor/successor relationships) those activities that have resource conflicts during the same time period. The pairwise comparison of all such activities in a large network with many critical resources presents a huge combinatorial problem. The only viable solution approaches are based upon heuristic decision rules. (A simple rule could be that the predecessor activity should be the one with the smaller LST.)

## Combined CDA and CRA

Combining CDA and CRA to achieve a feasible and optimal schedule is virtually impossible for all but the simplest networks. Although the CDA problem does have rigorous mathematical solutions, they are not incorporated in most commercial software. On the other hand, the software generally does incorporate heuristic-based solutions for the CRA problem. Therefore, one should use the software in an interactive decision-making manner.

**Example 190.4<sup>3/4</sup> CDA and CRA.** Assume that after the upgrade as shown in [Table 190.4](#), it is decided that the desired completion date is day 19, and also that activities B102 and B401 cannot be done simultaneously because of insufficient labor. Further, assume that B401 can be reduced from 5 days to 3 days at an additional cost of \$200 per day and that C302 can be reduced from 3 days to 2 days at an additional cost of \$400.

**Solution.** The solution approach to this problem is to work two cases: (1) B102 precedes B401, and compress B401 and/or C302 if they lie on a critical path, and (2) B401 precedes B102, and again compress B401 and/or C302. Case 1 would yield an overall project duration of 25 days, and one can readily see that it is impossible to reduce it 6 days (one can only gain a total of 3 days from activities B401 and C302). Case 2 (B401 precedes B102) yields an overall project duration of 21 days, with both B401 and C302 on the same critical path. One should choose the least expensive method to gain one day—that is, change B401 to 4 days for \$200. This yields a project duration of 20 days and an additional critical path. Therefore, reducing B401 another day to 3 days does not get the project to 19 days. Instead, the more expensive activity (C302) must be reduced from 3 to 2 days for \$400. The answer to the problem, then, is B401 goes from 5 to 4 days for \$200, and C302 goes from 3 to 2 days for \$400. Thus, the overall project duration is compressed from 21 to 19 days from a total additional cost of \$600.

## 190.5 Project Management Using CPM

---

CPM was first developed in the late 1950s by the Remington Rand Corporation and the DuPont Chemical Company. Since then, many software manufacturers have developed sophisticated computer-based management information systems using CPM. In addition to performing the CPM calculations discussed in this chapter, such systems can provide data for creating the historical file of an ongoing project, for developing estimating information for a future project, and for performance evaluation of both the project and the participating managers. CPM has even become a well-accepted means for analyzing and resolving construction disputes.

The successful use of CPM as a managerial tool involves not only the analytical aspects discussed in this chapter, but also the attitude that is displayed by those using it in actual practice. If CPM is used improperly as a weapon, rather than as a tool, there will be project management failure. Therefore, successful project management must include positive team building among all project participants, along with the proper application of the critical path method.

## Defining Terms

**Activity parameters:** The activity times (EST, EFT, LFT, and LST) and activity floats (TF and FF) calculated in the scheduling step.

**Controlling:** The third step in the interactive decision-making process, which monitors the accomplishments of the project by updating and upgrading and seeks feasibility and optimality by cost duration analysis and critical resource analysis.

**Cost duration analysis (CDA):** Reducing durations of selected activities in the least costly manner in order to achieve a predetermined project completion date.

**Critical path (CP):** String of activities from start to finish that have zero total float. There may be more than one CP, and the CPs may change after an update or upgrade.

**Critical resource analysis (CRA):** Sequencing selected activities in such a manner as to minimize the increase in project duration in order to resolve resource conflicts among competing activities.

**Planning:** The first step in the interactive decision-making process, which determines the work breakdown structure.

**Scheduling:** The second step in the interactive decision-making process, which determines the activity parameters by a forward and a backward pass through a network.

**Update:** Changing remaining activity durations due to progress only, then rescheduling.

**Upgrade:** Changing activity durations and interrelationships due to new information only, then rescheduling.

**Work breakdown structure (WBS):** Listing of the individual activities that make up the project, their durations, and their predecessor/successor relationships.

## References

- Antill, J. M. and Woodhead, R. W. 1990. *Critical Path Methods in Construction Practise*, 4th ed. John Wiley & Sons, New York.
- Hendrickson, C. and Au, T. 1989. *Project Management for Construction*. Prentice Hall, Englewood Cliffs, NJ.
- Moder, J. J., Philips, C. R., and Davis, E. W. 1983. *Project Management with CPM, PERT and Precedence Diagramming*, 3rd ed. Van Nostrand Reinhold, New York.

## Further Information

*Journal of Management in Engineering* and *Journal of Construction Engineering and Management*. Published by the American Society of Civil Engineers.

*Project Management Journal*. Published by the Project Management Institute.



*The Construction Specifier*. Published by the Construction Specifications Institute.  
*Journal of Industrial Engineering*. Published by the American Institute of Industrial Engineers.

Beck, P. A., Bordas, C. I. "Patents, Copyrights, Trademarks, and Licenses"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## Patents, Copyrights, Trademarks, and Licenses

---

- 191.1 Patents
- 191.2 Copyrights
- 191.3 Trademarks
- 191.4 Licenses

**Paul A. Beck**

*Paul A. Beck & Associates*

**Carol I. Bordas**

*Thorp, Reed & Armstrong*

The origins of U.S. intellectual property law—which includes patents, copyrights, and trademarks—can be traced to medieval England. Letters Patent were issued by the king of England under the Statute of Monopolies in 1623 to confer the exclusive privilege of working or making a new manufacture for a 14-year term to the true and first inventor. In 1710 the Statute of Anne was enacted to both limit and create copyright monopolies in the publishing industry. Within the medieval guilds, individual artisans such as jewelers placed marks on their products in order that a purchaser could identify the source of their product.

Today a patent, copyright, trademark, or license can be used to protect intellectual property rights. For example, a computer program used in a process for manufacturing a new drug can be protected by a patent, a copyright, a trademark, and a license. A patent can be obtained for the process of manufacturing and to protect the use of the computer program in manufacturing the drug. A trademark can be used to identify the origin of the computer program. A copyright is created in the source code, which prevents the copying of the program. The owner of the copyrighted program can enter into a license agreement in which the licensee can distribute and reproduce the program in return for royalties payable to the licensor-owner.

### 191.1 Patents

---

A patent can be viewed as a contract between the inventor and the government in which the government grants the inventor the right to exclude others from making, using, or selling the invention in exchange for the inventor fully disclosing the invention to the public. In order for an

inventor to obtain patent protection in the U.S., an inventor must apply for a patent and be granted a U.S. Letters Patent by the U.S. Patent and Trademark Office (USPTO). Three types of patents are granted by the USPTO. Utility patents are granted for processes, machines, manufactures, compositions of matter, and any improvements thereof that are useful, new, and unobvious to one of ordinary skill in the art. Design patents are granted for any new, original, and ornamental design of an article of manufacture. Plant patents are granted for any distinct and new variety of plant, including those asexually produced and those found in an uncultivated state.

A utility patent application includes (1) a description of the invention, (2) at least one **claim** that defines the boundaries of protection sought by the patent applicant, and (3) any drawings necessary to understand the invention. The description of the invention must disclose the **best mode** of practicing the invention known by the inventor at the time of filing the patent application.

In order to be granted a U.S. patent the utility patent application must describe an invention that is useful, novel, and unobvious. The invention is novel if the invention was conceived and reduced to practice before any **prior art**. The invention is unobvious if the invention would have been unobvious to one of ordinary skill in the art at the time of the filing of the patent application in view of the prior art [Chisum, 1994]. If the invention was patented or described in any printed publication, in public use, or on sale in the U.S. more than one year before the U.S. patent application filing date, the inventor is barred by law from being granted a patent.

There is no one patent that can grant worldwide protection for an invention. Generally, if an inventor seeks to protect an invention in a foreign country, the inventor must apply for a patent in that specific country. The term of a patent in a foreign country usually extends for 20 years from the filing date of the patent application.

Patent infringement is the making, using, or selling of a patented invention during the term of the patent without the permission of the owner of the patent. The patented invention is defined by the claims of the patent. When every element recited in a claim of the patent or an equivalent of the element is present in an accused infringing device, direct infringement is established.

## 191.2 Copyrights

---

A copyright prohibits others from copying a work. Copyrights protect expressions of original work of authorship fixed in a tangible medium. Works that can be copyrighted include literary works, musical works, dramatic works, pantomimes and choreographic works, pictorial, graphic and sculptural works, motion pictures and other audiovisual works, sound recordings, and architectural works [Nimmer, 1993]. Copyrights only protect nonfunctional artistic features of a work.

The owner of a copyright has the exclusive right to reproduce the work, prepare **derivative works** based on the work, distribute the work, perform the work, and display the work. Generally, the owner of the copyright is the author of the work unless the work was specially commissioned or was produced in the scope of the employment of the author, in which case the employer is the author and owner of the copyright and the work is considered a work for hire.

Because a copyright is created at the moment the expression is fixed in a tangible medium, a copyright need not be registered with the U.S. Copyright Office to be valid. However, if the copyright is registered within five years of the publication of the work, the copyright is presumed to be valid. Further, registration is a prerequisite for a copyright infringement action in the courts.

A federal copyright registration application includes (1) an application form, (2) a filing fee, and (3) two **best editions** of the copyrighted work. The copyright application is examined for completeness, consistency of information, and appropriateness of subject matter. The term of a registered copyright is generally the life of the author plus 50 years. If the copyrighted work is a joint work, the term of the copyright is the length of the last surviving author plus 50 years. If the work is a work for hire, the copyright term is 75 years from the first publication or 100 years from the year of creation, whichever comes first.

To establish infringement one must prove both (1) ownership of a valid copyright, and (2) copying of the original work. Copying is usually established if a substantial similarity between the accused work and the copyrighted work exists, and if the accused infringer had access to the copyrighted work. Although copyright notice is no longer required, if notice is affixed to the infringed work, innocent infringement cannot be successfully claimed by the infringing party. Copyright notice consists of the word "Copyright" or an abbreviation thereof, the year of publication, and the name of the copyright owner. One can affix copyright notice to a work that has not been registered with the Copyright Office.

The U.S. is a signatory to the **Berne Convention** and as a result copyrighted works of U.S. citizens enjoy the same copyright protection as other signatory countries accord their nationals.

## 191.3 Trademarks

---

A trademark is a badge of origin placed on goods to signify the origin of the goods. A word or several words, a symbol, a color, or a scent can be utilized as a trademark. A mark that is distinguishable from other trademarks should be chosen in order that there is no confusion as to the origin of the goods [McCarthy, 1994].

Trademarks can be registered in the USPTO as well as with a state agency within each of the 50 states and in most foreign countries. In the case of federal trademark registration, the trademark must be based on actual use or based on a *bona fide intent to use* the trademark. A federal trademark registration will generally be granted unless the trademark (1) includes immoral, deceptive, or scandalous matter; (2) includes a flag or coat of arms or other insignia of a municipality or nation; (3) includes a name, portrait, or signature identifying a particular living individual without written consent of the individual or includes the name, signature, or portrait of a deceased president of the U.S. during the life of his widow without written consent; (4) resembles a trademark registered in the USPTO that when used can cause confusion; or (5) when used with the goods is merely descriptive, deceptively misdescriptive, geographically misdescriptive, or is merely a surname. However, if the applicant for registration can establish that the mark is distinctive such that the public identifies the mark with the goods—that is, the mark gains **secondary meaning**<sup>3/4</sup> then the mark's being descriptive, misdescriptive, or a surname will not prevent the applicant from obtaining a trademark registration.

An application for federal trademark registration based on actual use consists of (1) a written application that includes a description of the goods to which the mark is affixed, the date of first use of the mark in commerce, and the manner in which the mark is used; (2) a drawing of the mark; (3) **specimens** of the mark as it is used on the goods; and (4) the filing fee. An application based on an intent to use contains all of these four items except that the date of the first use of the

trademark and the description of the manner of use of the mark are both supplied to the USPTO within six months after a trademark has been allowed. A trademark registration has a term of ten years from the date of registration, which can be renewed for ten-year periods indefinitely.

Trademark infringement is established if a likelihood of confusion between the accused mark and the protected trademark exists. The following are some factors that are considered when determining if there is a likelihood of confusion: similarity of marks, similarity of goods, or similarity and character of markets.

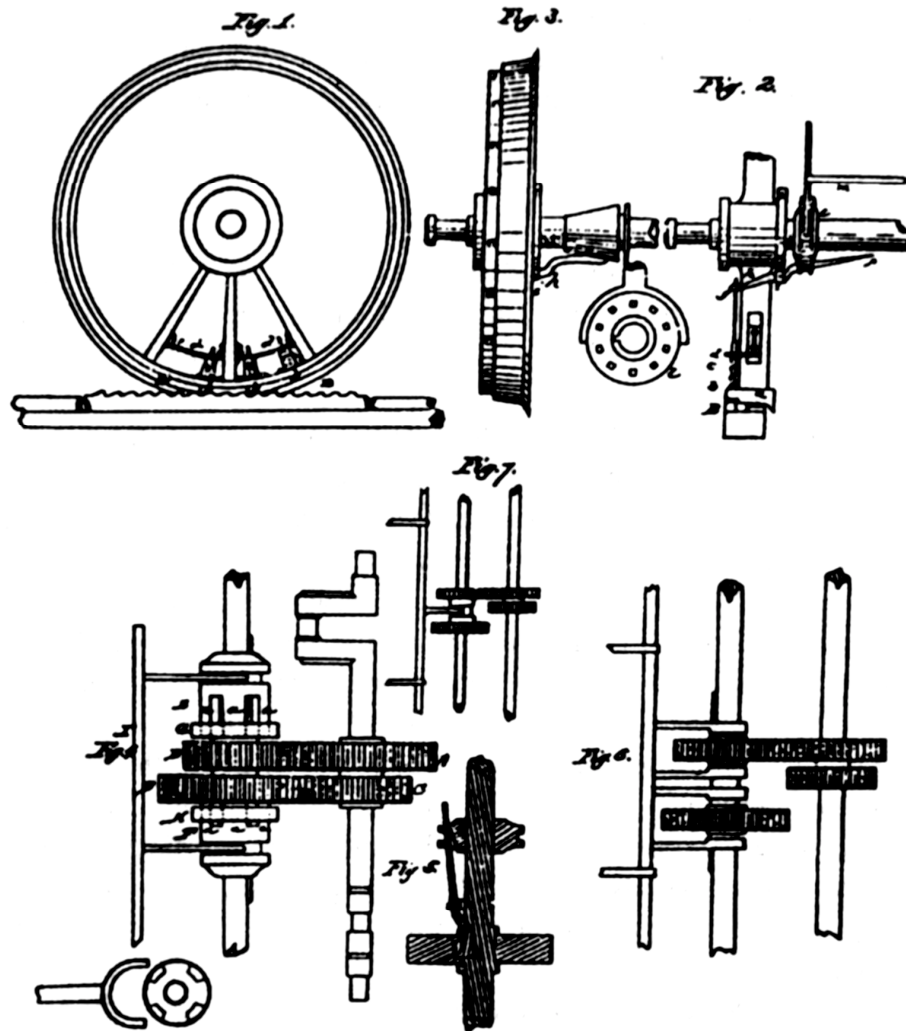
## 191.4 Licenses

---

A license can be granted by the owner of a patent, a copyright, or a trademark to permit another to use the owner's exclusive rights while retaining title in exchange for a **royalty**. A patent owner can license others to make, use, and sell the invention. The three rights can be licensed separately or in combination with one another. In other words the patent owner can license the right to make the invention to one individual and license the rights to use and sell the invention to a second individual. The license may be exclusive or nonexclusive. An **exclusive license** prohibits the patent owner from licensing the same right to another [Lipscomb, 1993], whereas a nonexclusive license permits the patent owner to enter into the same nonexclusive licensing agreement with more than one individual. Other terms and conditions can be added to a patent licensing agreement, such as a territory restriction.

A copyright license must take the form of a nonexclusive license. An exclusive license of the copyright owner's rights, even if limited in time and place, is considered a transfer of copyright ownership. An owner of a copyright can enter into a nonexclusive license agreement that grants any of the exclusive rights of a copyright. For example, the licensee can be granted the right to reproduce the copyright subject matter and distribute the copyrighted work but not display the copyrighted work. One joint author does not need the permission of the other joint author when he enters into a nonexclusive license agreement of his copyright. No recordation requirement for nonexclusive copyright licenses exists.

For a trademark license to be valid the licensor must control the nature and the quality of the goods or services sold under the mark. Controlling the use of the licensed trademark ensures that the licensee's goods are of equal quality to that of the licensor. If the quality of the goods and services designated by the trademark are changed, then the public could be deceived. If proper quality control is not practiced by the licensee, the trademark could be considered abandoned by the licensee.



# LOCOMOTIVE STEAM-ENGINE FOR RAIL AND OTHER ROADS

## John Ruggles

Patented July 13, 1836 #1

This was the first patent awarded under the Act of 1836. Ruggles, a senator from Maine, was in charge of the research leading up to the drafting of this sweeping legislation.

Because no examination of "prior art" was made between 1802 and 1836, patents were granted, in effect, to whomever applied. Patents were once again to be examined for novelty and usefulness. Specifications, drawings, and a model were to be furnished. The Patent Office was now a separate State Department bureau with a Commissioner and appointees to perform the necessary examinations and other clerical work. A sequential numbering system, still used today, made filing of the increasing number of patents easier and more coherent.

It seems more than a coincidence that Ruggles was awarded patent #1; perhaps it was in honor of his efforts to create the new patent system. He spent only one term in the Senate and returned to his law practice in Maine in 1841. (©1993, DewRay Products, Inc. Used with permission.)

## Defining Terms

- Berne Convention:** The Convention for the Protection of Literary and Artistic Works, signed at Berne, Switzerland, on 9 September 1986.
- Best edition:** The edition of a copyright work, published in the U.S. before the deposit, that the Library of Congress determines suitable.
- Best mode:** The specification must set forth the best mode contemplated by the inventor in carrying out the invention. The requirement is violated when the inventor knew of a superior method of carrying out the invention at the time of filing the application and concealed it.
- Claim:** An applicant must include one or more claims in a patent application that set forth the parameters of the invention. A claim recites a number of elements or limitations and covers only those inventions that contain all such elements and limitations.
- Derivative work:** A work based on one or more preexisting copyrighted works.
- Exclusive license:** A license that permits the licensee to use the patent, trademark, or copyright and prohibits another from using the patent, trademark, or copyright in the same manner.
- Prior art:** Those references that may be used to determine the novelty and nonobviousness of claimed subject matter in a patent application or a patent. Prior art includes all patents and publications.
- Royalty:** A payment to the owner of a patent, trademark, or copyright that is payable in proportion to the use of the owner's rights in the patent, trademark, or copyright.
- Secondary meaning:** A trademark acquires secondary meaning when the trademark is distinctive such that the public identifies the trademark with the goods. A descriptive trademark is protected only when secondary meaning is established. A generic trademark can never acquire secondary meaning.
- Specimen:** Trademark specimens consist of samples of promotional material bearing the trademark used for labeling of applicant's goods.

## References

- Chisum, D. S. 1994. Nonobviousness. In *Patents*, §5.01, June. Matthew Bender, New York.
- Lipscomb, E. B., III. 1993. Licenses. In *Lipscomb's Walker On Patents*, 3rd ed. Lawyers Co-operative, Rochester, NY.
- McCarthy, J. T. 1994. The fundamental principles of trademark protection. In *McCarthy on Trademarks and Unfair Competition*, 3rd ed., §2.07, June. Clark Boardman Callaghan, Deerfield, IL.
- Nimmer, M. B. and Nimmer, D. 1993. Subject matter of copyright. In *Nimmer on Copyright*,



§§2.04–2.10. Matthew Bender, New York.

## **Further Information**

The monthly *Journal of the Patent and Trademark Office Society* publishes legal articles in the fields of patents, trademarks and copyrights. For subscriptions, contact: P.O. Box 2600, Subscription Manager, Arlington, VA.

*Marketing Your Invention* is a booklet prepared by the American Bar Association Section of Patent, Trademark and Copyright Law. To obtain a copy, contact: American Bar Association Section of Patent, Trademark and Copyright Law, 750 North Lake Shore Drive, Chicago, IL.

The International Trademark Association produces and publishes a range of authoritative texts concerning trademark law. Among them are *Worldwide Trademark Transfer*, *U.S. Trademark Law*, and *State Trademark and Unfair Competition Law*. For more information regarding these publications, contact: International Trademark Association, 1133 Avenue of the Americas, New York, NY 10036-6710. Phone: (212)768-9887. Fax: 212.768.7796.

Shackelford, J. F. "Materials Engineering"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



The Prodigy is a state-of-the-art type of cementless implant. The main material used in this total hip replacement (THR) is a high-strength cobalt chrome. A clinically proven porous coating called Porocoat® is used on the ball and stem of the Prodigy. Porocoat® allows bone and tissue matter to grow into the porous material and enables the THR to be cementless.

The Prodigy stem and cobalt chrome ball mate with a titanium Duraloc acetabular cup. This cup has a congruent interface between the metal shell and a polyethylene liner. This liner is made of a polymer called Hylamer®, which offers enhanced creep resistance, increased wear resistance, improved resistance to oxidation, and superior quality control. (Photo courtesy of DePuy Inc.)

# XXIX

## Materials Engineering

---

**James F. Shackelford**

*University of California, Davis*

**192 Properties of Solids** *J. F. Shackelford*

Structure • Composition • Physical Properties • Mechanical Properties • Thermal Properties • Chemical Properties • Electrical and Optical Properties

**193 Failure Analysis** *J. F. Shackelford*

Types of Failures • Failure Analysis Methodology • Fracture Mechanics • Nondestructive Testing • Engineering Design for Failure Prevention

**194 Liquids and Gases** *B. E. Poling*

Viscosity • Thermal Conductivity • Heat Capacity • Vapor Pressure

**195 Biomaterials** *R. B. Martin*

History • Problems Associated with Implanted Devices • Immunology and Biocompatibility • Commonly Used Implant Materials • Metals • Polymers • Ceramics • Carbon Materials

THE PLACEMENT OF SECTIONS ON MATERIALS and mathematics at the end of the *Engineering Handbook* is symbolic of their general relationships to the full range of engineering disciplines. The central role of materials in the engineering profession is further suggested by the basic definition provided by the Accreditation Board for Engineering and Technology (ABET): "Engineering is the profession in which a knowledge of the mathematical and natural sciences gained by study, experience, and practice is applied with judgment to develop ways to utilize, economically, the *materials* and forces of nature for the benefit of mankind" [emphasis mine].

This section covers a wide range of issues dealing with engineering materials. First, the key properties of solids are reviewed. Representative tables of property data are provided to give general guidance in the selection of the appropriate materials for a given engineering design. We cannot provide a comprehensive source of data in this relatively brief introduction. We do give, however, a sense of the relative performance of various representative materials and references to more comprehensive sources such as *The CRC Materials Science and Engineering Handbook*.

Properties guide us in the selection of materials, but an increasingly important role for engineers is to evaluate problems which arise when the materials selected for an engineering application prove to be inadequate. In this regard, the systematic methodology of failure analysis is reviewed. Fracture mechanics and nondestructive testing play central roles in failure analysis. The growing body of work in this field is laying the foundation for failure prevention in future engineering designs.

It is important to remember that materials of significance to engineering are not all in the solid state. A discussion of liquids and gases outlines the nature of key fluids involved in a wide variety

of engineering systems.

Among the most intriguing applications of engineering materials are those used in biology and medicine. These "biomaterials" range from the polymers used in the cardiovascular system to the metallic alloys used in orthopedic joint replacement.

Shackelford, J. F. "Properties of Solids"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

- 192.1 Structure
- 192.2 Composition
- 192.3 Physical Properties
- 192.4 Mechanical Properties
- 192.5 Thermal Properties
- 192.6 Chemical Properties
- 192.7 Electrical and Optical Properties

**James F. Shackelford**

*University of California, Davis*

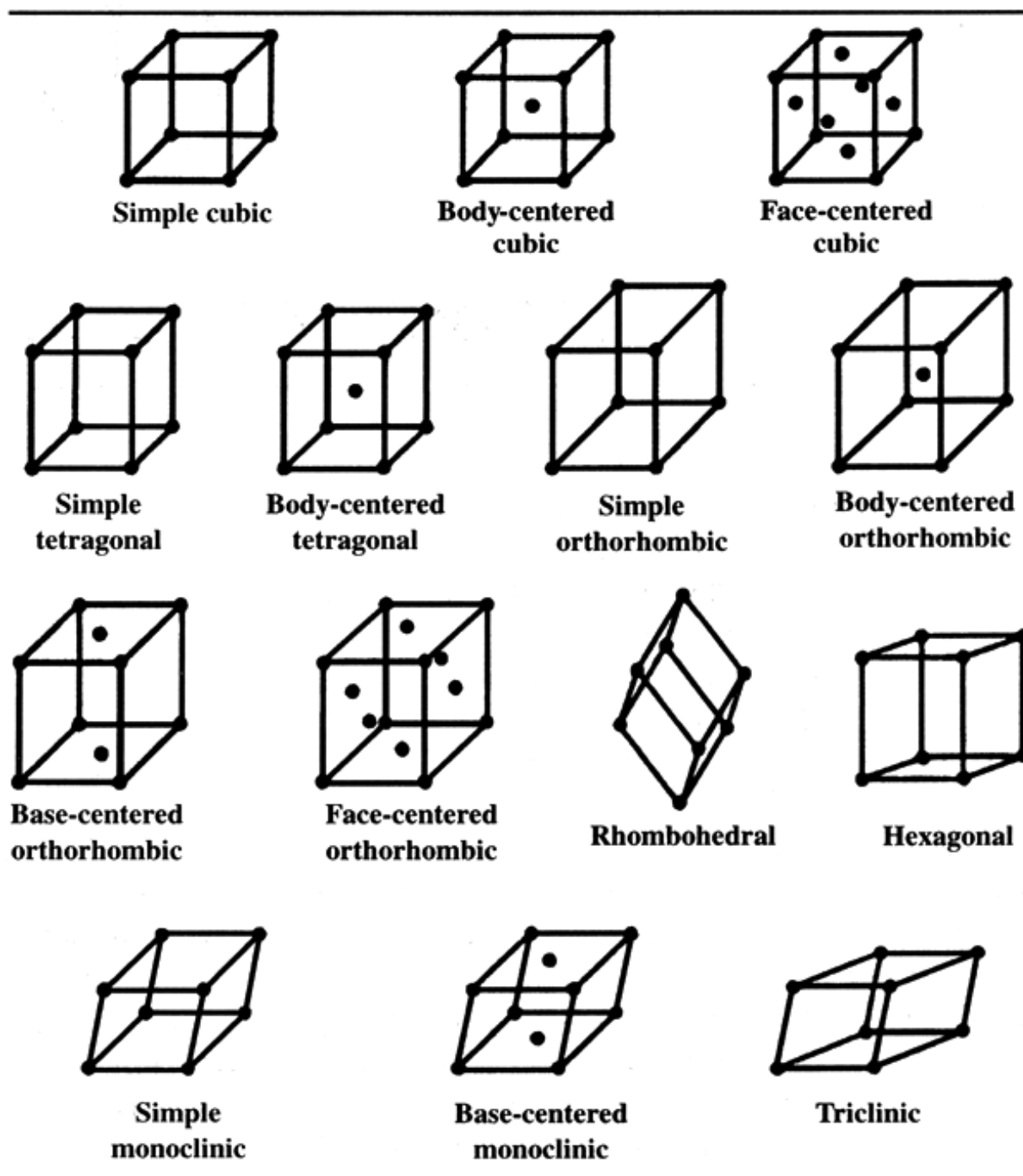
The term *materials science and engineering* refers to that branch of engineering dealing with the processing, selection, and evaluation of solid-state materials [Shackelford, 1992]. As such, this is a highly interdisciplinary field. This chapter reflects the fact that engineers outside of the specialized field of materials science and engineering largely need guidance in the selection of materials for their specific applications. A comprehensive source of property data for engineering materials is available in *The CRC Materials Science and Engineering Handbook* [Shackelford *et al.*, 1994]. This brief chapter will be devoted to defining key terms associated with the properties of engineering materials and providing representative tables of such properties. Because the underlying principle of the fundamental understanding of solid-state materials is the fact that atomic- or microscopic-scale structure is responsible for the nature of materials properties, we shall begin with a discussion of structure, which will be followed by a discussion of the importance of specifying the chemical composition of commercial materials. These discussions will be followed by the definition of the main categories of material properties.

## 192.1 Structure

---

A central tenet of materials science is that the behavior of materials (represented by their **properties**) is determined by their structure on the atomic and microscopic scales [Shackelford, 1992]. Perhaps the most fundamental aspect of the structure-property relationship is to appreciate the basic skeletal arrangement of atoms in **crystalline** solids. Table 192.1 illustrates the fundamental possibilities, known as the 14 **Bravais lattices**. All crystalline structures of real materials can be produced by "decorating" the unit cell patterns of Table 192.1 with one or more atoms and repetitively stacking the unit cell structure through three-dimensional space.

**Table 192.1** The Fourteen Bravais Lattices



## 192.2 Composition

The properties of commercially available materials are determined by chemical composition as



well as structure [Shackelford, 1992]. As a result, extensive numbering systems have been developed to label materials, especially metal **alloys**. Table 192.2 gives an example for gray cast irons.

**Table 192.2** Composition Limits of Selected Gray Cast Irons (%)

UNS	SAE Grade	C	Mn	Si	P	S
F10004	G1800	3.40 to 3.70	0.50 to 0.80	2.80 to 2.30	0.15	0.15
F10005	G2500	3.20 to 3.50	0.60 to 0.90	2.40 to 2.00	0.12	0.15
F10009	G2500	3.40 min	0.60 to 0.90	1.60 to 2.10	0.12	0.12
F10006	G3000	3.10 to 3.40	0.60 to 0.90	2.30 to 1.90	0.10	0.16
F10007	G3500	3.00 to 3.30	0.60 to 0.90	2.20 to 1.80	0.08	0.16
F10010	G3500	3.40 min	0.60 to 0.90	1.30 to 1.80	0.08	0.12
F10011	G3500	3.50 min	0.60 to 0.90	1.30 to 1.80	0.08	0.12
F10008	G4000	3.00 to 3.30	0.70 to 1.00	2.10 to 1.80	0.07	0.16
F10012	G4000	3.10 to 3.60	0.60 to 0.90	1.95 to 2.40	0.07	0.12

Data from ASM. 1984. *ASM Metals Reference Book*, 2nd ed., p. 166. American Society for Metals, Metals Park, OH.

## 192.3 Physical Properties

Among the most basic and practical characteristics of engineering materials are their physical properties. Table 192.3 gives the **density** of a wide range of materials in units of  $\text{Mg/m}^3$  ( $= \text{g/cm}^3$ ), whereas Table 192.4 gives the **melting points** for several common metals and ceramics.

**Table 192.3** Density of Selected Materials ( $\text{Mg/m}^3$ )

Metal		Ceramic		Glass		Polymer	
Ag	10.50	$\text{Al}_2\text{O}_3$	3.97–3.986	$\text{SiO}_2$	2.20	ABS	1.05–1.07
Al	2.7	BN (cub)	3.49	$\text{SiO}_2$ 10 wt% $\text{Na}_2\text{O}$	2.291	Acrylic	1.17–1.19
Au	19.28	BeO	3.01–3.03	$\text{SiO}_2$ 19.55 wt% $\text{Na}_2\text{O}$	2.383	Epoxy	1.80–2.00
Co	8.8	MgO	3.581	$\text{SiO}_2$ 29.20 wt% $\text{Na}_2\text{O}$	2.459	HDPE	0.96
Cr	7.19	SiC (hex)	3.217	$\text{SiO}_2$ 39.66 wt% $\text{Na}_2\text{O}$	2.521	Nylon, type 6	1.12–1.14
Cu	8.93	$\text{Si}_3\text{N}_4$ ( $\alpha$ )	3.184	$\text{SiO}_2$ 39.0 wt% CaO	2.746	Nylon 6/6	1.13–1.15
Fe	7.87	$\text{Si}_3\text{N}_4$ ( $\beta$ )	3.187			Phenolic	1.32–1.46
Ni	8.91	$\text{TiO}_2$ (rutile)	4.25			Polyacetal	1.425
Pb	11.34	$\text{UO}_2$	10.949–10.97			Polycarbonate	1.2

Pt	21.44	ZrO <sub>2</sub> (CaO)	5.5	Polyester	1.31
Ti	4.51	Al <sub>2</sub> O <sub>3</sub> MgO	3.580	Polystyrene	1.04
W	19.25	3Al <sub>2</sub> O <sub>3</sub> 2SiO <sub>2</sub>	2.6–3.26	PTFE	2.1–2.3

Selected data from Shackelford, J. F., Alexander, W., and Park, J. S. (Eds.) 1994. *CRC Materials Science and Engineering Handbook*, 2nd ed. CRC Press, Boca Raton, FL.

**Table 192.4** Melting Point of Selected Metals and Ceramics

Metal	M.P. (°C)	Ceramic	M.P. (°C)
Ag	962	Al <sub>2</sub> O <sub>3</sub>	2049
Al	660	BN	2727
Au	1064	B <sub>2</sub> O <sub>3</sub>	450
Co	1495	BeO	2452
Cr	1857	NiO	1984
Cu	1083	PbO	886
Fe	1535	SiC	2697
Ni	1453	Si <sub>3</sub> N <sub>4</sub>	2442
Pb	328	SiO <sub>2</sub>	1723
Pt	1772	WC	2627
Ti	1660	ZnO	1975
W	3410	ZrO <sub>2</sub>	2850

Selected data from Shackelford, J. F., Alexander, W., and Park, J. S. (Eds.) 1994. *CRC Materials Science and Engineering Handbook*, 2nd ed. CRC Press, Boca Raton, FL.

## 192.4 Mechanical Properties

Central to the selection of materials for structural applications is their behavior in response to mechanical loads. A wide variety of mechanical properties are available to help guide materials selection [Shackelford *et al.*, 1994]. The most basic of the mechanical properties are defined in terms of the **engineering stress** and the **engineering strain**. The engineering stress,  $\sigma$ , is defined as

$$\sigma = P/A_o \quad (192.1)$$

where  $P$  is the load on the sample with an original (zero stress) cross-sectional area  $A_o$ . The engineering strain,  $\varepsilon$ , is defined as

$$\varepsilon = [l - l_o]/l_o = \Delta l/l_o \quad (192.2)$$

where  $l$  is the sample length at a given load and  $l_o$  is the original (zero stress) length. The maximum engineering stress that can be withstood by the material during its load history is termed the *ultimate tensile strength*, or simply **tensile strength**, TS. An example of the tensile strength for

selected wrought (meaning "worked," as opposed to cast) aluminum alloys is given in [Table 192.5](#) . The "stiffness" of a material is indicated by the linear relationship between engineering stress and engineering strain for relatively small levels of load application. The **modulus of elasticity**,  $E$ , also known as **Young's modulus**, is given by the ratio

$$E = \sigma / \varepsilon \quad (192.3)$$

[Table 192.6](#) gives values of Young's modulus for selected compositions of glass materials. The "ductility" of a material is indicated by the percent elongation at failure ( $= 100 \times \varepsilon_{\text{failure}}$  ), representing the general ability of the material to be plastically (i.e., permanently) deformed. The percent elongation at failure for selected polymers is given in [Table 192.7](#).

**Table 192.5** Tensile Strength of Selected Wrought Aluminum Alloys

Alloy	Temper	TS (MPa)
1050	0	76
1050	H16	130
2024	0	185
2024	T361	495
3003	0	110
3003	H16	180
5050	0	145
5050	H34	195
6061	0	125
6061	T6, T651	310
7075	0	230
7075	T6, T651	570

Selected data from Shackelford, J. F., Alexander, W., and Park, J. S. (Eds.) 1994. *CRC Materials Science and Engineering Handbook*, 2nd ed. CRC Press, Boca Raton, FL.

**Table 192.6** Young's Modulus of Selected Glasses (GPa)

Type	$E$
SiO <sub>2</sub>	72.76–74.15
SiO <sub>2</sub> 20 mol % Na <sub>2</sub> O	62.0
SiO <sub>2</sub> 30 mol % Na <sub>2</sub> O	60.5
SiO <sub>2</sub> 35 mol % Na <sub>2</sub> O	60.2
SiO <sub>2</sub> 24.6 mol % PbO	47.1
SiO <sub>2</sub> 50.0 mol % PbO	44.1
SiO <sub>2</sub> 65.0 mol % PbO	41.2
SiO <sub>2</sub> 60 mol % B <sub>2</sub> O <sub>3</sub>	23.3
SiO <sub>2</sub> 90 mol % B <sub>2</sub> O <sub>3</sub>	20.9
B <sub>2</sub> O <sub>3</sub>	17.2–17.7
B <sub>2</sub> O <sub>3</sub> 10 mol % Na <sub>2</sub> O	31.4
B <sub>2</sub> O <sub>3</sub> 20 mol % Na <sub>2</sub> O	43.2

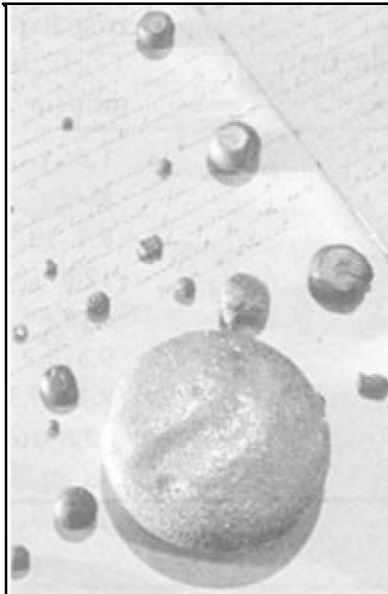
Selected data from Shackelford, J. F., Alexander, W., and Park, J. S. (Eds.) 1994. *CRC Materials Science and Engineering Handbook*, 2nd ed. CRC Press, Boca Raton, FL.

**Table 192.7** Total Elongation at Failure of Selected Polymers

Polymer	Elongation*
ABS	5–20
Acrylic	2–7
Epoxy	4.4
HDPE	700–1000
Nylon, type 6	30–100
Nylon 6/6	15–300
Phenolic	0.4–0.8
Polyacetal	25
Polycarbonate	110
Polyester	300
Polypropylene	100–600
PTFE	250–350

\*% in 50 mm section

Selected data from Shackelford, J. F., Alexander, W., and Park, J. S. (Eds.) 1994. *CRC Materials Science and Engineering Handbook*, 2nd ed. CRC Press, Boca Raton, FL.



#### ALUMINUM IN HISTORY

Original aluminum pellets shown on a page of hand-written minutes from a company meeting, circa 1890.

More than 7000 years ago, potters in ancient Persia made their strongest pots and bowls from a

clay that contained large amounts of a compound of aluminum and oxygen we know today as alumina. The Persians couldn't have known about aluminum, because it is not found in nature as a metal. However, they did discover one of the advantages that makes certain aluminum compounds so useful in ceramics today; it made their pots durable.

Three thousand years later, the Egyptians and Babylonians were using other aluminum compounds. Physicians used alum (a mixture of aluminum and sulphates) to control bleeding. Aluminum compounds also were used by these people for fabric dyes and cosmetics. The first man to produce aluminum metal may have been a Chinese chemist. Scientists believe this discovery took place about 300 B.C. A belt, decorated with aluminum, was found in the tomb of a Chinese general, known to have died about then.

However, the secret of producing aluminum in quantity eluded the world's finest scientists until the 19th century. The British chemist, Sir Humphry Davy, established the existence of aluminum in 1808 but failed to extract it from the ore. In 1825, the Danish physicist, Hans Christian Oersted, succeeded in producing a few droplets of aluminum.

For many years, a German chemist named Friederich (or Friedrich) Wöhler concentrated on improving Oersted's method. The credit for bringing aluminum out of the laboratory and into the marketplace (though its applications at the time were quite costly and specialized) generally goes to the French metallurgist, Henri Sainte-Claire DeVille. Improving on Wöhler's process, DeVille was able, in 1854, to produce aluminum the size of marbles. By 1869, DeVille was producing nearly two tons a year, and the price had dropped from \$545 to \$17 per pound. But aluminum was still far too costly to be made in quantities that people could afford.

It began to look as if only the wealthy would be able to purchase products made of aluminum. Aluminum tableware was used to serve honored guests at the French Court. Denmark's King Christian X wore an aluminum crown. In 1884, a six-pound aluminum cap was used to crown the top of the Washington Monument. It was probably the biggest single chunk of aluminum in the world at the time. The cap is still in place today.

As more aluminum was produced, scientific interest in the metal increased. Finally, in 1886, two inventors in their twenties, American Charles Martin Hall and Frenchman Paul L. T. Héroult, independently and within two months of each other, discovered a practical process for making aluminum inexpensively. That process is still used around the world today. (Courtesy of the Aluminum Company of America.)

## 192.5 Thermal Properties

---

Many applications of engineering materials depend on their response to a thermal environment. The **thermal conductivity**,  $k$ , is defined by **Fourier's law**:

$$k = -[dQ/dt]/[A(dT/dx)] \quad (192.4)$$

where  $dQ/dt$  is the rate of heat transfer across an area  $A$  due to a temperature gradient  $dT/dx$ . It is also important to note that the dimensions of a material will, in general, increase with temperature.

Increases in temperature lead to greater thermal vibration of the atoms and an increase in average separation distance of adjacent atoms. The **linear coefficient of thermal expansion**,  $\alpha$ , is given by

$$\alpha = dl/l dT \quad (192.5)$$

with  $\alpha$  having units of  $\text{mm}/(\text{mm}^\circ\text{C})$ . Examples of thermal conductivity and thermal expansion coefficient for alloy cast irons are given in [Table 192.8](#).

**Table 192.8** Thermal Conductivity and Thermal Expansion of Alloy Cast Irons

Alloy	Thermal Conductivity W/(m K)	Thermal Expansion Coefficient $\text{mm}/(\text{m}^\circ\text{C})$
Low-C white iron	22 <sup>a</sup>	12 <sup>b</sup>
Martensitic nickel-chromium iron	30 <sup>a</sup>	8 – 9 <sup>b</sup>
High-nickel gray iron	38–40	8.1–19.3
High-nickel ductile iron	13.4	12.6–18.7
Medium-silicon iron	37	10.8
High-chromium iron	20	9.3–9.9
High-nickel iron	37–40	8.1–19.3
Nickel-chromium-silicon iron	30	12.6–16.2
High-nickel (20%) ductile iron	13	18.7

<sup>a</sup> Estimated

<sup>b</sup> 10 to 260° C

Data from ASM. 1984. *ASM Metals Reference Book*, 2nd ed., p. 172. American Society for Metals, Metals Park, OH.

## 192.6 Chemical Properties

A wide variety of data are available to characterize the nature of the reaction between engineering materials and their chemical environments [[Shackelford et al., 1994](#)]. Perhaps no such data are more fundamental and practical than the **electromotive force series** of metals shown in [Table 192.9](#). The voltage associated with various half-cell reactions in standard aqueous environments are arranged in order, with the materials associated with more anodic reactions tending to be corroded in the presence of a metal associated with a more cathodic reaction.

**Table 192.9** Electromotive Force Series of Metals

Metal	Potential (V)	Metal	Potential (V)	Metal	Potential (V)
Anodic or Corroded End					
Li	–3.04	Al	–1.70	Pb	–0.13
Rb	–2.93	Mn	–1.04	H	0.00
K	–2.92	Zn	–0.76	Cu	0.52
Ba	–2.90	Cr	–0.60	Ag	0.80

Sr	-2.89	Cd	-0.40	Hg	0.85
Ca	-2.80	Ti	-0.33	Pd	1.0
Na	-2.71	Co	-0.28	Pt	1.2
Mg	-2.37	Ni	-0.23	Au	1.5
Be	-1.70	Sn	-0.14		

Cathodic or Noble Metal End

Data compiled by J. S. Park from Bolz, R. E. and Tuve, G. L. (Eds.) 1973. *CRC Handbook of Tables for Applied Engineering Science*. CRC Press, BocaRaton, FL.

## 192.7 Electrical and Optical Properties

To this point, we have concentrated on various properties dealing largely with the structural applications of engineering materials. In many cases the electromagnetic nature of the materials may determine their engineering applications. Perhaps the most fundamental relationship in this regard is **Ohm's law**, which states that the magnitude of current flow,  $I$ , through a circuit with a given resistance  $R$  and voltage  $V$  is related by:

$$V = IR \quad (192.6)$$

where  $V$  is in units of volts,  $I$  is in amperes, and  $R$  is in ohms. The resistance value depends on the specific sample geometry. In general,  $R$  increases with sample length,  $l$ , and decreases with sample area,  $A$ . As a result, the property more characteristic of a given material and independent of its geometry is **resistivity**,  $\rho$ , defined as

$$\rho = [RA]/l \quad (192.7)$$

The units for resistivity are ohm  $\cdot$  m (or  $\Omega \cdot \text{m}$ ). [Table 192.10](#) gives the values of electrical resistivity for various materials, indicating that metals typically have low resistivities (and correspondingly high electrical conductivities) and ceramics and polymers typically have high resistivities (and correspondingly low conductivities).

**Table 192.10** Electrical Resistivity of Selected Materials

Metal (Alloy Cast Iron)	$\rho$ ( $\Omega \cdot \text{m}$ )	Ceramic	$\rho$ ( $\Omega \cdot \text{m}$ )	Polymer	$\rho$ ( $\Omega \cdot \text{m}$ )
Low-C white cast iron	$0.53 \cdot 10^{-6}$	$\text{Al}_2\text{O}_3$	$> 10^{13}$	ABS	$2-4 \cdot 10^{13}$
Martensitic Ni-Cr iron	$0.80 \cdot 10^{-6}$	$\text{B}_4\text{C}$	$0.3-0.8 \cdot 10^{-2}$	Acrylic	$> 10^{13}$
High-Si iron	$0.50 \cdot 10^{-6}$	BN	$1.7 \cdot 10^{11}$	HDPE	$> 10^{15}$
High-Ni iron	$1.4-1.7 \cdot 10^{-6}$	BeO	$> 10^{15}$	Nylon 6/6	$10^{12}-10^{13}$
Ni-Cr-Si iron	$1.5-1.7 \cdot 10^{-6}$	MgO	$1.3 \cdot 10^{13}$	Phenolic	$10^7-10^{11}$
High-Al iron	$2.4 \cdot 10^{-6}$	SiC	$1-1 \cdot 10^{10}$	Polyacetal	$10^{13}$
Medium-Si ductile iron	$0.58-0.87 \cdot 10^{-6}$	$\text{Si}_3\text{N}_4$	$10^{11}$	Polypropylene	$> 10^{15}$
High-Ni (20%) ductile iron	$1.02 \cdot 10^{-6}$	$\text{SiO}_2$	$10^{16}$	PTFE	$> 10^{16}$

Selected data from Shackelford, J. F., Alexander, W., and Park, J. S. (Eds.) 1994. *CRC Materials Science and Engineering Handbook*, 2nd ed. CRC Press, Boca Raton, FL.

An important aspect of the electromagnetic nature of materials is their optical properties. Among the most fundamental optical characteristics of a light-transmitting material is the **index of refraction**,  $n$ , defined as

$$n = v_{\text{vac}}/v \quad (192.8)$$

where  $v_{\text{vac}}$  is the speed of light in vacuum (essentially equal to that in air) and  $v$  is the speed of light in a transparent material. The index of refraction for a variety of polymers is given in [Table 192.11](#).

**Table 192.11** Refractive Index of Selected Polymers

Polymer	$n$
Acrylic	1.485–1.500
Cellulose Acetate	1.46–1.50
Epoxy	1.61
HDPE	1.54
Polycarbonate	1.586
PTFE	1.35
Polyester	1.50–1.58
Polystyrene	1.6
SAN	1.565–1.569
Vinylidene chloride	1.60–1.63

Data compiled by J. S. Park from Lynch, C. T. (Ed.) 1975. *CRC Handbook of Materials Science, Volume 3*. CRC Press, Inc., Boca Raton, FL; and ASM. 1988. *Engineered Materials Handbook, Volume 2, Engineering Plastics*. ASM International, Metals Park, OH.

## Defining Terms

**Alloy:** Metal composed of more than one element.

**Bravais lattice:** One of the 14 possible arrangements of points in three-dimensional space.

**Crystalline:** Having constituent atoms stacked together in a regular, repeating pattern.

**Density:** Mass per unit volume.

**Electromotive force series:** Systematic listing of half-cell reaction voltages.

**Engineering strain:** Increase in sample length at a given load divided by the original (stress-free) length.

**Engineering stress:** Load on a sample divided by the original (stress-free) area.

**Fourier's law:** Relationship between rate of heat transfer and temperature gradient.

**Index of refraction:** Ratio of speed of light in vacuum to that in a transparent material.

**Linear coefficient of thermal expansion:** Material parameter indicating dimensional change as a function of increasing temperature.

**Melting point:** Temperature of transformation from solid to liquid upon heating.



**Ohm's law:** Relationship between voltage, current, and resistance in an electrical circuit.

**Property:** Observable characteristic of a material.

**Resistivity:** Electrical resistance normalized for sample geometry.

**Tensile strength:** Maximum engineering stress during a tensile test.

**Thermal conductivity:** Proportionality constant in Fourier's law.

**Young's modulus (modulus of elasticity):** Ratio of engineering stress to engineering strain for relatively small levels of load application.

## References

- ASM. 1984. *ASM Metals Reference Book*, 2nd ed. American Society for Metals, Metals Park, OH.
- ASM. 1988. *Engineered Materials Handbook, Volume 2, Engineering Plastics*. ASM International, Metals Park, OH.
- Bolz, R. E. and Tuve, G. L. (Eds.) 1973. *CRC Handbook of Tables for Applied Engineering Science*. CRC Press, Boca Raton, FL.
- Lynch, C. T. (Ed.) 1975. *CRC Handbook of Materials Science, Volume 3*. CRC Press, Boca Raton, FL.
- Shackelford, J. F. 1992. *Introduction to Materials Science for Engineers*, 3rd ed. Macmillan, New York.
- Shackelford, J. F., Alexander, W., and Park, J. S. (Eds.) 1994. *The CRC Materials Science and Engineering Handbook*, 2nd ed. CRC Press, Boca Raton, FL.

## Further Information

A general introduction to the field of materials science and engineering is available from a variety of introductory textbooks. In addition to Shackelford [1992], readily available references include:

- Askeland, D. R. 1994. *The Science and Engineering of Materials*, 2nd ed. PWS-Kent, Boston.
- Callister, W. D. 1994. *Materials Science and Engineering: An Introduction*, 3rd ed. John Wiley & Sons, New York.
- Flinn, R. A., and Trojan, P. K. 1990. *Engineering Materials and Their Applications*, 4th ed. Houghton Mifflin, Boston.
- Smith, W. F. 1990. *Principles of Materials Science and Engineering*, 2nd ed. McGraw-Hill, New York.
- Van Vlack, L. H. 1989. *Elements of Materials Science and Engineering*, 6th ed. Addison-Wesley, Reading, MA.

As noted earlier, *The CRC Materials Science and Engineering Handbook* [Shackelford *et al.*, 1994] is available as a comprehensive source of property data for engineering materials. In addition, ASM International has published between 1982 and 1992 the *ASM Handbook*, an 18-volume set concentrating on metals and alloys. ASM International has also published a 4-volume set entitled the *Engineered Materials Handbook*, covering composites, engineering plastics, adhesives and sealants, and ceramics and glasses.

Shackelford, J. F. "Failure Analysis"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

[193.1 Types of Failures](#)[193.2 Failure Analysis Methodology](#)[193.3 Fracture Mechanics](#)[193.4 Nondestructive Testing](#)

Radiographic Testing • Ultrasonic Testing • Other Methods of Nondestructive Testing

[193.5 Engineering Design for Failure Prevention](#)**James F. Shackelford**

*University of California, Davis*

Failure analysis and prevention are important, practical considerations relative to the applications of materials in engineering design [ASM, 1986]. This chapter begins by exploring the various types of failures that can occur. There is now a well-established, systematic methodology for the analysis of failures of engineering materials. Beyond **failure analysis** as a "postmortem" examination, the related issue of **failure prevention** is equally important as the basis for avoiding future disasters. Fracture mechanics provides a powerful technique for both analyzing failures and modifying engineering designs for the prevention of future failures. Similarly, nondestructive testing serves as one set of techniques for a comprehensive analysis of failures, as well as a primary strategy for the prevention of future failures, especially in the identification of critical flaws. The two primary techniques on which we will focus are radiographic testing (typically using X rays) and ultrasonic testing.

The issues of failure analysis and failure prevention are taking on increasing levels of importance with the growing awareness of the responsibilities of the professional engineer. Ethical and legal demands for the understanding of engineering failures and the prevention of such future catastrophes are moving the field of materials science and engineering into a central role in the broader topic of engineering design.

---

## 193.1 Types of Failures

---

A wide spectrum of failure modes has been identified. To illustrate the range of these types of failures, we will begin with a brief description of several of the most common failures in structural metals.

*Ductile fracture* is observed in a large number of the failures occurring in metals due to overload—that is, simply taking a material beyond the elastic limit, beyond its ultimate tensile strength, and, subsequently, to fracture. The microscopic result of ductile fracture is a characteristic

"dimpled" fracture surface morphology.

*Brittle fracture* is characterized by rapid crack propagation without significant plastic deformation on a macroscopic scale. The cleavage texture of a brittle fracture surface comes from both *transgranular* fracture, involving the cleavage of microstructural grains, and *intergranular* fracture, occurring by crack propagation between adjacent grains.

*Fatigue failure* is the result of cyclic stresses that would, individually, be well below the yield strength of the material. The fatigue mechanism of slow crack growth gives a distinctive "clamshell" fatigue fracture surface.

*Corrosion-fatigue failure* is due to the combined actions of a cyclic stress and a corrosive environment. In general, the fatigue strength (or fatigue life at a given stress) of the metal will be decreased in the presence of an aggressive, chemical environment.

*Stress-corrosion cracking* (SCC) is another combined mechanical and chemical failure mechanism in which a noncyclic tensile stress (below the yield strength of the metal) leads to the initiation and propagation of fracture in a relatively mild chemical environment. Stress-corrosion cracks may be intergranular, transgranular, or a combination thereof.

*Wear failure* encompasses a broad range of relatively complex surface-related damage phenomena. Both surface damage and wear debris can constitute "failure" of materials intended for sliding contact applications.

*Liquid erosion failure* is a special form of wear damage in which a liquid, rather than another solid, is responsible for the removal of material. Liquid-erosion damage typically results in a pitted or honeycomb-like surface region.

*Liquid-metal embrittlement* is another form of material failure caused by a liquid. In this case the solid loses some degree of ductility or fractures below its yield stress in conjunction with its surface being wetted by a lower-melting-point liquid metal. This failure mode occurs in specific solid-liquid metal combinations.

*Hydrogen embrittlement* is perhaps the most notorious form of catastrophic failure in high-strength steels. A few parts per million of hydrogen dissolved in these materials can lead to fine "hairline" cracks and a loss of ductility. A variety of commercial environments serve as sources of hydrogen gas that, in turn, dissociates into atomic hydrogen, which can readily diffuse into the alloy.

*Creep and stress-rupture failures* can occur above about one-half the absolute melting point of an alloy. *Creep* is defined as plastic deformation over an extended period of time under a fixed load. Failure of this type can occur near room temperature for many polymers and certain low-melting-point metals, such as lead, but may occur above 1000°C in many ceramics and certain high-melting-point metals, such as the superalloys.

*Complex failures* are those in which the failure occurs by the sequential operation of two distinct fracture mechanisms. An example would be initial cracking due to stress-corrosion cracking and, then, ultimate failure by fatigue after a cyclic load is introduced simultaneously with the removal of the corrosive environment. The possibility of such sequences should always be considered when conducting a failure analysis.

## 193.2 Failure Analysis Methodology

---

A systematic sequence of procedures has been developed for the analysis of the failure of an engineering material. Although the specific methodology will vary with the specific failure, the principal components of the investigation and analysis are given in [Table 193.1](#). This specific set of components is that given in the *ASM Handbook* volume on failure analysis and prevention [[ASM, 1986](#)]. Because of the general utility of **nondestructive testing** for failure prevention as well as failure analysis, the subject is covered in detail later in this chapter. **Fracture mechanics** is the general analysis of the failure of structural materials with preexisting flaws. This subject will be discussed separately in the next section. In regard to failure analysis methodology, fracture mechanics has provided a quantitative framework for evaluating structural reliability.

**Table 193.1** Principal Components of Failure Analysis Methodology

---

Collection of background data and samples
Preliminary examination of the failed part
Nondestructive testing
Mechanical testing
Selection, preservation, and cleaning of fracture surfaces
Macroscopic (1 to 100 × ) examination of fracture surfaces
Microscopic (> 100 × ) examination of fracture surfaces
Application of fracture mechanics
Simulated-service testing
Analyzing the evidence, formulating conclusions, and writing the report

---

Adapted from ASM. 1986. *ASM Handbook, Volume 11: Failure Analysis and Prevention*. ASM International, Materials Park, OH.

## 193.3 Fracture Mechanics

---

The **fracture toughness**,  $K_{Ic}$ , is the most widely used material parameter associated with fracture mechanics [[Shackelford, 1992](#)] and is, in general, of the simple form

$$K_{Ic} = \sigma_f (\pi a)^{1/2} \quad (193.1)$$

where  $\sigma_f$  is the overall applied stress at failure and  $a$  is the length of a surface crack (or one-half the length of an internal crack). Fracture toughness has units of  $\text{MPa}\cdot\text{m}^{1/2}$ . Typical values of fracture toughness for various metals and alloys range between 20 and 200  $\text{MPa}\cdot\text{m}^{1/2}$ . Values of fracture toughness for ceramics and glass are typically in the range of 1 to 9  $\text{MPa}\cdot\text{m}^{1/2}$ , values for polymers are typically 1 to 4  $\text{MPa}\cdot\text{m}^{1/2}$ , and values for composites are typically 10 to 60  $\text{MPa}\cdot\text{m}^{1/2}$ . Although Eq. (193.1) is widely used for predicting flaw-induced or "fast" fracture, one must keep in mind that the subscript  $I$  refers to "mode  $I$ " loading—that is, simple, uniaxial tension. Fortunately, mode  $I$  conditions predominate in most practical systems, and the wide use of  $K_{Ic}$  is justified. Another aspect of fracture mechanics is the relationship to fatigue and other mechanical phenomena in which flaw size increases incrementally prior to failure.

## 193.4 Nondestructive Testing

As with fracture mechanics, nondestructive testing can serve to analyze an existing failure or be used to prevent future failures [Bray and Stanley, 1989]. The dominant techniques within this field are radiography and ultrasonics.

### Radiographic Testing

X rays compose a portion of the electromagnetic spectrum. Although diffraction techniques allow dimensions on the order of the X-ray wavelength (typically less than one nanometer) to be determined, X radiography produces a "shadow graph" of the internal structure of a part with a much coarser resolution (typically on the order of 1 mm). The common chest X ray is a routine example of this technology. In industrial applications, X radiography is widely used for the inspection of castings and weldments.

A key factor in this test is the thickness of material through which the X-ray beam can penetrate. For a given material being inspected by a given energy X-ray beam, the intensity of the beam,  $I$ , transmitted through a thickness of material,  $x$ , is given by Beer's law:

$$I = I_0 e^{-\mu x} \quad (193.2)$$

where  $I_0$  is the incident beam intensity and  $\mu$  is the linear absorption coefficient for the material. The intensity is proportional to the number of photons in the beam and should not be confused with the energy of photons in the beam. The absorption coefficient is a function of the beam energy and of the elemental composition of the material. Experimental values for the linear absorption coefficient of iron as a function of energy are given in Table 193.2. The general trend is a steady drop in the magnitude of  $\mu$  with increasing beam energy until above 1 MeV, where  $\mu$  levels off. Below 1 MeV the absorption of the beam is due to mechanisms of photon absorption and photon scattering. Above 1 MeV, however, an additional absorption mechanism comes into play (electron-positron pair production).

**Table 193.2** Linear Absorption Coefficient of Iron as a Function of X-Ray Beam Energy

Energy (MeV)	$\mu$ (mm <sup>-1</sup> )
0.05	1.52
0.10	0.293
0.50	0.0662
1.00	0.0417
2.00	0.0334
4.00	0.0260

Selected data from Bray, D. E. and Stanley, R. K. 1989. *Nondestructive Evaluation*. McGraw-Hill, New York.

The dependence of the linear absorption coefficient on elemental composition is illustrated by the data of Table 193.3. The general trend of data is that  $\mu$  for a given beam energy increases with

atomic number. As a result, low–atomic-number metals such as aluminum are relatively transparent and high–atomic-number metals such as lead are relatively opaque.

**Table 193.3** Linear Absorption Coefficient of Various Elements for an X-Ray Beam with Energy = 100 keV (= 0.1 MeV)

Element	Atomic Number	$\mu$ (mm <sup>-1</sup> )
Titanium	22	0.124
Iron	26	0.293
Nickel	28	0.396
Copper	29	0.410
Zinc	30	0.356
Tungsten	74	8.15
Lead	82	6.20

Selected data from Bray, D. E. and Stanley, R. K. 1989. *Nondestructive Evaluation*. McGraw-Hill, New York.

## Ultrasonic Testing

X radiography was seen to be based on a portion of the electromagnetic spectrum with relatively short wavelengths in comparison to the visible region. Similarly, ultrasonic testing is based on a portion of the acoustic spectrum with frequencies above those of the audible range (20 to 20 000 cycles/second or Hz). Typical ultrasonic inspections are made in the 1 to 25 MHz range. A key distinction between X radiography and ultrasonic testing is that the ultrasonic waves are mechanical in nature and require a transmitting medium, whereas electromagnetic waves can be transmitted in a vacuum.

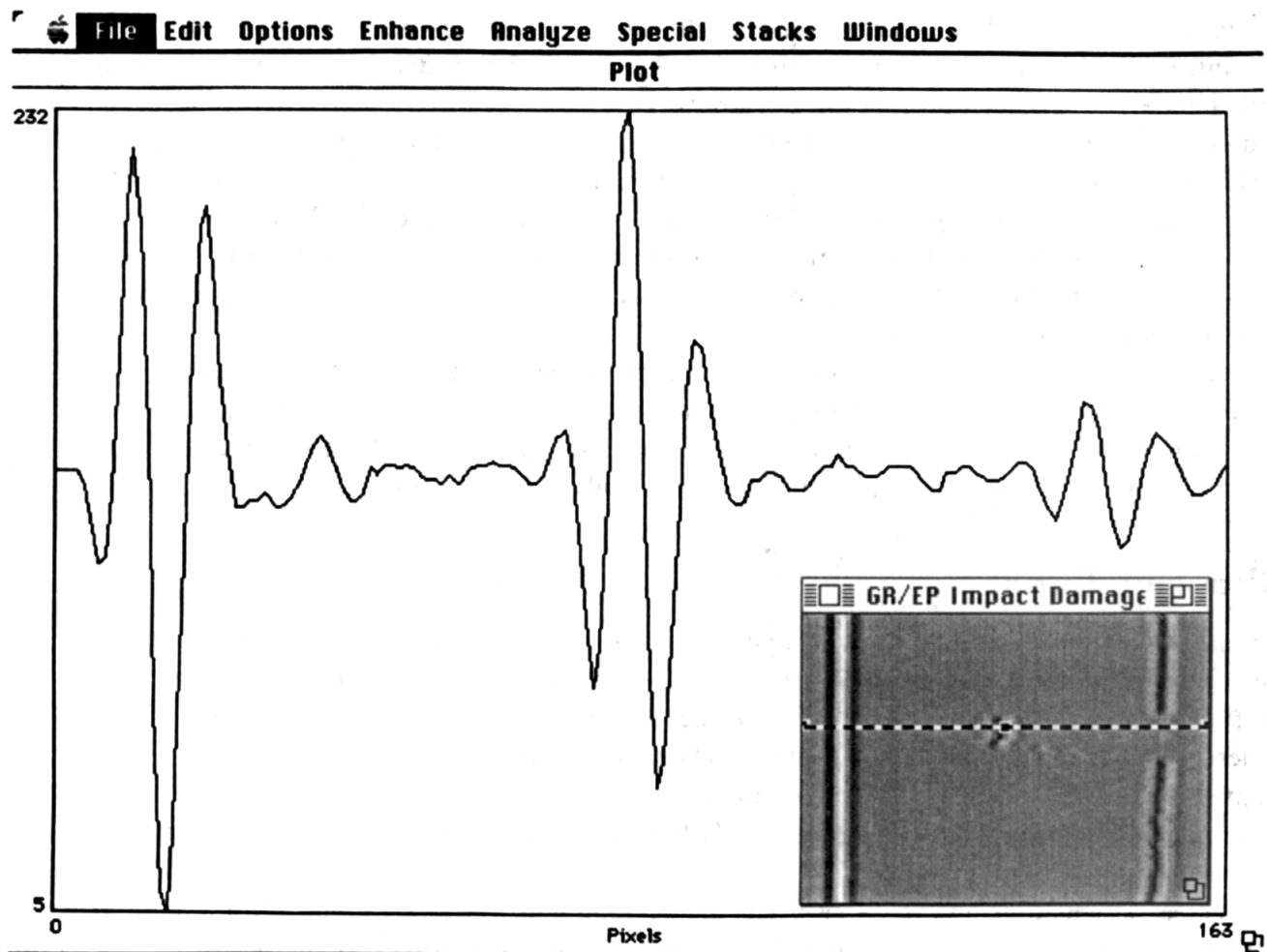
Whereas attenuation of the X-ray beam in the solid was a dominant consideration for X radiography, typical engineering materials are relatively transparent to ultrasonic waves. The key consideration for ultrasonic testing is the reflection of the ultrasonic waves at dissimilar material interfaces. The reflection coefficient,  $R$ , defined as the ratio of reflected beam intensity,  $I_r$ , to incident beam intensity,  $I_i$ , is given by

$$R = I_r/I_i = [(Z_2 - Z_1)/(Z_2 + Z_1)]^2 \quad (193.3)$$

where  $Z$  is the acoustic impedance, defined as the product of the material's density and velocity of sound, with the subscripts 1 and 2 referring to the two dissimilar materials on either side of the interface. The high degree of reflectivity by a typical flaw, such as an internal crack, is the basis for defect inspection. [Figure 193.1](#) illustrates a typical "pulse echo" ultrasonic inspection based on this principle. The oscillations in the figure represent voltage fluctuations in a piezoelectric transducer that is used to both send and detect the ultrasonic signal. The horizontal scale is proportional to time in a time frame of several microseconds. The initial pulse on the left is reflected back from a flaw and seen in the mid-range, with a small pulse on the right representing a reflection from the back side of the sample. The insert is a C-scan in which numerous adjacent A-scan pulses are viewed together to represent the spatial location of the flaw (impact damage in a NASA

graphite/epoxy wind tunnel blade). The dashed line corresponds to the displayed A-scan. The limitations of this method include the difficulty of applying the techniques in complex-shaped parts and the loss of information due to microstructural complexities such as porosity and precipitates.

**Figure 193.1** Typical ultrasonic pulse echo A-scan of a structural defect. (Ultrasonic imaging software provided courtesy of D. Bailey, McClellan Air Force Base.)



## Other Methods of Nondestructive Testing

A wide spectrum of additional methods are available in the field of nondestructive testing. The following four methods are among the most widely used for failure analysis and prevention.

*Eddy current testing* is a versatile technique for inspecting electrically conductive materials. The impedance of an inspection coil is affected by the presence of an adjacent test piece, in which alternating (eddy) currents have been induced by the coil. The net impedance is a function of the composition or geometry of the test piece. The popularity of eddy current testing is due to the convenient, rapid, and noncontact nature of the method. By varying the test frequency the method



can be used for both surface and subsurface flaws. Limitations include the qualitative nature and the need for an electrically conductive test piece.

*Magnetic-particle testing* is a simple, traditional technique widely used due to its convenience and low cost. Its primary limitation is the restriction to magnetic materials. (This is not as restrictive as one might initially assume, given the enormous volume of structural steels used in engineering.) The basic mechanism of this test involves the attraction of a fine powder of magnetic particles (Fe or  $\text{Fe}_3\text{O}_4$ ) to the "leakage flux" around a discontinuity such as a surface or near-surface crack.

*Liquid-penetrant testing* is, like magnetic-particle testing, an inexpensive and convenient technique for surface defect inspection. It is largely used on nonmagnetic materials for which magnetic-particle inspection is not possible. The basic mechanism for this method is the capillary action of a fine powder on the surface of a sample in which a high-visibility liquid has penetrated into surface defects. The limitations of this technique include the inability to inspect subsurface flaws and the loss of resolution on porous materials.

*Acoustic-emission testing* has assumed a unique role in failure prevention. This nondestructive test, in addition to being able to locate defects, can provide an early warning of impending failure due to those defects. In contrast to conventional ultrasonic testing, in which a transducer provides the source of ultrasound, acoustic-emission is the set of ultrasonic waves produced by defects within the microstructure of a material in response to an applied stress. With the material as the source of ultrasound, transducers serve only as receivers. In general, the rate of acoustic-emission events rises sharply just prior to failure. By continuously monitoring these emissions, the structural load can be removed in time to prevent that failure. The primary example of this preventative application is in the continuous surveillance of pressure vessels.

## 193.5 Engineering Design for Failure Prevention

---

Finally, it is important to note that engineering designs can be improved as a result of the application of concepts raised in sections 193.1 and 193.2—that is, failure analysis can lead to failure prevention. An example of this approach includes the avoidance of structural discontinuities that can serve as stress concentrators.

### Defining Terms

**Failure analysis:** The systematic examination of a failed engineering part to determine the nature of the failure.

**Failure prevention:** The application of the principles of failure analysis to prevent future catastrophes.

**Fracture mechanics:** Analysis of failure of structural materials with preexisting flaws.

**Fracture toughness:** Critical value of the stress-intensity factor at a crack tip necessary to produce catastrophic failure.

**Nondestructive testing:** A family of techniques for observing material flaws and characteristics without impairing the future usefulness of the part under inspection.

## References

- ASM. 1986. *ASM Handbook, Volume 11: Failure Analysis and Prevention*. ASM International, Materials Park, OH.
- Bray, D. E. and Stanley, R. K. 1989. *Nondestructive Evaluation*. McGraw-Hill, New York.
- Shackelford, J. F. 1992. *Introduction to Materials Science for Engineers*, 3rd ed. Macmillan, New York.

## Further Information

In addition to the references, readily available sources of further information include:

- ASM. 1987. *ASM Handbook, Volume 12: Fractography*. ASM International, Materials Park, OH.
- ASM. 1989. *ASM Handbook, Volume 17: Nondestructive Evaluation and Quality Control*. ASM International, Materials Park, OH.

Poling, B. E. "Liquids and Gases"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## 194.1 Viscosity

## 194.2 Thermal Conductivity

## 194.3 Heat Capacity

## 194.4 Vapor Pressure

**Bruce E. Poling***University of Toledo*

There are a number of thermodynamic and kinetic properties of gases and liquids that are important in engineering applications. Properties that are discussed in this chapter include viscosity, thermal conductivity, heat capacity, and vapor pressure. In the material that follows, a brief description of each of these four properties is given. However, engineers often do not want a description; rather, they want a number or an equation. The strategy that should be used to accomplish this latter task depends on the particular situation, but many features of this strategy are independent of the desired property. Thus, a general strategy is presented first.

For gases at low pressures and pure liquids, experimental data for many substances are available and have generally been correlated by equations and sometimes by nomographs. Many such correlations, summaries of data sources, and tabulations of values can be found in the references at the end of the chapter. The types of information available in the various references are summarized in Table 194.1. If information for a new compound or information at temperatures or pressures outside the range where experimental data have been measured is required, estimation methods such as those described in Reid *et al.* [1987] may be used. Reid *et al.* [1987] also presents recommended methods for finding properties for gases that cannot be considered ideal (at high pressure) and for mixtures of compounds.

**Table 194.1** Information Available in Various References

Reference Number <sup>a</sup>		1	2, 3	4	5	6, 7	8	9	10
Viscosity									
	Vapor		V	V,N					V
	Liquid	V	V	N	E				V
Thermal conductivity									
	Vapor	V	V	V	E				V
	Liquid	V	V	V	E				V
Heat capacity									
	Ideal Gas	V	V		E		E	V	V

	Liquid		V	V,N		E	V	
Vapor pressure		E	V	V	E	E,V	E	V

E = equation, V = values, N = nomograph

<sup>a</sup>1. Dean, 1992. 2. Lide, 1993. 3. Vargaftik, 1975. 4. Perry and Green, 1984. 5. Reid *et al.*, 1987. 6. Boublík *et al.*, 1984. 7. Shuzo, 1976. 8. Yaws, 1992. 9. Stull *et al.*, 1969. 10. Kaye *et al.*, 1986.

## 194.1 Viscosity

The **viscosity** of a fluid is a measure of its thickness, or how easily it flows. The viscosity,  $\eta$ , of a substance is defined as the shear stress per unit area divided by the velocity gradient. Fluids for which the viscosity is independent of shear stress are called *Newtonian fluids* and this class includes gases and most common liquids. Polymer solutions, inks, coatings and paints are often non-Newtonian—that is, their thickness and flow characteristics change with shear stress. Viscosity has dimensions of mass per length per time. A common viscosity unit is the poise, defined as one gram per centimeter per second. The kinematic viscosity is defined as the ratio of viscosity to density. A common unit for the kinematic viscosity is the stoke, defined as  $\text{cm}^2/\text{s}$ .

For pure liquids and pure gases at low pressure (near one atm) the viscosity is a function of temperature but is insensitive to changes in pressure. The viscosity of many common gases at low pressure can be found from the nomograph in Perry [1984]; for liquids, constants for 375 compounds appear in Reid *et al.* [1987]. One of the equations used in this tabulation is

$$\ln \eta = A + B/T + CT + DT^2 \quad (194.1)$$

where  $\eta$  is the viscosity in centipoise and  $T$  is the absolute temperature in kelvins. Constants to be used in Eq. (194.1) for liquid water are listed in Table 194.2.

**Table 194.2** Constants for Chapter Equations for Water

Property	Liquid Viscosity	Vapor Thermal Conductivity	Liquid Thermal Conductivity	Ideal Gas Heat Capacity	Vapor Pressure
Eq. in text	(194.1)	(194.2)	(194.3)	(194.5)	(194.9)
$A$	−24.71	$7.341 \cdot 10^{-3}$	−0.03838	32.24	−7.76451
$B$	4209	$−1.013 \cdot 10^{-5}$	$5.254 \cdot 10^{-3}$	0.001924	1.45838
$C$	0.04527	$1.801 \cdot 10^{-7}$	$−6.369 \cdot 10^{-6}$	$1.055 \cdot 10^{-5}$	−2.7758
$D$	$−3.376 \cdot 10^{-5}$	$−9.100 \cdot 10^{-11}$		$−3.596 \cdot 10^{-9}$	−1.23303
$T$ range, K	273–643	273–1070	273–623	273–1500	275–647.3

## 194.2 Thermal Conductivity

The **thermal conductivity** of a fluid is a measure of the rate at which heat is transferred through the fluid by conduction, that is, in the absence of convection. Units used for thermal conductivity

are  $W/(m\ K)$ . These may be converted to English or cgs units by

$$W/(m\ K) \times 0.5778 = \text{Btu}/(\text{h ft } ^\circ\text{R})$$

$$W/(m\ K) \times 0.8604 = \text{kcal}/(\text{cm h K})$$

Yaws and coworkers have used the following equation to correlate the thermal conductivity of 62 different gases:

$$\lambda = A + BT + CT^2 + DT^3 \quad (194.2)$$

and have correlated the thermal conductivity of liquids with the equation:

$$\lambda = A + BT + CT^2 \quad (194.3)$$

In both Eqs. (194.2) and (194.3),  $\lambda$  is the thermal conductivity in  $W/(m\ K)$  and  $T$  is in kelvins. Constants to be used for water vapor in Eq. (194.2) and liquid water in Eq. (194.3) appear in [Table 194.2](#). Constants for the other compounds are listed in both Miller *et al.* [1976] and Reid *et al.* [1987].

## 194.3 Heat Capacity

---

Generally speaking, the **heat capacity** of a fluid is the amount of heat required to increase a unit mass of the substance by one degree. Heat capacities are used to calculate sensible heat effects and are important in the design of heat exchangers. From a historic point of view, the heat capacity of water is of particular significance because, at one time, a calorie was defined as the amount of heat required to heat one gram of water one degree centigrade. The term *specific heat* is the ratio of the heat capacity of a substance to that of water. Although the definition of the calorie is now in terms of joules, and although the heat capacity of water varies slightly with temperature, for engineering purposes the heat capacity of liquid water may still be taken as  $1\ \text{cal}/(\text{g } ^\circ\text{C})$  over the range of 0 to  $100^\circ\text{C}$ . This is particularly convenient because the four sets of units— $\text{cal}/(\text{g } ^\circ\text{C})$ ,  $\text{cal}/(\text{g K})$ ,  $\text{Btu}/(\text{lb } ^\circ\text{F})$ , and  $\text{Btu}/(\text{lb } ^\circ\text{R})$ —are all numerically equivalent. Note that the temperature units in these four sets refer to temperature changes, so  $^\circ\text{C}$  and  $\text{K}$ , for example, are equivalent. Because the heat capacity of water is  $1\ \text{cal}/(\text{g } ^\circ\text{C})$ , the specific heat is numerically equal to the heat capacity when the units are  $\text{cal}/(\text{g } ^\circ\text{C})$ . For gases particularly, heat capacities are often given in molar units. Thus, if a heat capacity with units of  $\text{cal}/(\text{g } ^\circ\text{C})$  is multiplied by a substance's molecular weight, the molar heat capacity in  $\text{cal}/(\text{g mol } ^\circ\text{C})$  is obtained. The units,  $\text{cal}/(\text{g } ^\circ\text{C})$ , may be converted to  $\text{J}/(\text{g } ^\circ\text{C})$  by multiplying by 4.186.

From a rigorous thermodynamic point of view, the term *heat capacity* means heat capacity at constant pressure,  $C_p$ , and is the change in enthalpy with respect to temperature in a constant pressure experiment. However, other heat capacities can be defined. The heat capacity at constant volume,  $C_v$ , is the change in enthalpy with respect to temperature at constant volume, and for

liquids and solids these two heat capacities are essentially the same number. For ideal gases

$$C_p = C_v + R \quad (194.4)$$

Two other heat capacities that are sometimes seen are the change in enthalpy of a saturated liquid with respect to temperature (this path is at neither constant  $P$  nor  $V$  but, rather, along the vapor pressure curve) and the amount of energy required to effect a temperature change while maintaining the liquid in a saturated state. Fortunately, for liquids not close to the critical point and for ideal gases, these latter two heat capacities are essentially the same number as  $C_p$ .

For engineering applications, the heat capacities most often needed are **ideal gas heat capacities** and heat capacities of a condensed phase, that is, liquid or solid. The ideal gas heat capacity,  $C_p^o$ , for a monatomic gas in the ideal gas state, (examples of monatomic compounds include mercury, neon, helium, and argon) is 20.8 J/(g mol K). This is  $2.5R$ , where  $R$  is the gas constant. For compounds that contain more than one atom,  $C_p^o$  is a function of temperature (but not pressure) and its value generally increases with increasing temperature.  $C_p^o$  is correlated with temperature by equations such as

$$C_p^o = A + BT + CT^2 + DT^3 \quad (194.5)$$

and constants for many compounds are tabulated in textbooks, in Reid *et al.* [1987], and in Yaws [1992]. The constants to be used for water in the ideal gas state are listed in Table 194.2.  $T$  is in kelvins and  $C_p^o$  is in J/(g mol K).

The heat capacity of a mixture of ideal gases,  $C_{p,\text{mix}}^o$  may be calculated with the equation

$$C_{p,\text{mix}}^o = \sum_i x_i C_{p,i}^o \quad (194.6)$$

where  $x_i$  is the mole fraction of component  $i$  and  $C_{p,i}^o$  is for pure  $i$ .

For liquids, heat capacities for specific substances can be determined with the nomographs in Perry and Green [1984] or the constants in Yaws [1992]. Liquid heat capacities are weak functions of temperature below the normal boiling point. For some compounds  $C_p$  in this temperature range increases slowly with temperature, but for others  $C_p$  passes through a shallow minimum. The  $C_p$  behavior for water is an example of this latter behavior.  $C_p$  for water passes through a minimum at about 35°C; the values at 0 and 100°C are each about 1% higher than this minimum value. Above the normal boiling point,  $C_p$  for liquids begins to rise and for all compounds approaches infinity as the critical point is approached.

For gases in the nonideal state, and for liquids near the critical point, the heat capacity may be calculated by corresponding states correlations of the type

$$C_p = C_p^o + \Delta C_p^o \quad (194.7)$$

Values of  $\Delta C_p^o$  are functions of both temperature and pressure and may be obtained from tables in Reid *et al.* [1987].

## 194.4 Vapor Pressure

---

The **vapor pressure**, or saturation pressure, of a pure liquid is that pressure where both liquid and vapor are in equilibrium. The vapor pressure is a function only of temperature, and each substance has a unique vapor pressure curve. Three important points on this curve are the normal boiling point, the triple point, and the critical point. The normal boiling point corresponds to that temperature where the vapor pressure is one atm, and the triple point is that one point on the vapor pressure curve where vapor, liquid, and solid can exist in equilibrium. The triple point temperature is essentially the same number as the melting point temperature because pressure has little effect on melting points. At the critical point, the vapor and liquid phases become identical, and, above the critical temperature, the two phases are no longer distinct. The vapor pressure increases rapidly with temperature, and, in fact, the log of the vapor pressure varies nearly linearly with the reciprocal of the absolute temperature. Thus, if vapor pressures are known at two temperatures, reliable vapor pressures can be determined at other temperatures by interpolating on  $\log P_{vp}$  versus  $1/T$ , where  $P_{vp}$  is the vapor pressure and  $T$  is in kelvins. However, this linear relationship is not exact and extrapolations over large temperature ranges can sometimes lead to unacceptable errors.

A number of equations have been developed for correlating vapor pressures. Perhaps the most common is the Antoine equation:

$$\ln P_{vp} = A + \frac{B}{C + T} \quad (194.8)$$

More recently, the Wagner equation has generally been accepted as one of the best equations for correlating vapor pressures. This equation has the form

$$\ln(P_{vp}/P_c) = (A\tau + B\tau^{1.5} + C\tau^3 + D\tau^6)/T_r \quad (194.9)$$

In both Eqs. (194.8) and (194.9), the constants for a particular compound are determined by fitting vapor pressure data. In Eq. (194.9),  $P_c$  is the critical pressure,  $P_{vp}$  is the vapor pressure and will have the same units as  $P_c$ ,  $T_r$  is the reduced temperature,  $T/T_c$ ,  $\tau = 1 - T_r$ , and both  $T$  and  $T_c$  are in kelvins. Some authors have claimed that Eq. (194.9) is improved if the exponents 3 and 6 are replaced with 2.5 and 5. Constants for these equations are listed for 519 compounds in Reid *et al.* [1987]; constants to be used in Eq. (194.9) for water are listed in Table 194.2. Boublík *et al.* [1984] is an excellent source for experimental vapor pressure data. Vapor pressure values can also be found in Perry and Green [1984], Lide [1993], and Shuzo [1976].

### Defining Terms

**Heat capacity:** Generally, constant pressure heat capacity, which is defined as the change in enthalpy with respect to temperature at constant pressure.



**Ideal gas heat capacity:** The heat capacity of a substance at the specified temperature and in the ideal gas state. Unless otherwise stated, this generally means constant pressure heat capacity.

**Thermal conductivity:** The proportionality constant between the heat transfer rate (by conduction) per unit area and the temperature gradient; a measure of how fast heat is transferred by conduction through a substance.

**Vapor pressure:** The pressure at which both the liquid and vapor phases of a pure substance are in equilibrium.

**Viscosity:** The proportionality constant between the shear stress per unit area and the velocity gradient; a measure of the resistance to deformation.

## References

- Boublík, T. V., Fried, V., and Hála, E. 1984. *The Vapour Pressures of Pure Substances*, 2nd ed. Elsevier, New York.
- Dean, J. A. 1992. *Lange's Handbook of Chemistry*, 14th ed. McGraw-Hill, New York.
- Kaye, G. W. C. and Laby, T. H. 1986. *Tables of Physical and Chemical Constants*, 15th ed. Longman, New York.
- Lide, D. R. 1993. *CRC Handbook of Chemistry and Physics*, 74th ed. CRC Press, Boca Raton, FL.
- Miller, J. W., Jr., McGinley, J. J., and Yaws, C. L. 1976, *Chem. Eng.*, 83:133.
- Miller, J. W., Jr., Shah, P. N., and Yaws, C. L., 1976, *Chem. Eng.*, 83:153.
- Perry, R. H., and Green, D. 1984. *Perry's Chemical Engineers' Handbook*, 6th ed. McGraw-Hill, New York.
- Reid, R. C., Prausnitz, J. M., and Poling, B. E. 1987. *The Properties of Gases and Liquids*, 4th ed. McGraw-Hill, New York.
- Shuzo, O. 1976. *Computer Aided Data Book of Vapor Pressure*, Data Book, Tokyo.
- Stull, D. R., Westrum, E. F., Jr., and Sinke, G. C. 1969. *The Chemical Thermodynamics of Organic Compounds*, 2nd ed. John Wiley & Sons, New York.
- Vargaftik, N. B. 1975. *Tables on the Thermophysical Properties of Liquids and Gases*. John Wiley & Sons, New York.
- Yaws, C. L. 1992. *Thermodynamic and Physical Property Data*, Gulf, Houston, TX.

## Further Information

Introductory-level material on heat capacities and vapor pressures can generally be found in thermodynamics textbooks such as *Introduction to Chemical Engineering Thermodynamics* by J. M. Smith and H. C. Van Ness, 4th edition, published by McGraw-Hill in 1987. Introductory-level material on viscosities and thermal conductivities can be found in textbooks dealing with fluid flow and heat transfer, respectively. One such textbook is *Unit Operations of Chemical Engineering*, by W. L. McCabe, J. C. Smith, and P. Harriott, 5th edition, published by McGraw-Hill in 1993. As mentioned in the text, [Table 194.1](#) lists the type of information available in the books listed in the reference section.

Martin, R. B. "Biomaterials"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 195

## Biomaterials

---

- 195.1 History
- 195.2 Problems Associated with Implanted Devices
- 195.3 Immunology and Biocompatibility
- 195.4 Commonly Used Implant Materials
- 195.5 Metals
- 195.6 Polymers
- 195.7 Ceramics
- 195.8 Carbon Materials

### **R. Bruce Martin**

*University of California, Davis*

The term **biomaterials** usually refers to human-made materials used to construct prosthetic or other medical devices for implantation in a human being or materials that otherwise come in contact with the tissues of internal organs (e.g., tubing to carry blood to and from a heart-lung machine). Despite advances in many areas of materials science, it would be a mistake to assume that modern technology has the ability to replace any part of a living organism with an artificial material (or organ) that will be superior to the original structure. Although it is possible to imagine situations in which this might be true in some limited sense, the organism as a whole will rarely work better than when the original organ was in place. For example, a segment of bone may be replaced with a similar structure of titanium alloy having greater strength but lacking the bone's ability to adapt to changing loads and repair fatigue damage over time.

### **195.1 History**

---

The history of biomaterials can be divided into three eras. Prior to 1850 nonmetallic materials, such as wood and ivory, and common metals, such as iron, gold, silver, and copper, were used to fabricate simple prosthetic devices, such as teeth and noses, and to hold fractured bones together while they healed. In 1829 Levert experimented with lead, gold, silver, and platinum wire in dogs, but these metals clearly did not have the desired mechanical attributes. Furthermore, without anesthesia, human patients could not endure long surgeries in order to implant significant prostheses or fixation devices.

The second era of biomaterials was defined by the rapid development of surgery as something other than an emergency procedure, between 1850 and 1925. The advent of anesthesia just prior to the middle of the 19th century precipitated this development. Also, X rays were discovered by

Roentgen and found immediate application in orthopedics in the late 1800s, revealing for the first time the true nature of many skeletal problems. Finally, the acceptance of the aseptic surgical procedures propounded by Lister gradually but dramatically reduced the rate of postsurgical infections.

The period from 1925 to the present is the third era, in which the primary advances in the various surgical specialties have resulted from three important developments. The first was the development of cobalt chrome and stainless steel alloys in the 1930s and 1940s, respectively. The second was the development of polymer chemistry and plastics in the 1940s and 1950s. The third was the discovery of ways to produce useful quantities of penicillin and other antibiotics. The ability to further reduce surgical infection rates and to fabricate many devices that were compatible with biological tissues significantly advanced the ability of surgeons to treat a great variety of problems. Most of the biomaterials commonly in use today were developed more than 25 years ago; the intervening years have been ones of gradual refinement.

## **195.2 Problems Associated with Implanted Devices**

---

There are many problems that a biomaterial should ideally overcome. It must be formable into the desired shape using economical methods and without compromising its mechanical properties. It must not corrode in the presence of body fluids. This property entails the avoidance of crevice and fretting corrosion. A biomaterial must not poison the patient; therefore, it must either be free of toxic substances, or those substances must be adequately locked into the material's structure. A biomaterial must also be easy to sterilize (using steam under pressure, radiation, or ethylene oxide gas) without damaging its properties. The material must not break, either due to an occasional acute overload or due to fatigue from repeated functional loads. The strength and fatigue properties must combine with the shape of the implant to keep stresses within safe limits, particularly where stress concentrations cannot be avoided. Usually, the material must not simply work for a year or two, but for many years. Toxic ions that may be gradually released must not accumulate or lead to a long-term immunological response. A **connective tissue** implant must not perturb the stresses in adjacent tissues into a state that prohibits normal repair of fatigue damage. Finally, implants must be made from materials that can be removed and replaced if they fail. Materials that are impervious to attack by the body may be difficult for the surgeon to remove as well.



### Charnley Total Hip Replacement

*R. B. Martin*

This radiograph is of a Charnley total hip replacement prosthesis implanted in a human patient. The metal femoral component would typically be made from cobalt-chrome alloy and grouted into the femur with polymethylmethacrylate. The latter was Sir John Charnley's innovation, borrowed from dentistry, which made this procedure practical. The hemispherical polyethylene acetabular cup can be seen as the less radiodense material surrounding the top of the femoral component. It has a metal backing which shows as a radio-opaque boundary between the cup and the pelvic bone. The tangle of wire to the left (lateral) side of the prosthesis serves to reattach the greater trochanter to the femur. This ridge of bone is the attachment site for the hip's abductor muscles, which needed to be moved aside to gain access to the joint using the approach standard at that time. It is easier for the body to repair bone than muscle, so it was better to cut the bone than to cut the muscles. Today, other surgical approaches are often used which obviate this problem. (Photo from Waugh, William: *John Charnley: The Man and the Hip*, Springer-Verlag, New York, 1990. With permission.)

## 195.3 Immunology and Biocompatibility

---

All organisms try to keep out foreign matter; if they fail in this, they work very hard either to destroy the invading object if its molecules look destructible or to isolate it by encapsulating it in fibrous tissue if it looks impregnable. The problem of detecting foreign material is very complex, and the problem of defeating an immune system that successfully copes with this task looms very large. Today, people pay surgeons to put foreign objects into them, but their bodies still stick to the old principle of "encapsulate or destroy." In some cases **fibrous encapsulation** does not have an adverse effect on an implant, but, as a general rule, biomaterials should be able to avoid the body's natural inclination to either encapsulate them or break them down. Since doing this by destroying the body's defenses (i.e., the immune system) is a poor tactic, the best way to proceed is to make the implant invisible to the host's chemical sensors. This is difficult to do, since most materials will be easily recognized as "outsiders" and will have a difficult time avoiding the consequences. One useful principle to remember is that materials with molecules that look like biological molecules will be more readily attacked by destructive cells. For example, nylon and polyethylene both have a core of carbon atoms with hydrogens attached along their lengths. The main difference between the two is that polyethylene has a CH<sub>3</sub> terminal group, whereas nylon has an NH<sub>2</sub> terminal group. Since proteins have the latter kind of terminus, nylon tends to be more susceptible than polyethylene to degradation by cells of the immune system.

On the other hand, it may be very difficult to induce the body to attach tissue to impregnable materials such as polyethylene. In fact, no polymers have been found that are both immune to degradation and amenable to tissue adhesion. Some other materials are very attractive to connective tissue cells; for example, hydroxyapatite is nearly identical to bone mineral, and will quickly become integrated with bone when implanted in the skeleton. Unfortunately, hydroxyapatite is not nearly as tough and strong as bone. The commonly used materials that are tough and strong—metals—cannot be destroyed by the immune system, so the body tends to encapsulate them in fibrous tissue. This frustrates attempts to rigidly attach them to bone.

## 195.4 Commonly Used Implant Materials

---

Human-made biomaterials may be broadly categorized into metals, polymers, ceramics, and carbon composites. Each of these categories may contain materials in several forms, such as solids, membranes, fibers, or coatings, and serve many purposes, including replacing structural organs or organs that carry out chemical exchanges, housing electronic devices, repairing damaged or congenitally defective cardiovascular or connective tissue, or delivering drugs.

[Tables 195.1](#) and [195.2](#) show mechanical properties of metals and polymers commonly used as biomaterials. [Table 195.3](#) shows the properties of some connective tissues that these materials are designed to replace.

**Table 195.1** Mechanical Properties of Metals and Ceramics Commonly Used in Biomedical Implants<sup>a</sup>

	Elastic Modulus (GPa)	Maximum Strain (%)	Tensile Strength (MPa)
Stainless steel	193	10	1000
Cast cobalt-Cr	235	8	670
Wrought cobalt-Cr	235	12	1170
Ti-6Al-4V alloy	117	10	900
Pure titanium	100	15	550
Al <sub>2</sub> O <sub>3</sub> <sup>b</sup>	380	0	50
Apatite	62	0	690
C-Si composite	21	0	690

<sup>a</sup>Values in Tables 195.1 through 195.4 are approximations compiled from the references listed at the end of this chapter. Given the variability of biological tissues, and variations in manufacturing processes, precise values cannot be given for mechanical properties.

<sup>b</sup>Compressive strength = 4 GPa.

**Table 195.2** Mechanical Properties of Polymers Commonly Used in Biomedical Implants

Material	Elastic Modulus (MPa)	Maximum Strain (%)	Tensile Strength (MPa)
Silicone rubber <sup>a</sup>	2.4	700	—
Polyether urethane	—	700	41
Bion polymer <sup>a</sup>	1.5	350	13
UHMWPE	500	350	35
PMMA	2000	2	30 <sup>c</sup>
TCF <sup>b</sup>	20000	—	250

<sup>a</sup>Modulus at 100% strain.

<sup>b</sup>Carbon fiber–reinforced traizin resin.

<sup>c</sup>Compressive strength = 90 MPa.

**Table 195.3** Mechanical Properties of Major Connective Tissues

Material	Tensile Strength (MPa)	Elastic Modulus (MPa)	Extensibility (%)
Bone	150	20000	1.5
Cartilage (costal)	1	14	8.0
Collagen (tendon)	75	1300	9.0
Keratin <sup>a</sup>	50	5000	2.0

<sup>a</sup>For the alpha-form region of tensile tests of wool.

## 195.5 Metals

*Cobalt-chromium alloys* were the first corrosion resistant alloys to be developed and have proven very effective in surgical implants, beginning in 1936 when Venable reported the use of such an alloy in orthopedics. They are usually regarded as the metal of choice in skeletal prostheses. They have historically been available in cast and wrought forms; as with stainless steel, the wrought material is substantially stronger. The mechanical properties and compositions of each of the alloys

discussed in this section are shown in [Tables 195.1](#) and [195.4](#), respectively. A modification of Co-Cr alloy contains 35% nickel. It can be forged and heat-treated to obtain tensile strengths significantly above those of stainless steel and Co-Cr alloy—as high as 1800 MPa.

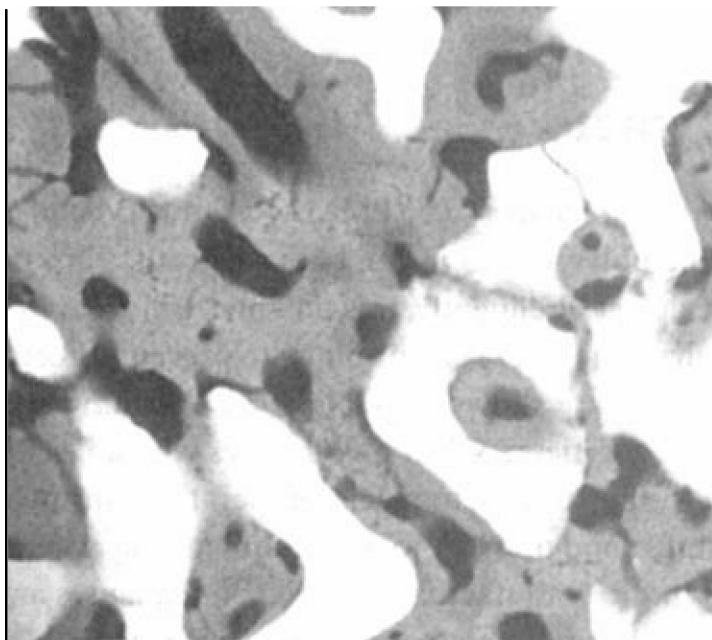
**Table 195.4** Approximate Compositions of Common Biomaterial Alloys

Alloy	Approximate composition
Cobalt-chrome	60% Co, 28% Cr, 6% Mo, 3% Ni, 3% other
Stainless steel	62% Fe, 18% Cr, 14% Ni, 3% Mo, 0.03% C, 3% other
Titanium	90% Ti, 6% Al, 4% V, 0.4% other

*Stainless steel* (usually 316L) is a workhorse industrial alloy that has been very successful as surgical implant material. Both stainless steel and cobalt-chrome alloys owe their corrosion resistance to the formation of a ceramiclike  $\text{CrO}_2$  coating on the surface, and it is important that this coating not be scratched during implantation. The ductility of these alloys can be increased by heat treatment, and their strength can be increased by cold working. Cast stainless steels are unsuitable for orthopedic applications because of their large grain sizes and low fatigue strengths. Type 316LVM (low carbon, vacuum melt) material is preferred.

*Titanium alloys* are primarily Ti-6Al-4V (i.e., 6% aluminum, 4% vanadium, and 90% titanium). (Pure titanium is also used, primarily in dental applications.) This metal is becoming increasingly popular in orthopedics because its strength is as good as the two previous choices, but it is only half as stiff ([Table 195.1](#)). This is potentially important because a large elastic modulus mismatch between implant and bone causes stress concentrations in some places and tends to **stress-shield** the bone in others. However, the modulus of Ti-6Al-4V is still several times greater than bone ([Tables 195.1](#) and [195.3](#)), and marked changes in stress shielding when this alloy is used have not yet been demonstrated.





**Microradiograph of Bone Material**

*R. B. Martin*

Microradiograph of bone (the gray material) which has grown into a porous coralline hydroxyapatite material (white) 16 weeks after implantation into a "window" defect in the radius of a dog. The hydroxyapatite material is used as a bone graft substitute to fill large defects created by trauma or tumor removal. It may be manufactured from coral by a thermochemical process which converts the calcium carbonate manufactured by the marine organism to calcium phosphate. Similar implants may be created by other means; the advantage of using coral is that it has an open porous structure which conveniently accommodates bone ingrowth. While hydroxyapatite is mechanically quite brittle and weak, it is the same mineral normally found in bone and is extremely "osteoconductive"—bone cells readily form new bone on its surfaces with intimate contact between the two materials. Ideally, it would be desirable for the hydroxyapatite to eventually be resorbed and entirely replaced by bone. There are currently insufficient data to know whether or not this will occur. For further information about the materials shown in this microradiograph, see the following:

Martin, R. B., Chapman, M. W., Holmes, R. E., Sartoris, D. J., Shors, E. C., Gordon, J. E., Heitter, D. O., Sharkey, N. A., and Zissimos, A. G. Effects of bone ingrowth on the strength and non-invasive assessment of a coralline hydroxyapatite material. *Biomaterials*. 10:481–488, 1989.

## 195.6 Polymers

*Polymethylmethacrylate* (PMMA) is an acrylic plastic from which parts may be machined. It also may be polymerized in the body to grout a hip prosthesis stem into the medullary canal or fill a bony defect. Frequently, this polymer will have an additive: (1) barium to increase its radiographic

visualization, or (2) an antibiotic to prevent infection following surgery. The mechanical properties of PMMA are shown in [Table 195.2](#); barium and antibiotic additions do not substantially affect these properties. PMMA polymerizes with an exothermic reaction that causes the mass of "dough" used in hip surgery to reach temperatures in the vicinity of 90°C. Also, PMMA has the effect of causing blood pressure to drop momentarily when it is implanted during hip surgery. PMMA is also used in the manufacture of hard contact lenses and of intraocular lenses for cataract patients. For this application, polymerization is initiated with heat to produce an extremely clear material with good optical properties. Soft contact lenses are usually hydrogels made of homo- or copolymers of hydroxyethyl methacrylate; other methacrylates and silicones have also been used, however.

*Ultrahigh-molecular-weight polyethylene* (UHMWPE) has a very simple chemical structure, consisting of chains of carbon atoms with hydrogen atoms attached to the sides. In the ultrahigh-molecular-weight form, these chains achieve a molecular weight of 1 to 4 million. When this material is used as a bearing surface against one of the metal alloys already mentioned, ordinary body fluids are usually sufficient to lubricate the artificial joint and reduce wear problems. Polyethylene has also been used to replace the ossicles of the inner ear.

*Polydimethylsiloxane* (silicone rubber) is a widely used polymer that was first applied biomedically in a hydrocephalus shunt in 1955. It has a long history of biological compatibility and clinical testing in a great variety of applications, and a "medical grade" (Silastic) has been developed with superior **biocompatibility** and mechanical properties. It can be sterilized by steam, radiation, or ethylene oxide. This polymer can be manufactured with various degrees of cross-linking to adapt its mechanical properties for other purposes as well. Silicone rubber is commonly used for catheters. Rubbery sheets of silicone have been used in hernia repair, and a gel silicone has often been used for breast prostheses or augmentation. (The recent controversy over the safety of the latter application is a good example of the difficulty in assessing the biocompatibility of biomaterials. Different individuals may respond quite differently to the same material.) Harder versions have served as balls in ball-and-cage heart valves. Silicone rubber is widely used in pacemakers and biomedical research to form seals against body fluids for electrical leads. It is used as the functional membrane in both kidney dialysis and extracorporeal blood oxygenator machines. It has also been used in drug delivery implants, where it serves as a membrane through which the drug slowly passes. Silicone also serves as a structural material in prosthetic heart valves and to replace the auricle or the ossicles of the ear. In orthopedics, silicone rubber is formed into soft, flexible "strap hinges" for the replacement of arthritic finger, wrist, or toe joints. However, its fatigue resistance is not as good as it should be for this application.

*Dacron* has been used for blood vessel prostheses since the pioneering work of DeBakey in 1951. Dacron is thrombogenic, so the pores in the fabric soon become filled with coagulated blood, which is then replaced by a tissue called *neointima*, which serves as a biological wall between the Dacron and the blood. However, if the vessel diameter is smaller than 6 mm, the neointima will occlude the tube. Dacron has also been used in prosthetic heart valves.

*Polytetrafluoroethylene* (PTFE, Teflon, or Gore-Tex) has been used experimentally to try to produce blood vessel prostheses smaller than 6 mm in diameter, and the neointima appears to be thinner with this material. However, it too is unsatisfactory for most human blood vessels, which are smaller than 3 mm. PTFE is also used to make prosthetic heart valves, ligaments, and artificial

ossicles for the ear.

*Polyether urethane* has been used for years in blood bags and tubing for kidney dialysis machines. It has also been tried as a material for blood vessel replacement. It is frequently found in intraaortic balloon pumps and in artificial hearts, where it lines the chambers and forms the pumping diaphragm. The primary requirement in the latter application is fatigue resistance; a prosthesis that is to last ten years must flex about 360 million times. Materials that have reportedly been tried for this purpose and found wanting include several other polymers used in tubing for kidney dialysis machines: polyvinylchloride, silicone rubber, and natural and synthetic rubbers.

*Polyalkylsulfones* are used in blood oxygenator membranes.

*Hexsyn* and a variation, *Bion*, are brand names of a recently developed elastomer that apparently has good biological compatibility and an extraordinarily high fatigue life (more than 300 million cycles to failure, compared to 600 000 for silicone rubber in an ASTM D430 flexure test). It has been tested in human finger joint prostheses.

*Epoxy resins* have been used to encapsulate electronic implants.

*Polydepsipeptides* and *polylactic acid* polymers have been tested with some success as biodegradable implants. Implant materials that disintegrate gradually in body fluids are useful as sutures and potentially useful as bone plates, to fill defects in bone or other tissues, and for the delivery of embedded drugs.

## 195.7 Ceramics

---

Ceramics generally have hydrophilic surfaces amenable to intimate bonding with tissues. They are very biocompatible but brittle relative to biological materials, including bone. Their primary application may ultimately be as a coating for metals to promote attachment to bone.

*Hydroxyapatite* [ $\text{Ca}_{10}(\text{PO}_4)_6(\text{OH})_2$ ] and *tricalcium phosphate* [TCP,  $\text{Ca}_3(\text{PO}_4)_2$ ] in various forms have proven to be very biocompatible and form intimate bonds with bone. Unfortunately, their tensile strengths are not great. They are primarily viewed as materials that may be made in porous or granular forms and used to fill bone defects in lieu of a bone graft. TCP may eventually be resorbed and replaced by bone; hydroxyapatite is less resorbable.

*Alumina*, a polycrystalline form of  $\text{Al}_2\text{O}_3$ , is a widely used industrial ceramic. It has relatively good strength characteristics but is very stiff and brittle compared to bone. It is sintered from alumina powder under pressure at  $1600^\circ\text{C}$  to form structures with relatively smooth surfaces. A principal use of this material is to form the heads of total hip prostheses (i.e., the ball of the "ball-and-socket" joint).

*Bioglass* ( $\text{Na}_2\text{O}-\text{CaO}-\text{P}_2\text{O}_5-\text{SiO}_2$ ) is a glass with soluble additives designed to form a silica gel at its surface and thereby aid chemical bonding to bone. To date, however, this material has not been widely accepted as an orthopedic biomaterial.

## 195.8 Carbon Materials

---

Carbon is the basis for organic chemistry. Thus, in principle, this element may serve as a good starting point for biocompatible materials. Graphite has a weak, anisotropic crystalline structure,

but isostatic (or turbostratic) carbon is isotropic and relatively strong. The two most useful forms of turbostratic carbon in biomedical applications are vitreous (or glassy) carbon and pyrolytic low-temperature isotropic (LTI) carbon. These are both isotropic materials, but the latter is more wear-resistant and stronger. Turbostratic carbons, carbon-fiber PMMA, and carbon-silicon composites have excellent biocompatibility and can have elastic moduli similar to bone. However, it typically turns out that in order to reduce the stiffness of these materials to that of bone, the strength must be reduced to less than that available in metals. In the last decade a number of these materials have been tested for various biomedical applications, but few have been widely used. Carbon-silicon and carbon-coated graphite have, however, been used to some extent in heart valves, and pyrolytic carbon coating has been shown to improve bone growth into porous metal surfaces.

## Defining Terms

**Biocompatibility:** The ability to remain in direct contact with the tissues of a living person without significantly affecting those tissues, or being affected by them, other than in a prescribed way. For example, a biocompatible material used in a blood vessel wall would contain the blood under pressure but would not adversely interact with its constituents or other adjacent tissues.

**Biomaterial:** An engineering material suitable for use in situations where it may come into direct contact with the internal tissues of the body. Usually, this means in a surgically implanted device, but it may also include devices that contain or process blood or other tissues that will be returned to the body. Sometimes, *biomaterials* is used to refer to biological structural materials.

**Connective tissue:** Tissue that primarily serves a mechanical rather than a metabolic or chemical function. The volume of the cells in such tissue is small relative to the extracellular components. For example, tendons are composed primarily of extracellular collagen, with a small fraction of their volume occupied by the cells that produce the collagen.

**Fibrous encapsulation:** The formation of connective tissues around a foreign object in order to isolate it from the rest of the body, both physically and chemically.

**Stress-shield:** To reduce the stress in a region of bone by placing it in parallel with an implant that is stiffer. Bone is able to adjust its structure to the applied loads, and if the imposed load is reduced, the affected bone will atrophy. This may leave the prosthesis inadequately supported and lead to failure.

## References

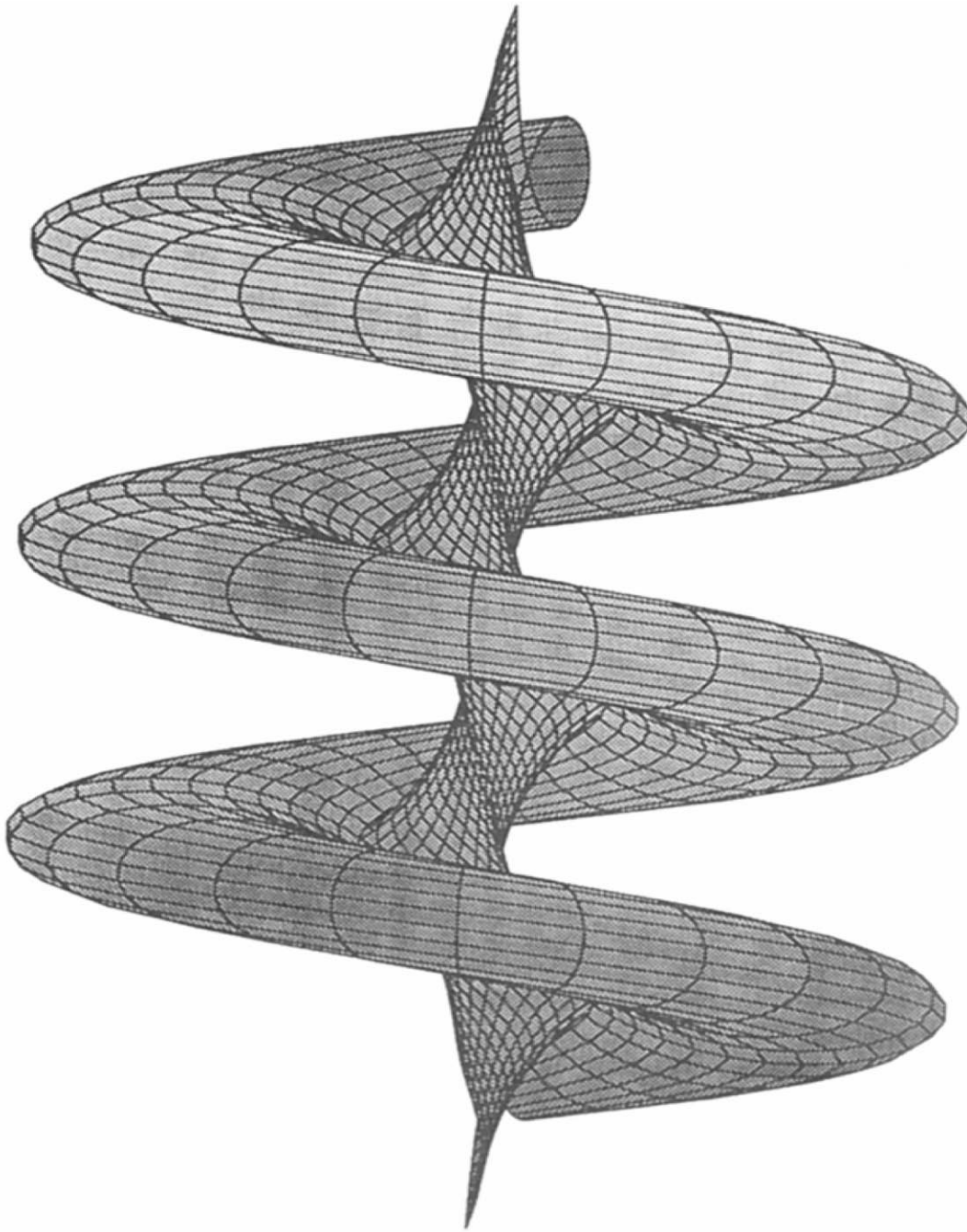
- Gebelein, C. G. 1984. *Polymeric Materials and Artificial Organs*. American Chemical Society, Washington, D.C.
- Hastings, G. W. and Ducheyne, P. 1984. *Natural and Living Biomaterials*. CRC Press, Boca Raton, FL.
- Hench, L. L. and Ethridge, E. C. 1982. *Biomaterials. An Interfacial Approach*. Academic Press, New York.

- Kossowsky, R. and Kossovsky, N. 1986. *Materials Sciences and Implant Orthopaedic Surgery*. Martinus Nijhoff, Boston, MA.
- Lin, O. C. C. and Chao, E. Y. S. 1985. Perspectives on biomaterials. *Proceedings of the 1985 International Symposium on Biomaterials*. Taipei, Taiwan, 25–27 February. Elsevier, Amsterdam.
- Nahum, A. M. and Melvin, J. 1985. *The Biomechanics of Trauma*. Appleton-Century-Crofts, Norwalk, CT.
- National Research Council. *Internal Structural Prostheses. Report of a Workshop on Fundamental Studies for Internal Structural Prostheses*. National Academy of Sciences, Washington, DC.

### **Further Information**

- Black, J. 1988. *Orthopaedic Biomaterials in Research and Practice*. Churchill Livingstone, New York.
- Kambik, H. E., and Toshimitsu, A. 1994. *Biomaterials' Mechanical Properties*. ASTM, Philadelphia, PA.
- Szycher, M. 1992. *High Performance Biomaterials: A Comprehensive Guide to Medical and Pharmaceutical Applications*. Technomic, Lancaster, PA.
- Szycher, M. 1992. *Szycher's Dictionary of Biomaterials and Medical Devices*. Technomic, Lancaster, PA.

Ames, W. F. "Mathematics"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000



This three-dimensional figure represents a single soliton surface for the nonlinear Schrödinger equation  $iq_t + q^2q/2 + q_{xx} = 0$ . The Schrödinger equation arises as a model equation in mathematical physics of atomic behavior. The nonlinear version shown here is one of the next steps in a hierarchy of theories.

This three-dimensional projection was generated using the MAPLE software package. MAPLE is one of three important mathematical computer packages which offer a variety of analytical and numerical software for use by scientists, engineers, and mathematicians. MAPLE was started in 1980 by a group of professors at the University of Waterloo in Canada. The latest version of MAPLE software is well used in the Western world and is strongly supported by the same people at the University of Waterloo.

This figure was developed by W. K. Schief and C. Rogers at the Center for Dynamical Systems and Nonlinear Studies at Georgia Institute of Technology and the University of New South Wales in Sydney, Australia. (Figure courtesy of Schief and Rogers.)

# XXX

## Mathematics

---

**William F. Ames**

*Georgia Institute of Technology*

**196 General Mathematics**

Trigonometry • Series • Differential Calculus • Integral Calculus • Special Functions

**197 Linear Algebra Matrices** *G. Cain*

Basic Definitions • Algebra of Matrices • Systems of Equations • Vector Spaces • Rank and Nullity • Orthogonality and Length • Determinants • Eigenvalues and Eigenvectors

**198 Vector Algebra and Calculus** *G. Cain*

Basic Definitions • Coordinate Systems • Vector Functions • Gradient, Curl, and Divergence • Integration • Integral Theorems

**199 Complex Variables** *G. Cain*

Basic Definitions and Arithmetic • Complex Functions • Analytic Functions • Integration • Series • Singularities • Conformal Mapping

**200 Difference Equations** *W. F. Ames*

First-Order Equations • Second-Order Equations • Linear Equations with Constant Coefficients • Generating Function ( $z$  Transform)

**201 Differential Equations** *W. F. Ames*

Ordinary Differential Equations • Partial Differential Equations

**202 Integral Equations** *W. F. Ames*

Classification and Notation • Relation to Differential Equations • Methods of Solution

**203 Approximation Methods** *W. F. Ames*

Perturbation • Iterative Methods

**204 Integral Transforms** *W. F. Ames*

Laplace Transform • Convolution Integral • Fourier Transform • Fourier Cosine Transform

**205 Chaos, Fractals, and Julia Sets** *A. Deliu*

Chaos • Fractals • Julia Sets

**206 Calculus of Variations** *W. F. Ames*

The Euler Equation • The Variation • Constraints

**207 Probability and Statistics** *Y. L. Tong*

Elementary Probability • Random Sample and Sampling Distributions • Normal Distribution-Related Sampling Distributions • Confidence Intervals • Testing Statistical Hypotheses • A Numerical Example

**208 Optimization** *G. Cain*

Linear Programming • Unconstrained Nonlinear Programming • Constrained Nonlinear Programming

**209 Numerical Methods** *W. F. Ames*

Linear Algebra Equations • Nonlinear Equations in One Variable • General Methods for Nonlinear



Equations in One Variable • Numerical Solution of Simultaneous Nonlinear Equations • Interpolation and Finite Differences • Numerical Differentiation • Numerical Integration • Numerical Solution of Ordinary Differential Equations • Numerical Solution of Integral Equations • Numerical Methods for Partial Differential Equations • Discrete and Fast Fourier Transform • Software

**210 Dimensional Analysis** *W. F. Ames*

Units and Variables • Method of Dimensions

**211 Computer Graphics Visualization** *R. S. Gallagher*

The Display of Objects in 3-D • Scalar Display Techniques • Vector and Tensor Field Display • Continuum Volume Visualization • Animation over Time • Summary

MATHEMATICS HAS BEEN DEFINED AS "the logic of drawing conclusions from arbitrary assumptions." And these conclusions remain valid forevermore! But I hasten to add that mathematics does not supply the assumptions. The engineer does that, but then the conclusions are unambiguous. These mathematical models of the physical world have been remarkably useful. Witness the plethora of equations of all sorts, such as the Navier-Stokes equations, the harmonic oscillator equation, the Korteweg de Vries equation, and so on. All have proven effective in predicting real-world phenomena. But we should not forget that they are not the phenomena, but only *models* of it. Further physical information often leads to improved mathematical models. And that is one of the reasons why a large mathematics section is included in this handbook. A second reason is to provide the reader with a readily available reference for basic and advanced mathematical ideas.

Beginning with a brief review of *analytic geometry* and *calculus*, both differential and integral, the advanced mathematical ideas concentrate on areas that have already been extremely useful. The properties of *vectors* and *matrices* are met first, including operators such as the gradient and curl. *Complex variables* are next in line because of their special nature in the solution of equations, matrix operations, and the solution of certain classes of differential equations.

Increasingly, *difference equations* are being employed to model discrete phenomena as well as in the numerical simulation of the *differential equations*, which are encountered in the following section. Of course, differential equations are the preeminent modeling tools in many fields of engineering. Less familiar are the *integral equations* which arise in certain optimization problems and in the numerical analysis of a variety of equations.

Since many equations are not solvable exactly, a variety of *approximate* (iteration, perturbation, weighted residual) and *numerical methods* have been developed. The latter section is the largest since the developed algorithms are useful on our essential computing machines—the new best friend of the engineer. A brief description of some computer software and packages is appended to this section. Other methods, to help in the solutions of problems, are given in the chapters on *integral transforms* and *optimization*. An important area in optimization is the *calculus of variations*, an old area that grew up with mechanics. *Dimensional analysis* still plays an important role, often offering a quick analysis of a new situation which needs *visualization*.

Of course, no engineer who needs to analyze data could do so without the ideas of *probability and statistics*. Theories of the physical world are continually evolving, so the areas of *chaos*, *fractals*, and *Julia sets* are included as a glimpse into the future. Finally, tables of various kinds are appended to aid the engineer in the search for information.

“General Mathematics”  
*The Engineering Handbook.*  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

**196.1 Trigonometry**

Triangles • Trigonometric Functions of an Angle • Trigonometric Identities • Inverse Trigonometric Functions

**196.2 Series**

Bernoulli and Euler Numbers • Series of Functions • Error Function • Series Expansion

**196.3 Differential Calculus**

Notation • Slope of a Curve • Angle of Intersection of Two Curves • Radius of Curvature • Relative Maxima and Minima • Points of Inflection of a Curve • Taylor's Formula • Indeterminant Forms • Numerical Methods • Functions of Two Variables • Partial Derivatives

**196.4 Integral Calculus**

Indefinite Integral • Definite Integral • Properties • Common Applications of the Definite Integral • Cylindrical and Spherical Coordinates • Double Integration • Surface Area and Volume by Double Integration • Centroid

**196.5 Special Functions**

Hyperbolic Functions • Bessel Functions • Legendre Polynomials • Laguerre Polynomials • Hermite Polynomials • Orthogonality • Functions with  $x^2/a^2 \pm y^2/b^2$  • Functions with  $(x^2/a^2 + y^2/b^2 \pm c^2)^{1/2}$

---

**196.1 Trigonometry**

---

**Triangles**

In any triangle (in a plane) with sides  $a$ ,  $b$ , and  $c$  and corresponding opposite angles  $A$ ,  $B$ , and  $C$ ,

$$\frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C} \quad (\text{Law of sines})$$

$$a^2 = b^2 + c^2 - 2bc \cos A \quad (\text{Law of cosines})$$

$$\frac{a+b}{a-b} = \frac{\tan \frac{1}{2}(A+B)}{\tan \frac{1}{2}(A-B)} \quad (\text{Law of Tangents})$$

$$\sin \frac{1}{2}A = \sqrt{\frac{(s-b)(s-c)}{bc}} \quad \text{where } s = \frac{1}{2}(a+b+c)$$

$$\cos \frac{1}{2}A = \sqrt{\frac{s(s-a)}{bc}}$$

$$\tan \frac{1}{2}A = \sqrt{\frac{(s-b)(s-c)}{s(s-a)}}$$

$$\begin{aligned}\text{Area} &= \frac{1}{2}bc \sin A \\ &= \sqrt{s(s-a)(s-b)(s-c)}\end{aligned}$$

If the vertices have coordinates  $(x_1, y_1)$ ,  $(x_2, y_2)$ ,  $(x_3, y_3)$ , the area is the *absolute value* of the expression

$$\frac{1}{2} \begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix}$$

## Trigonometric Functions of an Angle

With reference to [Fig. 196.1](#),  $P(x, y)$  is a point in any one of the four quadrants and  $A$  is an angle whose initial side is coincident with the positive  $x$  axis and whose terminal side contains the point  $P(x, y)$ . The distance from the origin  $P(x, y)$  is denoted by  $r$  and is positive. The trigonometric functions of the angle  $A$  are defined as:

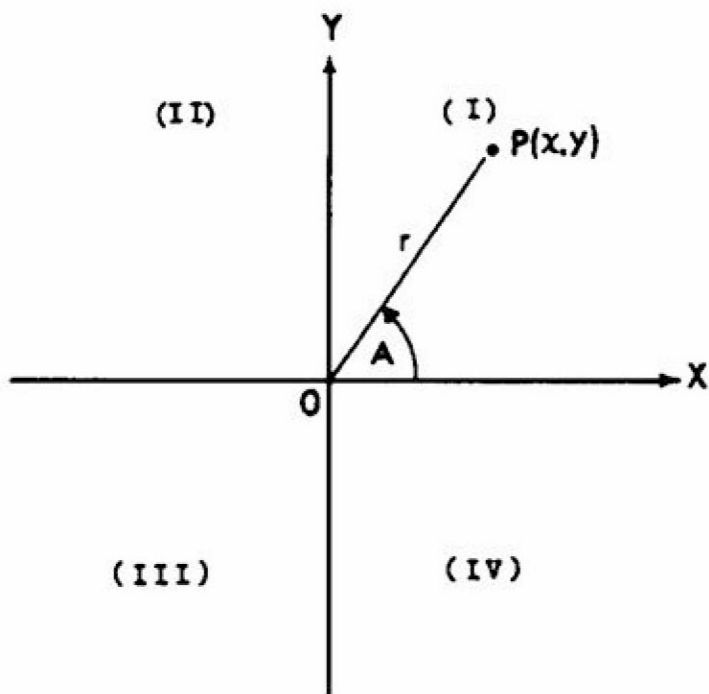
$$\begin{aligned}\sin A &= \text{sine } A &&= y/r \\ \cos A &= \text{cosine } A &&= x/r \\ \tan A &= \text{tangent } A &&= y/x \\ \text{ctn } A &= \text{cotangent } A &&= x/y \\ \sec A &= \text{secant } A &&= r/x \\ \csc A &= \text{cosecant } A &&= r/y\end{aligned}$$

Angles are measured in degrees or radians;  $180^\circ = \pi$  radians; 1 radian =  $180/\pi$  degrees. The trigonometric functions of  $0^\circ$ ,  $30^\circ$ ,  $45^\circ$ , and integer multiples of these are directly computed.

	$0^\circ$	$30^\circ$	$45^\circ$	$60^\circ$	$90^\circ$	$120^\circ$	$135^\circ$	$150^\circ$	$180^\circ$
sin	0	$\frac{1}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{3}}{2}$	1	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	0
cos	1	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	$-\frac{\sqrt{2}}{2}$	$-\frac{\sqrt{3}}{2}$	-1
tan	0	$\frac{\sqrt{3}}{3}$	1	$\sqrt{3}$	$\infty$	$-\sqrt{3}$	-1	$-\frac{\sqrt{3}}{3}$	0

ctn	$\infty$	$\sqrt{3}$	1	$\frac{\sqrt{3}}{3}$	0	$-\frac{\sqrt{3}}{3}$	-1	$-\sqrt{3}$	$\infty$
sec	1	$\frac{2\sqrt{3}}{3}$	$\sqrt{2}$	2	$\infty$	-2	$-\sqrt{2}$	$\frac{2\sqrt{3}}{3}$	-1
csc	$\infty$	$\frac{3}{2}$	$\sqrt{2}$	$\frac{2\sqrt{3}}{3}$	1	$\frac{2\sqrt{3}}{3}$	$\sqrt{2}$	$\frac{3}{2}$	$\infty$

**Figure 196.1** The trigonometric point. Angle  $A$  is taken to be positive when the rotation is counterclockwise and negative when the rotation is clockwise. The plane is divided into quadrants as shown.



## Trigonometric Identities

$$\sin A = \frac{1}{\csc A}$$

$$\cos A = \frac{1}{\sec A}$$

$$\tan A = \frac{1}{\cot A} = \frac{\sin A}{\cos A}$$

$$\csc A = \frac{1}{\sin A}$$

$$\sec A = \frac{1}{\cos A}$$

$$\cot A = \frac{1}{\tan A} = \frac{\cos A}{\sin A}$$

$$\sin^2 A + \cos^2 A = 1$$

$$1 + \tan^2 A = \sec^2 A$$

$$1 + \cot^2 A = \csc^2 A$$

$$\sin(A \pm B) = \sin A \cos B \pm \cos A \sin B$$

$$\cos(A \pm B) = \cos A \cos B \mp \sin A \sin B$$

$$\tan(A \pm B) = \frac{\tan A \pm \tan B}{1 \mp \tan A \tan B}$$

$$\sin 2A = 2 \sin A \cos A$$

$$\sin 3A = 3 \sin A - 4 \sin^3 A$$

$$\sin nA = 2 \sin(n-1)A \cos A - \sin(n-2)A$$

$$\cos 2A = 2 \cos^2 A - 1 = 1 - 2 \sin^2 A$$

$$\cos 3A = 4 \cos^3 A - 3 \cos A$$

$$\cos nA = 2 \cos(n-1)A \cos A - \cos(n-2)A$$

$$\begin{aligned}\sin A + \sin B &= 2 \sin \frac{1}{2}(A+B) \cos \frac{1}{2}(A-B) \\ \sin A - \sin B &= 2 \cos \frac{1}{2}(A+B) \sin \frac{1}{2}(A-B) \\ \cos A + \cos B &= 2 \cos \frac{1}{2}(A+B) \cos \frac{1}{2}(A-B) \\ \cos A - \cos B &= -2 \sin \frac{1}{2}(A+B) \sin \frac{1}{2}(A-B)\end{aligned}$$

$$\begin{aligned}\tan A \pm \tan B &= \frac{\sin(A \pm B)}{\cos A \cos B} \\ \operatorname{ctn} A \pm \operatorname{ctn} B &= \pm \frac{\sin(A \pm B)}{\sin A \sin B}\end{aligned}$$

$$\begin{aligned}\sin A \sin B &= \frac{1}{2} \cos(A-B) - \frac{1}{2} \cos(A+B) \\ \cos A \cos B &= \frac{1}{2} \cos(A-B) + \frac{1}{2} \cos(A+B) \\ \sin A \cos B &= \frac{1}{2} \sin(A+B) + \frac{1}{2} \sin(A-B)\end{aligned}$$

$$\begin{aligned}\sin \frac{A}{2} &= \pm \sqrt{\frac{1 - \cos A}{2}} \\ \cos \frac{A}{2} &= \pm \sqrt{\frac{1 + \cos A}{2}} \\ \tan \frac{A}{2} &= \frac{1 - \cos A}{\sin A} = \frac{\sin A}{1 + \cos A} = \pm \sqrt{\frac{1 - \cos A}{1 + \cos A}}\end{aligned}$$

$$\begin{aligned}\sin^2 A &= \frac{1}{2}(1 - \cos 2A) \\ \cos^2 A &= \frac{1}{2}(1 + \cos 2A) \\ \sin^3 A &= \frac{1}{4}(3 \sin A - \sin 3A) \\ \cos^3 A &= \frac{1}{4}(\cos 3A + 3 \cos A)\end{aligned}$$

$$\sin ix = \frac{1}{2}i(e^x - e^{-x}) = i \sinh x$$

$$\cos ix = \frac{1}{2}(e^x + e^{-x}) = \cosh x$$

$$\tan ix = \frac{i(e^x - e^{-x})}{e^x + e^{-x}} = i \tanh x$$

$$e^{x+iy} = e^x(\cos y + i \sin y)$$

$$(\cos x \pm i \sin x)^n = \cos nx \pm i \sin nx$$

## Inverse Trigonometric Functions

The inverse trigonometric functions are multiple valued, and this should be taken into account in the use of the following formulas.

$$\begin{aligned}\sin^{-1} x &= \cos^{-1} \sqrt{1-x^2} \\ &= \tan^{-1} \frac{x}{\sqrt{1-x^2}} = \operatorname{ctn}^{-1} \frac{\sqrt{1-x^2}}{x} \\ &= \sec^{-1} \frac{1}{\sqrt{1-x^2}} = \operatorname{csc}^{-1} \frac{1}{x} \\ &= -\sin^{-1}(-x)\end{aligned}$$

$$\begin{aligned}\cos^{-1} x &= \sin^{-1} \sqrt{1-x^2} \\ &= \tan^{-1} \frac{\sqrt{1-x^2}}{x} = \operatorname{ctn}^{-1} \frac{x}{\sqrt{1-x^2}} \\ &= \sec^{-1} \frac{1}{x} = \operatorname{csc}^{-1} \frac{1}{\sqrt{1-x^2}} \\ &= \pi - \cos^{-1}(-x)\end{aligned}$$

$$\begin{aligned}\tan^{-1} x &= \operatorname{ctn}^{-1} \frac{1}{x} \\ &= \sin^{-1} \frac{x}{\sqrt{1+x^2}} = \cos^{-1} \frac{1}{\sqrt{1+x^2}} \\ &= \sec^{-1} \sqrt{1+x^2} = \operatorname{csc}^{-1} \frac{\sqrt{1+x^2}}{x}\end{aligned}$$



## 196.2 Series

---

### Bernoulli and Euler Numbers

A set of numbers,  $B_1, B_3, \dots, B_{2n-1}$  (Bernoulli numbers) and  $B_2, B_4, \dots, B_{2n}$  (Euler numbers) appear in the series expansions of many functions. A partial listing follows; these are computed from the following equations:

$$B_{2n} - \frac{2n(2n-1)}{2!} B_{2n-2} + \frac{2n(2n-1)(2n-2)(2n-3)}{4!} B_{2n-4} - \dots + (-1)^n = 0$$

and

$$\begin{aligned} \frac{2^{2n}(2^{2n}-1)}{2n} B_{2n-1} &= (2n-1) B_{2n-2} \\ &\quad - \frac{(2n-1)(2n-2)(2n-3)}{3!} B_{2n-4} + \dots + (-1)^{n-1} \end{aligned}$$

$$B_1 = 1/6 \qquad B_2 = 1$$

$$B_3 = 1/30 \qquad B_4 = 5$$

$$B_5 = 1/42 \qquad B_6 = 61$$

$$B_7 = 1/30 \qquad B_8 = 1385$$

$$B_9 = 5/66 \qquad B_{10} = 50\,521$$

$$B_{11} = 691/2730 \qquad B_{12} = 2\,702\,765$$

$$B_{13} = 7/6 \qquad B_{14} = 199\,360\,981$$

$$\vdots$$

$$\vdots$$

### Series of Functions

In the following, the interval of convergence is indicated; otherwise it is all  $x$ . Logarithms are to the base  $e$ . Bernoulli and Euler numbers ( $B_{2n-1}$  and  $B_{2n}$ ) appear in certain expressions.

$$\begin{aligned}
(a+x)^n &= a^n + na^{n-1}x \\
&+ \frac{n(n-1)}{2!}a^{n-2}x^2 + \frac{n(n-1)(n-2)}{3!}a^{n-3}x^3 + \dots \\
&+ \frac{n!}{(n-j)!j!}a^{n-j}x^j + \dots
\end{aligned}
\quad [x^2 < a^2]$$

$$(a-bx)^{-1} = \frac{1}{a} \left[ 1 + \frac{bx}{a} + \frac{b^2x^2}{a^2} + \frac{b^3x^3}{a^3} + \dots \right] \quad [b^2x^2 < a^2]$$

$$(1 \pm x)^n = 1 \pm nx + \frac{n(n-1)}{2!}x^2 \pm \frac{n(n-1)(n-2)x^3}{3!} + \dots \quad [x^2 < 1]$$

$$(1 \pm x)^{-n} = 1 \mp nx + \frac{n(n+1)}{2!}x^2 \mp \frac{n(n+1)(n+2)}{3!}x^3 + \dots \quad [x^2 < 1]$$

$$(1 \pm x)^{1/2} = 1 \pm \frac{1}{2}x - \frac{1}{2 \cdot 4}x^2 \pm \frac{1 \cdot 3}{2 \cdot 4 \cdot 6}x^3 - \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6 \cdot 8}x^4 \pm \dots \quad [x^2 < 1]$$

$$(1 \pm x)^{-1/2} = 1 \mp \frac{1}{2}x + \frac{1 \cdot 3}{2 \cdot 4}x^2 \mp \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6}x^3 + \frac{1 \cdot 3 \cdot 5 \cdot 7}{2 \cdot 4 \cdot 6 \cdot 8}x^4 \pm \dots \quad [x^2 < 1]$$

$$(1 \pm x^2)^{1/2} = 1 \pm \frac{1}{2}x^2 - \frac{x^4}{2 \cdot 4} \pm \frac{1 \cdot 3}{2 \cdot 4 \cdot 6}x^6 - \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6 \cdot 8}x^8 \pm \dots \quad [x^2 < 1]$$

$$(1 \pm x)^{-1} = 1 \mp x + x^2 \mp x^3 + x^4 \mp x^5 + \dots \quad [x^2 < 1]$$

$$(1 \pm x)^{-2} = 1 \mp 2x + 3x^2 \mp 4x^3 + 5x^4 \mp \dots \quad [x^2 < 1]$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots$$

$$e^{-x^2} = 1 - x^2 + \frac{x^4}{2!} - \frac{x^6}{3!} + \frac{x^8}{4!} - \dots$$

$$a^x = 1 + x \log a + \frac{(x \log a)^2}{2!} + \frac{(x \log a)^3}{3!} + \dots$$

$$\log x = (x-1) - \frac{1}{2}(x-1)^2 + \frac{1}{3}(x-1)^3 - \dots \quad [0 < x < 2]$$

$$\log x = \frac{x-1}{x} + \frac{1}{2} \left( \frac{x-1}{x} \right)^2 + \frac{1}{3} \left( \frac{x-1}{x} \right)^3 + \dots \quad \left[ x > \frac{1}{2} \right]$$

$$\begin{aligned}\log x &= 2 \left[ \left( \frac{x-1}{x+1} \right) + \frac{1}{3} \left( \frac{x-1}{x+1} \right)^3 + \frac{1}{5} \left( \frac{x-1}{x+1} \right)^5 + \cdots \right] & [x > 0] \\ \log(1+x) &= x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + \cdots & [x^2 < 1] \\ \log \left( \frac{1+x}{1-x} \right) &= 2 \left[ x + \frac{1}{3}x^3 + \frac{1}{5}x^5 + \frac{1}{7}x^7 + \cdots \right] & [x^2 < 1] \\ \log \left( \frac{x+1}{x-1} \right) &= 2 \left[ \frac{1}{x} + \frac{1}{3} \left( \frac{1}{x} \right)^3 + \frac{1}{5} \left( \frac{1}{x} \right)^5 + \cdots \right] & [x^2 > 1] \\ \sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \\ \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots \\ \tan x &= x + \frac{x^3}{3} + \frac{2x^5}{15} + \frac{17x^7}{315} \\ &\quad + \cdots + \frac{2^{2n}(2^{2n}-1)B_{2n-1}x^{2n-1}}{(2n)!} & \left[ x^2 < \frac{\pi^2}{4} \right] \\ \operatorname{ctn} x &= \frac{1}{x} - \frac{x}{3} - \frac{x^3}{45} - \frac{2x^5}{945} - \cdots - \frac{B_{2n-1}(2x)^{2n}}{(2n)!x} - \cdots & [x^2 < \pi^2] \\ \sec x &= 1 + \frac{x^2}{2!} + \frac{5x^4}{4!} + \frac{61x^6}{6!} + \cdots + \frac{B_{2n}x^{2n}}{(2n)!} + \cdots & \left[ x^2 < \frac{\pi^2}{4} \right] \\ \csc x &= \frac{1}{x} + \frac{x}{3!} + \frac{7x^3}{3 \cdot 5!} + \frac{31x^5}{3 \cdot 7!} + \cdots \\ &\quad + \frac{2(2^{2n+1}-1)}{(2n+2)!} B_{2n+1} x^{2n+1} + \cdots & [x^2 < \pi^2] \\ \sin^{-1} x &= x + \frac{x^3}{6} + \frac{(1 \cdot 3)x^5}{(2 \cdot 4)5} + \frac{(1 \cdot 3 \cdot 5)x^7}{(2 \cdot 4 \cdot 6)7} + \cdots & [x^2 < 1] \\ \tan^{-1} x &= x - \frac{1}{3}x^3 + \frac{1}{5}x^5 - \frac{1}{7}x^7 + \cdots & [x^2 < 1] \\ \sec^{-1} x &= \frac{\pi}{2} - \frac{1}{x} - \frac{1}{6x^3} - \frac{1 \cdot 3}{(2 \cdot 4)5x^5} - \frac{1 \cdot 3 \cdot 5}{(2 \cdot 4 \cdot 6)7x^7} - \cdots & [x^2 > 1] \\ \sinh x &= x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \cdots \\ \cosh x &= 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} + \frac{x^8}{8!} + \cdots\end{aligned}$$

$$\begin{aligned}\tanh x &= (2^2 - 1)2^2 B_1 \frac{x}{2!} - (2^4 - 1)2^4 B_3 \frac{x^3}{4!} \\ &\quad + (2^6 - 1)2^6 B_5 \frac{x^5}{6!} - \dots\end{aligned}\quad \left[ x^2 < \frac{\pi^2}{4} \right]$$

$$\operatorname{ctnh} x = \frac{1}{x} \left( 1 + \frac{2^2 B_1 x^2}{2!} - \frac{2^4 B_3 x^4}{4!} + \frac{2^6 B_5 x^6}{6!} - \dots \right) \quad [x^2 < \pi^2]$$

$$\operatorname{sech} x = 1 - \frac{B_2 x^2}{2!} + \frac{B_4 x^4}{4!} - \frac{B_6 x^6}{6!} + \dots \quad \left[ x^2 < \frac{\pi^2}{4} \right]$$

$$\operatorname{csch} x = \frac{1}{x} - (2 - 1)2B_1 \frac{x}{2!} + (2^3 - 1)2B_3 \frac{x^3}{4!} - \dots \quad [x^2 < \pi^2]$$

$$\sinh^{-1} x = x - \frac{1}{2} \frac{x^3}{3} + \frac{1 \cdot 3}{2 \cdot 4} \frac{x^5}{5} - \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6} \frac{x^7}{7} + \dots \quad [x^2 < 1]$$

$$\tanh^{-1} x = x + \frac{x^3}{3} + \frac{x^5}{5} + \frac{x^7}{7} + \dots \quad [x^2 < 1]$$

$$\operatorname{ctnh}^{-1} x = \frac{1}{x} + \frac{1}{3x^3} + \frac{1}{5x^5} + \dots \quad [x^2 > 1]$$

$$\operatorname{csch}^{-1} x = \frac{1}{x} - \frac{1}{2 \cdot 3x^3} + \frac{1 \cdot 3}{2 \cdot 4 \cdot 5x^5} - \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6 \cdot 7x^7} + \dots \quad [x^2 > 1]$$

$$\int_0^x e^{-t^2} dt = x - \frac{1}{3}x^3 + \frac{x^5}{5 \cdot 2!} - \frac{x^7}{7 \cdot 3!} + \dots$$

## Error Function

The following function, known as the error function,  $\operatorname{erf} x$ , arises frequently in applications:

$$\operatorname{erf} x = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

The integral cannot be represented in terms of a finite number of elementary functions; therefore, values of  $\operatorname{erf} x$  have been compiled in tables. The following is the series for  $\operatorname{erf} x$ :

$$\operatorname{erf} x = \frac{2}{\sqrt{\pi}} \left[ x - \frac{x^3}{3} + \frac{x^5}{5 \cdot 2!} - \frac{x^7}{7 \cdot 3!} + \dots \right]$$

There is a close relation between this function and the area under the standard normal curve. For evaluation it is convenient to use  $z$  instead of  $x$ ; then  $\operatorname{erf} z$  may be evaluated from the area  $F(z)$  by use of the relation

$$\operatorname{erf} z = 2F(\sqrt{2}z)$$

### Example

$$\operatorname{erf}(0.5) = 2F[(1.414)(0.5)] = 2F(0.707)$$

By interpolation,  $F(0.707) = 0.260$ ; thus,  $\operatorname{erf}(0.5) = 0.520$ .

## Series Expansion

The expression in parentheses following certain series indicates the region of convergence. If not otherwise indicated, it is understood that the series converges for all finite values of  $x$ .

### Binomial

$$\begin{aligned} (x + y)^n &= x^n + nx^{n-1}y + \frac{n(n-1)}{2!}x^{n-2}y^2 \\ &\quad + \frac{n(n-1)(n-2)}{3!}x^{n-3}y^3 + \cdots \end{aligned} \quad (y^2 < x^2)$$

$$(1 \pm x)^n = 1 \pm nx + \frac{n(n-1)x^2}{2!} \pm \frac{n(n-1)(n-2)x^3}{3!} + \cdots \quad (x^2 < 1)$$

$$(1 \pm x)^{-n} = 1 \mp nx + \frac{n(n+1)x^2}{2!} \mp \frac{n(n+1)(n+2)x^3}{3!} + \cdots \quad (x^2 < 1)$$

$$(1 \pm x)^{-1} = 1 \mp x + x^2 \mp x^3 + x^4 \mp x^5 + \cdots \quad (x^2 < 1)$$

$$(1 \pm x)^{-2} = 1 \mp 2x + 3x^2 \mp 4x^3 + 5x^4 \mp 6x^5 + \cdots \quad (x^2 < 1)$$

### Reversion of Series

Let a series be represented by

$$y = a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5 + a_6x^6 + \cdots \quad (a_1 \neq 0)$$

To find the coefficients of the series

$$x = A_1 y + A_2 y^2 + A_3 y^3 + A_4 y^4 + \dots$$

$$A_1 = \frac{1}{a_1} \quad A_2 = -\frac{a_2}{a_1^3} \quad A_3 = \frac{1}{a_1^5} (2a_2^2 - a_1 a_3)$$

$$A_4 = \frac{1}{a_1^7} (5a_1 a_2 a_3 - a_1^2 a_4 - 5a_2^3)$$

$$A_5 = \frac{1}{a_1^9} (6a_1^2 a_2 a_4 + 3a_1^2 a_3^2 + 14a_2^4 - a_1^3 a_5 - 21a_1 a_2^2 a_3)$$

$$A_6 = \frac{1}{a_1^{11}} (7a_1^3 a_2 a_5 + 7a_1^3 a_3 a_4 + 84a_1 a_2^3 a_3 - a_1^4 a_6 \\ - 28a_1^2 a_2^2 a_4 - 28a_1^2 a_2 a_3^2 - 42a_2^5)$$

$$A_7 = \frac{1}{a_1^{13}} (8a_1^4 a_2 a_6 + 8a_1^4 a_3 a_5 + 4a_1^4 a_4^2 + 120a_1^2 a_2^3 a_4 \\ + 180a_1^2 a_2^2 a_3^2 + 132a_2^6 - a_1^5 a_7 \\ - 36a_1^3 a_2^2 a_5 - 72a_1^3 a_2 a_3 a_4 \\ - 12a_1^3 a_3^3 - 330a_1 a_2^4 a_3)$$

## Taylor

1.

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \frac{(x-a)^3}{3!}f'''(a) \\ + \dots + \frac{(x-a)^n}{n!}f^{(n)}(a) + \dots \quad (\text{Taylor's series})$$

(Increment Form)

2.

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + \frac{h^3}{3!}f'''(x) + \dots \\ = f(h) + xf'(h) + \frac{x^2}{2!}f''(h) + \frac{x^3}{3!}f'''(h) + \dots$$

3. If  $f(x)$  is a function possessing derivatives of all orders throughout the interval  $a \leq X \leq b$ , then there is a value  $X$ , with  $a < X < b$ , such that

$$\begin{aligned}
f(b) &= f(a) + (b-a)f'(a) + \frac{(b-a)^2}{2!}f''(a) + \cdots \\
&\quad + \frac{(b-a)^{n-1}}{(n-1)!}f^{(n-1)}(a) + \frac{(b-a)^n}{n!}f^{(n)}(a) \\
f(a+h) &= f(a) + hf'(a) + \frac{h^2}{2!}f''(a) + \cdots + \frac{h^{n-1}}{(n-1)!}f^{(n-1)}(a) \\
&\quad + \frac{h^n}{n!}f^{(n)}(a+\theta h), \quad b = a+h, \quad 0 < \theta < 1
\end{aligned}$$

or

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \cdots + (x-a)^{n-1}\frac{f^{(n-1)}(a)}{(n-1)!} + R_n$$

where

$$R_n = \frac{f^{(n)}[a+\theta(x-a)]}{n!}(x-a)^n, \quad 0 < \theta < 1.$$

The above forms are known as Taylor's series with the remainder term.

4. Taylor's series for a function of two variables:

$$\begin{aligned}
\text{If } \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right) f(x,y) &= h\frac{\partial f(x,y)}{\partial x} + k\frac{\partial f(x,y)}{\partial y}; \\
\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^2 f(x,y) &= h^2\frac{\partial^2 f(x,y)}{\partial x^2} + 2hk\frac{\partial^2 f(x,y)}{\partial x\partial y} + k^2\frac{\partial^2 f(x,y)}{\partial y^2}
\end{aligned}$$

etc., and if

$$\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^n f(x,y) \Big|_{\substack{x=a \\ y=b}}$$

where the bar and subscripts mean that after differentiation we are to replace  $x$  by  $a$  and  $y$  by  $b$ ,

$$\begin{aligned}
f(a+h, b+k) &= f(a,b) + \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right) f(x,y) \Big|_{\substack{x=a \\ y=b}} + \cdots \\
&\quad + \frac{1}{n!} \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^n f(x,y) \Big|_{\substack{x=a \\ y=b}} + \cdots
\end{aligned}$$

## MacLaurin

$$f(x) = f(0) + xf'(0) + \frac{x^2}{2!}f''(0) + \frac{x^3}{3!}f'''(0) + \cdots + x^{n-1}\frac{f^{(n-1)}(0)}{(n-1)!} + R_n$$

where

$$R_n = \frac{x^n f^{(n)}(\theta x)}{n!}, \quad 0 < \theta < 1$$

## Exponential

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots \quad (\text{all real values of } x)$$

$$a^x = 1 + x \log_e a + \frac{(x \log_e a)^2}{2!} + \frac{(x \log_e a)^3}{3!} + \cdots$$

$$e^x = e^a \left[ 1 + (x - a) + \frac{(x - a)^2}{2!} + \frac{(x - a)^3}{3!} + \cdots \right]$$



## Logarithmic

$$\log_e x = \frac{x-1}{x} + \frac{1}{2} \left( \frac{x-1}{x} \right)^2 + \frac{1}{3} \left( \frac{x-1}{x} \right)^3 + \dots \quad (x > \tfrac{1}{2})$$

$$\log_e x = (x-1) - \tfrac{1}{2}(x-1)^2 + \tfrac{1}{3}(x-1)^3 - \dots \quad (2 \geq x > 0)$$

$$\log_e x = 2 \left[ \frac{x-1}{x+1} + \frac{1}{3} \left( \frac{x-1}{x+1} \right)^3 - \frac{1}{5} \left( \frac{x-1}{x+1} \right)^5 + \dots \right] \quad (x > 0)$$

$$\log_e(1+x) = x - \tfrac{1}{2}x^2 + \tfrac{1}{3}x^3 - \tfrac{1}{4}x^4 + \dots \quad (-1 < x \leq 1)$$

$$\log_e(n+1) - \log_e(n-1) = 2 \left[ \frac{1}{n} + \frac{1}{3n^3} + \frac{1}{5n^5} + \dots \right]$$

$$\log_e(a+x) = \log_e a + 2 \left[ \frac{x}{2a+x} + \frac{1}{3} \left( \frac{x}{2a+x} \right)^3 + \frac{1}{5} \left( \frac{x}{2a+x} \right)^5 + \dots \right] \\ (a > 0, -a < x < +\infty)$$

$$\log_e \frac{1+x}{1-x} = 2 \left[ x + \frac{x^3}{3} + \frac{x^5}{5} + \dots + \frac{x^{2n-1}}{2n-1} + \dots \right] \quad (-1 < x < 1)$$

$$\log_e x = \log_e a + \frac{(x-a)}{a} - \frac{(x-a)^2}{2a^2} + \frac{(x-a)^3}{3a^3} - \dots \\ (0 < x \leq 2a)$$

## Trigonometric

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots \quad (\text{all real values of } x)$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots \quad (\text{all real values of } x)$$

$$\tan x = x + \frac{x^3}{3} + \frac{2x^5}{15} + \frac{17x^7}{315} + \frac{62x^9}{2835} + \dots$$

$$+ \frac{(-1)^{n-1} 2^{2n} (2^{2n} - 1) B_{2n}}{(2n)!} x^{2n-1} + \dots$$

( $x^2 < \pi^2/4$ , and  $B_n$  represents the  $n$ th Bernoulli number )

$$\cot x = \frac{1}{x} - \frac{x}{3} - \frac{x^2}{45} - \frac{2x^5}{945} - \frac{x^7}{4725} - \dots$$

$$- \frac{(-1)^{n+1} 2^{2n}}{(2n)!} B_{2n} x^{2n-1} + \dots$$

( $x^2 < \pi^2$ , and  $B_n$  represents the  $n$ th Bernoulli number )

## 196.3 Differential Calculus

---

### Notation

For the following equations, the symbols  $f(x)$ ,  $g(x)$ , etc., represent functions of  $x$ . The value of a function  $f(x)$  at  $x = a$  is denoted  $f(a)$ . For the function  $y = f(x)$  the derivative of  $y$  with respect to  $x$  is denoted by one of the following:

$$\frac{dy}{dx}, \quad f'(x), \quad D_x y, \quad y'$$

Higher derivatives are as follows:

$$\frac{d^2 y}{dx^2} = \frac{d}{dx} \left( \frac{dy}{dx} \right) = \frac{d}{dx} f'(x) = f''(x)$$

$$\frac{d^3 y}{dx^3} = \frac{d}{dx} \left( \frac{d^2 y}{dx^2} \right) = \frac{d}{dx} f''(x) = f'''(x)$$

$\vdots$

and values of these at  $x = a$  are denoted  $f'(a)$ ,  $f'''(a)$ , and so on (see Table of Derivatives in Appendix).

### Slope of a Curve

The tangent line at point  $P(x,y)$  of the curve  $y = f(x)$  has a slope  $f'(x)$  provided that  $f'(x)$  exists at  $P$ .

The slope at  $P$  is defined to be that of the tangent line at  $P$ . The tangent line at  $P(x_1, y_1)$  is given by

$$y - y_1 = f'(x_1)(x - x_1)$$

The *normal line* to the curve at  $P(x_1, y_1)$  has slope  $-1/f'(x_1)$  and thus obeys the equation

$$y - y_1 = [-1/f'(x_1)](x - x_1)$$

(The slope of a vertical line is not defined.)

## Angle of Intersection of Two Curves

Two curves,  $y = f_1(x)$  and  $y = f_2(x)$ , that intersect at a point  $P(X, Y)$  where derivatives  $f'_1(X), f'_2(X)$  exist, have an angle ( $\alpha$ ) of intersection given by

$$\tan \alpha = \frac{f'_2(X) - f'_1(X)}{1 + f'_2(X) \cdot f'_1(X)}$$

If  $\tan \alpha > 0$ , then  $\alpha$  is the acute angle; if  $\tan \alpha < 0$ , then  $\alpha$  is the obtuse angle.

## Radius of Curvature

The radius of curvature  $R$  of the curve  $y = f(x)$  at the point  $P(x, y)$  is

$$R = \frac{\{1 + [f'(x)]^2\}^{3/2}}{f''(x)}$$

In polar coordinates  $(\theta, r)$  the corresponding formula is

$$R = \frac{\left[ r^2 + \left( \frac{dr}{d\theta} \right)^2 \right]^{3/2}}{r^2 + 2 \left( \frac{dr}{d\theta} \right)^2 - r \frac{d^2r}{d\theta^2}}$$

The *curvature*  $K$  is  $1/R$ .

## Relative Maxima and Minima

The function  $f$  has a relative maximum at  $x = a$  if  $f(a) \geq f(a+c)$  for all values of  $c$  (positive or negative) that are sufficiently near zero. The function  $f$  has a relative minimum at  $x = b$  if  $f(b) \leq f(b+c)$  for all values of  $c$  that are sufficiently close to zero. If the function  $f$  is defined on the

closed interval  $x_1 \leq x \leq x_2$  and has a relative maximum or minimum at  $x = a$ , where  $x_1 < a < x_2$ , and if the derivative  $f'(x)$  exists at  $x = a$ , then  $f'(a) = 0$ . It is noteworthy that a relative maximum or minimum may occur at a point where the derivative does not exist. Further, the derivative may vanish at a point that is neither a maximum nor a minimum for the function. Values of  $x$  for which  $f'(x) = 0$  are called "critical values." To determine whether a critical value of  $x$ , say  $x_c$ , is a relative maximum or minimum for the function at  $x_c$ , one may use the second derivative test:

1. If  $f''(x_c)$  is positive,  $f(x_c)$  is a minimum.
2. If  $f''(x_c)$  is negative,  $f(x_c)$  is a maximum.
3. If  $f''(x_c)$  is zero, no conclusion may be made.

The sign of the derivative as  $x$  advances through  $x_c$  may also be used as a test. If  $f'(x)$  changes from positive to zero to negative, then a maximum occurs at  $x_c$ , whereas a change in  $f'(x)$  from negative to zero to positive indicates a minimum. If  $f'(x)$  does not change sign as  $x$  advances through  $x_c$ , then the point is neither a maximum nor a minimum.

## Points of Inflection of a Curve

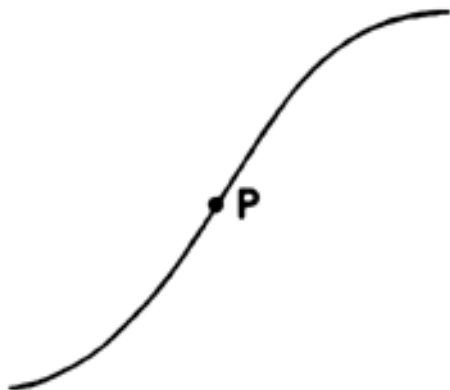
The sign of the second derivative of  $f$  indicates whether the graph of  $y = f(x)$  is concave upward or concave downward:

$f''(x) > 0$ : concave upward

$f''(x) < 0$ : concave downward

A point of the curve at which the direction of concavity changes is called a point of inflection (Fig. 196.2). Such a point may occur where  $f''(x) = 0$  or where  $f''(x)$  becomes infinite. More precisely, if the function  $y = f(x)$  and its first derivative  $y' = f'(x)$  are continuous in the interval  $a \leq x \leq b$ , and if  $y'' = f''(x)$  exists in  $a < x < b$ , then the graph of  $y = f(x)$  for  $a < x < b$  is concave upward if  $f''(x)$  is positive and concave downward if  $f''(x)$  is negative.

**Figure 196.2** Point of inflection.



## Taylor's Formula

If  $f$  is a function that is continuous on an interval that contains  $a$  and  $x$ , and if its first  $(n+1)$

derivatives are continuous on this interval, then

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 \\ + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R$$

where  $R$  is called the *remainder*. There are various common forms of the remainder:  
*Lagrange's Form*

$$R = f^{(n+1)}(\beta) \cdot \frac{(x-a)^{n+1}}{(n+1)!}, \quad \beta \text{ between } a \text{ and } x$$

*Cauchy's Form*

$$R = f^{(n+1)}(\beta) \cdot \frac{(x-a)^n(x-a)}{n!}, \quad \beta \text{ between } a \text{ and } x$$

*Integral Form*

$$R = \int_a^x \frac{(x-t)^n}{n!} f^{(n+1)}(t) dt$$

## Indeterminant Forms

If  $f(x)$  and  $g(x)$  are continuous in an interval that includes  $x = a$ , and if  $f(a) = 0$  and  $g(a) = 0$ , the limit  $\lim_{x \rightarrow a} [f(x)/g(x)]$  takes the form "0/0", called an *indeterminant form*. *L'Hôpital's rule* is

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = \lim_{x \rightarrow a} \frac{f'(x)}{g'(x)}$$

Similarly, it may be shown that if  $f(x) \rightarrow \infty$  and  $g(x) \rightarrow \infty$  as  $x \rightarrow a$ , then

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = \lim_{x \rightarrow a} \frac{f'(x)}{g'(x)}$$

(The above holds for  $x \rightarrow \infty$ .)

## Examples

$$\lim_{x \rightarrow 0} \frac{\sin x}{x} = \lim_{x \rightarrow 0} \frac{\cos x}{1} = 1$$

$$\lim_{x \rightarrow \infty} \frac{x^2}{e^x} = \lim_{x \rightarrow \infty} \frac{2x}{e^x} = \lim_{x \rightarrow \infty} \frac{2}{e^x} = 0$$

## Numerical Methods

1. *Newton's method* for approximating roots of the equation  $f(x) = 0$ : A first estimate  $x_1$  of the root is made; then, provided that  $f'(x_1) \neq 0$ , a better approximation is  $x_2$ :

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

The process may be repeated to yield a third approximation,  $x_3$ , to the root:

$$x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}$$

provided  $f'(x_2)$  exists. The process may be repeated. (In certain rare cases the process will not converge.)

2. *Trapezoidal rule for areas* (Fig. 196.3): For the function  $y = f(x)$  defined on the interval  $(a, b)$  and positive there, take  $n$  equal subintervals of width  $\Delta x = (b - a)/n$ . The area bounded by the curve between  $x = a$  and  $x = b$  [or definite integral of  $f(x)$ ] is approximately the sum of trapezoidal areas, or

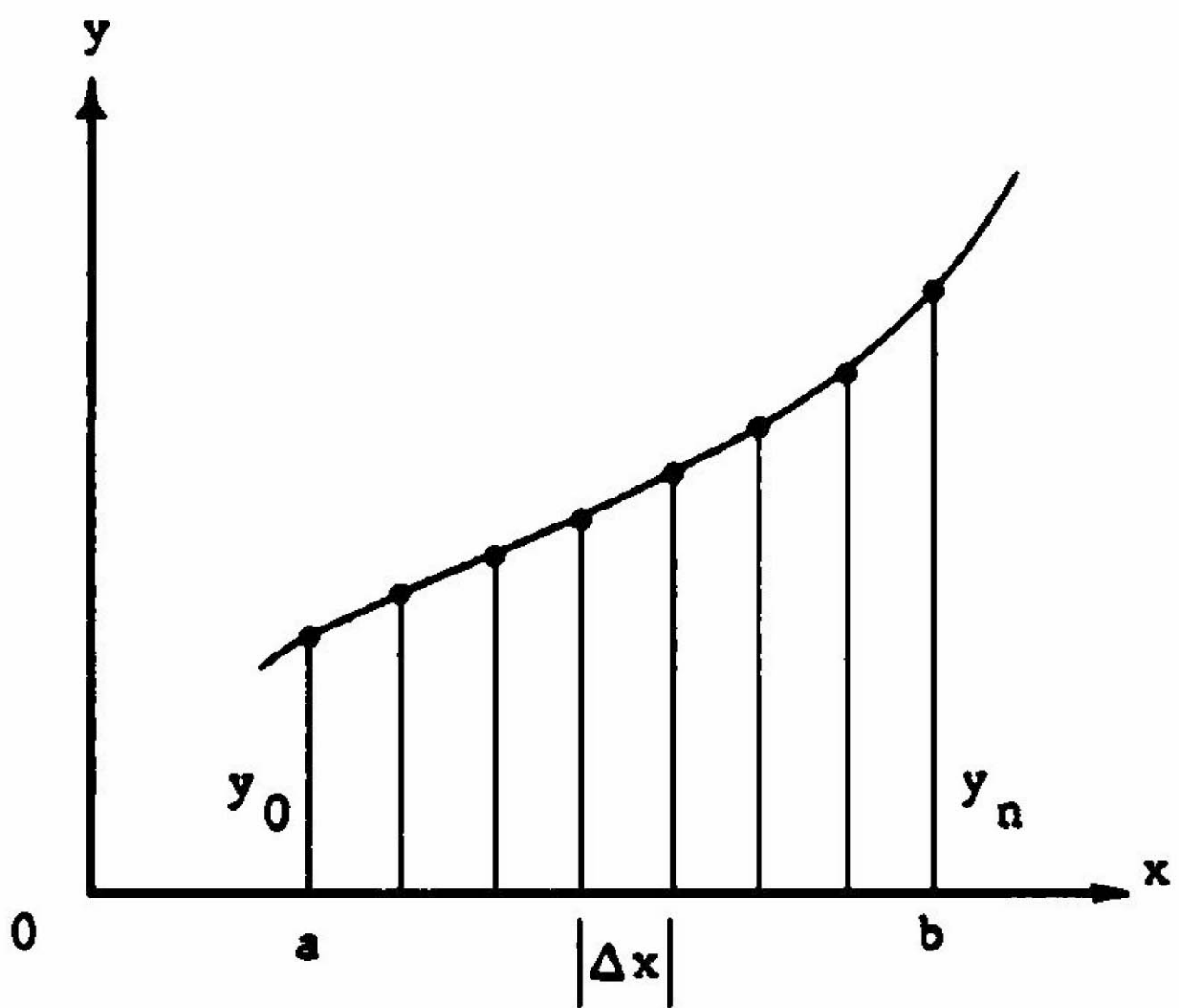
$$A \sim \left(\frac{1}{2}y_0 + y_1 + y_2 + \cdots + y_{n-1} + \frac{1}{2}y_n\right)(\Delta x)$$

Estimation of the error ( $E$ ) is possible if the second derivative can be obtained:

$$E = \frac{b-a}{12} f''(c)(\Delta x)^2$$

where  $c$  is some number between  $a$  and  $b$ .

**Figure 196.3** Trapezoidal rule for area.



## Functions of Two Variables

For the function of two variables, denoted  $z = f(x, y)$ , if  $y$  is held constant, say at  $y = y_1$ , then the resulting function is a function of  $x$  only. Similarly,  $x$  may be held constant at  $x_1$ , to give the resulting function of  $y$ .

### The Gas Laws

A familiar example is afforded by the ideal gas law relating the pressure  $p$ , the volume  $V$ , and the absolute temperature  $T$  of an ideal gas:

$$pV = nRT$$

where  $n$  is the number of moles and  $R$  is the gas constant per mole,  $8.31 \text{ (J} \cdot \text{K}^{-1} \cdot \text{mole}^{-1})$ . By rearrangement, any one of the three variables may be expressed as a function of the other two. Further, either one of these two may be held constant. If  $T$  is held constant, then we get the form known as Boyle's law:

$$p = kV^{-1} \quad (\text{Boyle's law})$$

where we have denoted  $nRT$  by the constant  $k$  and, of course,  $V > 0$ . If the pressure remains constant, we have Charles' law:

$$V = bT \quad (\text{Charles' law})$$

where the constant  $b$  denotes  $nR/p$ . Similarly, volume may be kept constant:

$$p = aT$$

where now the constant, denoted  $a$ , is  $nR/V$ .

## Partial Derivatives

The physical example afforded by the ideal gas law permits clear interpretations of processes in which one of the variables is held constant. More generally, we may consider a function  $z = f(x,y)$  defined over some region of the  $xy$  plane in which we hold one of the two coordinates, say  $y$ , constant. If the resulting function of  $x$  is differentiable at a point  $(x,y)$ , we denote this derivative by one of the notations

$$f_x, \quad \delta f/dx, \quad \delta z/dx$$

called the *partial derivative with respect to  $x$* . Similarly, if  $x$  is held constant and the resulting function of  $y$  is differentiable, we get the *partial derivative with respect to  $y$* , denoted by one of the following:

$$f_y, \quad \delta f/dy, \quad \delta z/dy$$

### Example.

Given  $z = x^4y^3 - y \sin x + 4y$ , then

$$\delta z/dx = 4(xy)^3 - y \cos x$$

$$\delta z/dy = 3x^4y^2 - \sin x + 4$$



## 196.4 Integral Calculus

---

### Indefinite Integral

If  $F(x)$  is differentiable for all values of  $x$  in the interval  $(a, b)$  and satisfies the equation  $dy/dx = f(x)$ , then  $F(x)$  is an integral of  $f(x)$  with respect to  $x$ . The notation is  $F(x) = \int f(x) dx$  or, in differential form,  $dF(x) = f(x) dx$ .

For any function  $F(x)$  that is an integral of  $f(x)$ , it follows that  $F(x) + C$  is also an integral. We thus write

$$\int f(x) dx = F(x) + C$$

### Definite Integral

Let  $f(x)$  be defined on the interval  $[a, b]$  which is partitioned by points  $x_1, x_2, \dots, x_j, \dots, x_{n-1}$  between  $a = x_0$  and  $b = x_n$ . The  $j$ th interval has length  $\Delta x_j = x_j - x_{j-1}$ , which may vary with  $j$ . The sum  $\sum_{j=1}^n f(v_j) \Delta x_j$ , where  $v_j$  is arbitrarily chosen in the  $j$ th subinterval, depends on the numbers  $x_0, \dots, x_n$  and the choice of the  $v$  as well as  $f$ ; but if such sums approach a common value as all  $\Delta x$  approach zero, then this value is the definite integral of  $f$  over the interval  $(a, b)$  and is denoted  $\int_a^b f(x) dx$ . The *fundamental theorem of integral calculus* states that

$$\int_a^b f(x) dx = F(b) - F(a),$$

where  $F$  is any continuous indefinite integral of  $f$  in the interval  $(a, b)$ .

### Properties

$$\int_a^b [f_1(x) + f_2(x) + \dots + f_j(x)] dx = \int_a^b f_1(x) dx + \int_a^b f_2(x) dx + \dots + \int_a^b f_j(x) dx$$

$$\int_a^b c f(x) dx = c \int_a^b f(x) dx, \quad \text{if } c \text{ is a constant}$$

$$\int_a^b f(x) dx = - \int_b^a f(x) dx$$

$$\int_a^b f(x) dx = \int_a^c f(x) dx + \int_c^b f(x) dx$$

## Common Applications of the Definite Integral

### Area (Rectangular Coordinates)

Given the function  $y = f(x)$  such that  $y > 0$  for all  $x$  between  $a$  and  $b$ , the area bounded by the curve  $y = f(x)$ , the  $x$  axis, and the vertical lines  $x = a$  and  $x = b$  is

$$A = \int_a^b f(x) dx$$

### Length of Arc (Rectangular Coordinates)

Given the smooth curve  $f(x, y) = 0$  from point  $(x_1, y_1)$  to point  $(x_2, y_2)$ , the length between these points is

$$L = \int_{x_1}^{x_2} \sqrt{1 + (dy/dx)^2} dx$$

$$L = \int_{y_1}^{y_2} \sqrt{1 + (dx/dy)^2} dy$$

### Mean Value of a Function

The mean value of a function  $f(x)$  continuous on  $[a, b]$  is

$$\frac{1}{(b-a)} \int_a^b f(x) dx$$

### Area (Polar Coordinates)

Given the curve  $r = f(\theta)$ , continuous and nonnegative for  $\theta_1 \leq \theta \leq \theta_2$ , the area enclosed by this curve and the radial lines  $\theta = \theta_1$  and  $\theta = \theta_2$  is given by

$$A = \int_{\theta_1}^{\theta_2} \frac{1}{2} [f(\theta)]^2 d\theta$$

### Length of Arc (Polar Coordinates)

Given the curve  $r = f(\theta)$  with continuous derivative  $f'(\theta)$  on  $\theta_1 \leq \theta \leq \theta_2$ , the length of arc from  $\theta = \theta_1$  to  $\theta = \theta_2$  is

$$L = \int_{\theta_1}^{\theta_2} \sqrt{[f(\theta)]^2 + [f'(\theta)]^2} d\theta$$

### Volume of Revolution

Given a function  $y = f(x)$  continuous and nonnegative on the interval  $(a, b)$ , when the region bounded by  $f(x)$  between  $a$  and  $b$  is revolved about the  $x$  axis, the volume of revolution is

$$V = \pi \int_a^b [f(x)]^2 dx$$

### Surface Area of Revolution(Revolution about the $x$ axis, between $a$ and $b$ )

If the portion of the curve  $y = f(x)$  between  $x = a$  and  $x = b$  is revolved about the  $x$  axis, the area  $A$  of the surface generated is given by the following:

$$A = \int_a^b 2\pi f(x) \{1 + [f'(x)]^2\}^{1/2} dx$$

### Work

If a variable force  $f(x)$  is applied to an object in the direction of motion along the  $x$  axis between  $x = a$  and  $x = b$ , the work done is

$$W = \int_a^b f(x) dx$$

## Cylindrical and Spherical Coordinates

1. Cylindrical coordinates (Fig. 196.4):

$$x = r \cos \theta$$

$$y = r \sin \theta$$

Element of volume  $dV = r dr d\theta dz$  .

2. Spherical coordinates (Fig. 196.5):

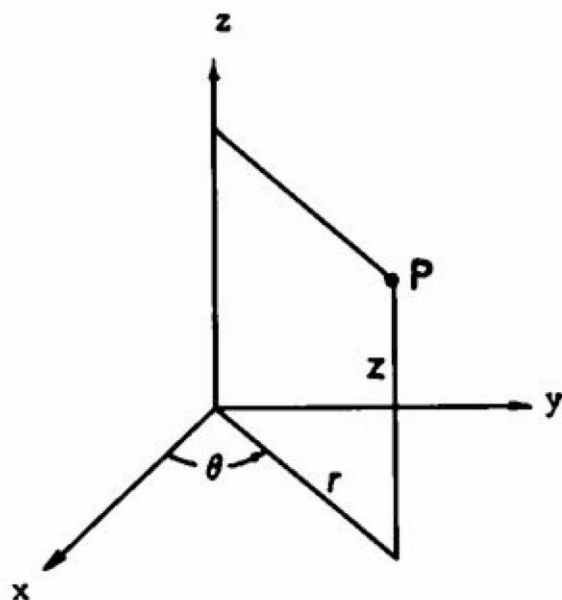
$$x = \rho \sin \phi \cos \theta$$

$$y = \rho \sin \phi \sin \theta$$

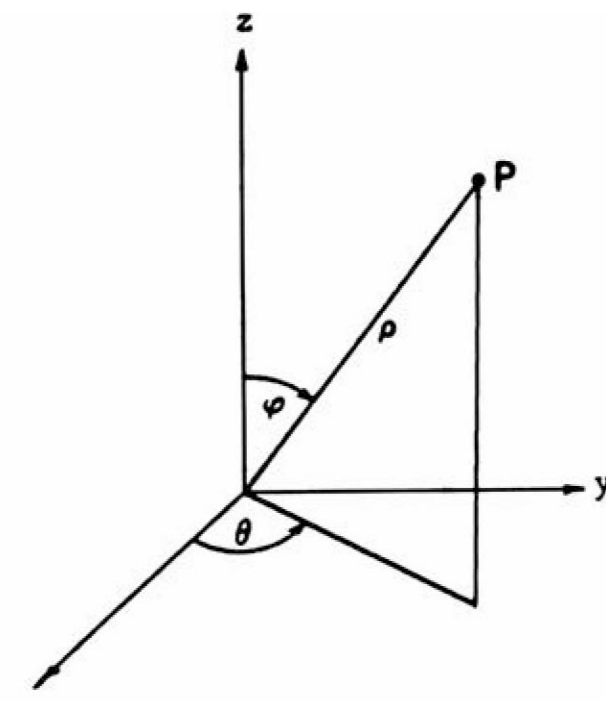
$$z = \rho \cos \phi$$

Element of volume  $dV = \rho^2 \sin \phi d\rho d\phi d\theta$  .

**Figure 196.4** Cylindrical coordinates.



**Figure 196.5** Spherical coordinates.



## Double Integration

The evaluation of a double integral of  $f(x, y)$  over a plane region  $R$ ,

$$\iint_R f(x, y) dA$$

is practically accomplished by iterated (repeated) integration. For example, suppose that a vertical straight line meets the boundary of  $R$  in at most two points so that there is an upper boundary,  $y = y_2(x)$ , and a lower boundary,  $y = y_1(x)$ . Also, it is assumed that these functions are continuous from  $a$  to  $b$  (see [Fig. 196.6](#)). Then

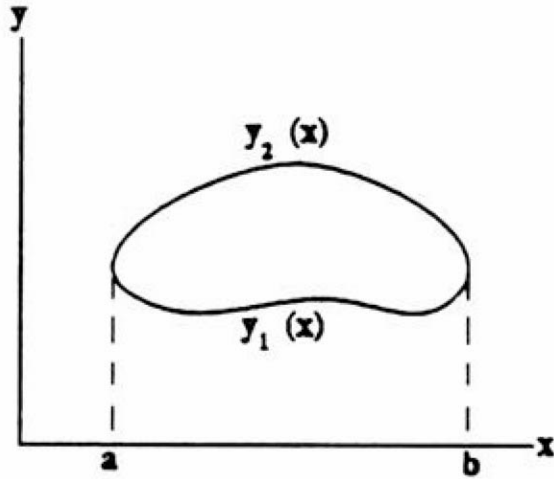
$$\iint_R f(x, y) dA = \int_a^b \left( \int_{y_1(x)}^{y_2(x)} f(x, y) dy \right) dx$$

If  $R$  has left-hand boundary,  $x = x_1(y)$ , and a right-hand boundary,  $x = x_2(y)$ , which are continuous from  $c$  to  $d$  (the extreme values of  $y$  in  $R$ ), then

$$\iint_R f(x, y) dA = \int_c^d \left( \int_{x_1(y)}^{x_2(y)} f(x, y) dx \right) dy$$

Such integrations are sometimes more convenient in polar coordinates,  $x = r \cos \theta$ ,  $y = r \sin \theta$ ,  $dA = r dr d\theta$ .

**Figure 196.6** Region  $R$  bounded by  $y_2(x)$  and  $y_1(x)$ .



## Surface Area and Volume by Double Integration

For the surface given by  $z = f(x, y)$ , which projects onto the closed region  $R$  of the  $xy$  plane, one may calculate the volume  $V$  bounded above by the surface and below by  $R$ , and the surface area  $S$  by the following:

$$V = \int \int_R z \, dA = \int \int_R f(x, y) \, dx \, dy$$

$$S = \int \int_R [1 + (\delta z / \delta x)^2 + (\delta z / \delta y)^2]^{1/2} \, dx \, dy$$

[In polar coordinates,  $(r, \theta)$ , we replace  $dA$  by  $r \, dr \, d\theta$  .]

## Centroid

The centroid of a region  $R$  of the  $xy$  plane is a point  $(x', y')$  where

$$x' = \frac{1}{A} \int \int_R x \, dA, \quad y' = \frac{1}{A} \int \int_R y \, dA$$

and  $A$  is the area of the region.

**Example.** For the circular sector of angle  $2\alpha$  and radius  $R$ , the area  $A$  is  $\alpha R^2$ ; the integral needed for  $x'$ , expressed in polar coordinates, is

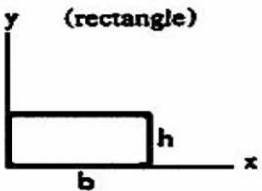
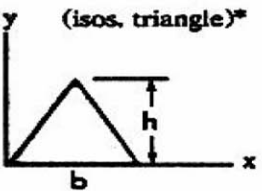
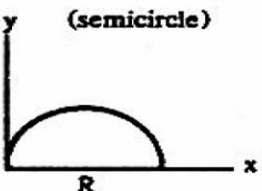
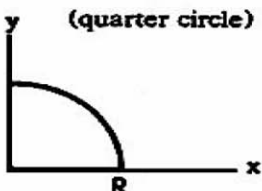
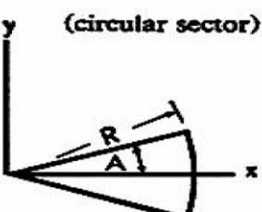
$$\begin{aligned} \int \int x \, dA &= \int_{-\alpha}^{\alpha} \int_0^R (r \cos \theta) r \, dr \, d\theta \\ &= \left[ \frac{R^3}{3} \sin \theta \right]_{-\alpha}^{+\alpha} = \frac{2}{3} R^3 \sin \alpha \end{aligned}$$

and thus,

$$x' = \frac{\frac{2}{3}R^3 \sin \alpha}{\alpha R^2} = \frac{2}{3}R \frac{\sin \alpha}{\alpha}$$

Centroids of some common regions are shown in [Table 196.1](#).

**Table 196.1** Centroids

	Area	$x'$	$y'$
Rectangle (rectangle) 	$bh$	$b/2$	$h/2$
Isosceles triangle* (isos. triangle)* 	$bh/2$	$b/2$	$h/3$
Semicircle (semicircle) 	$\pi R^2/2$	$R$	$4R/3\pi$
Quarter circle (quarter circle) 	$\pi R^2/4$	$4R/3\pi$	$4R/3\pi$
Circular sector (circular sector) 	$R^2 A$	$2R \sin A/3A$	0

\* $y' = h/3$  for any triangle of altitude  $h$ .

## 196.5 Special Functions

---

### Hyperbolic Functions

$$\sinh x = \frac{e^x - e^{-x}}{2}$$

$$\operatorname{csch} x = \frac{1}{\sinh x}$$

$$\cosh x = \frac{e^x + e^{-x}}{2}$$

$$\operatorname{sech} x = \frac{1}{\cosh x}$$

$$\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$\operatorname{ctnh} x = \frac{1}{\tanh x}$$

$$\sinh(-x) = -\sinh x$$

$$\operatorname{ctnh}(-x) = -\operatorname{ctnh} x$$

$$\cosh(-x) = \cosh x$$

$$\operatorname{sech}(-x) = \operatorname{sech} x$$

$$\tanh(-x) = -\tanh x$$

$$\operatorname{csch}(-x) = -\operatorname{csch} x$$

$$\tanh x = \frac{\sinh x}{\cosh x}$$

$$\operatorname{ctnh} x = \frac{\cosh x}{\sinh x}$$

$$\cosh^2 x - \sinh^2 x = 1$$

$$\cosh^2 x = \frac{1}{2}(\cosh 2x + 1)$$

$$\sinh^2 x = \frac{1}{2}(\cosh 2x - 1)$$

$$\operatorname{ctnh}^2 x - \operatorname{csch}^2 x = 1$$

$$\operatorname{csch}^2 x - \operatorname{sech}^2 x = \operatorname{csch}^2 x \operatorname{sech}^2 x$$

$$\tanh^2 x + \operatorname{sech}^2 x = 1$$

$$\sinh(x + y) = \sinh x \cosh y + \cosh x \sinh y$$

$$\cosh(x + y) = \cosh x \cosh y + \sinh x \sinh y$$

$$\sinh(x - y) = \sinh x \cosh y - \cosh x \sinh y$$

$$\cosh(x - y) = \cosh x \cosh y - \sinh x \sinh y$$

$$\tanh(x + y) = \frac{\tanh x + \tanh y}{1 + \tanh x \tanh y}$$

$$\tanh(x - y) = \frac{\tanh x - \tanh y}{1 - \tanh x \tanh y}$$

## Bessel Functions

Bessel functions, also called cylindrical functions, arise in many physical problems as solutions of the differential equation

$$x^2 y'' + xy' + (x^2 - n^2)y = 0$$

which is known as Bessel's equation. Certain solutions, known as *Bessel functions of the first kind of order  $n$* , are given by

$$J_n(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k! \Gamma(n+k+1)} \left(\frac{x}{2}\right)^{n+2k}$$

$$J_{-n}(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k! \Gamma(-n+k+1)} \left(\frac{x}{2}\right)^{-n+2k}$$

In the above it is noteworthy that the gamma function must be defined for the negative argument  $q$ :  $\Gamma(q) = \Gamma(q+1)/q$ , provided that  $q$  is not a negative integer. When  $q$  is a negative integer,  $1/\Gamma(q)$  is defined to be zero. The functions  $J_{-n}(x)$  and  $J_n(x)$  are solutions of Bessel's equation for all real  $n$ . It is seen, for  $n = 1, 2, 3, \dots$ , that

$$J_{-n}(x) = (-1)^n J_n(x)$$

and, therefore, these are not independent; hence, a linear combination of these is not a general solution. When, however,  $n$  is not a positive integer, a negative integer, or zero, the linear combination with arbitrary constants  $c_1$  and  $c_2$ ,

$$y = c_1 J_n(x) + c_2 J_{-n}(x)$$

is the general solution of the Bessel differential equation.

The zero-order function is especially important as it arises in the solution of the heat equation (for a "long" cylinder):

$$J_0(x) = 1 - \frac{x^2}{2^2} + \frac{x^4}{2^2 4^2} - \frac{x^6}{2^2 4^2 6^2} + \dots$$

while the following relations show a connection to the trigonometric functions:

$$J_{1/2}(x) = \left[ \frac{2}{\pi x} \right]^{1/2} \sin x$$

$$J_{-1/2}(x) = \left[ \frac{2}{\pi x} \right]^{1/2} \cos x$$



The following recursion formula gives  $J_{n+1}(x)$  for any order in terms of lower-order functions:

$$\frac{2n}{x} J_n(x) = J_{n-1}(x) + J_{n+1}(x)$$

## Legendre Polynomials

If Laplace's equation,  $\nabla^2 V = 0$ , is expressed in spherical coordinates, it is

$$r^2 \sin \theta \frac{\delta^2 V}{\delta r^2} + 2r \sin \theta \frac{\delta V}{\delta r} + \sin \theta \frac{\delta^2 V}{\delta \theta^2} + \cos \theta \frac{\delta V}{\delta \theta} + \frac{1}{\sin \theta} \frac{\delta^2 V}{\delta \phi^2} = 0$$

and any of its solutions,  $V(r, \theta, \phi)$ , are known as *spherical harmonics*. The solution as a product

$$V(r, \theta, \phi) = R(r)\Theta(\theta)$$

which is independent of  $\phi$ , leads to

$$\sin^2 \theta \Theta'' + \sin \theta \cos \theta \Theta' + [n(n+1) \sin^2 \theta] \Theta = 0$$

Rearrangement and substitution of  $x = \cos \theta$  leads to

$$(1-x^2) \frac{d^2 \Theta}{dx^2} - 2x \frac{d\Theta}{dx} + n(n+1) \Theta = 0$$

known as *Legendre's equation*. Important special cases are those in which  $n$  is zero or a positive integer, and, for such cases, Legendre's equation is satisfied by polynomials called Legendre polynomials,  $P_n(x)$ . A short list of Legendre polynomials, expressed in terms of  $x$  and  $\cos \theta$ , is given below. These are given by the following general formula:

$$P_n(x) = \sum_{j=0}^L \frac{(-1)^j (2n-2j)!}{2^n j! (n-j)! (n-2j)!} x^{n-2j}$$

where  $L = n/2$  if  $n$  is even and  $L = (n-1)/2$  if  $n$  is odd.

$$P_0(x) = 1$$

$$P_1(x) = x$$

$$P_2(x) = \frac{1}{2}(3x^2 - 1)$$

$$P_3(x) = \frac{1}{2}(5x^3 - 3x)$$

$$P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3)$$

$$P_5(x) = \frac{1}{8}(63x^5 - 70x^3 + 15x)$$

$$P_0(\cos \theta) = 1$$

$$P_1(\cos \theta) = \cos \theta$$

$$P_2(\cos \theta) = \frac{1}{4}(3 \cos 2\theta + 1)$$

$$P_3(\cos \theta) = \frac{1}{8}(5 \cos 3\theta + 3 \cos \theta)$$

$$P_4(\cos \theta) = \frac{1}{64}(35 \cos 4\theta + 20 \cos 2\theta + 9)$$

Additional Legendre polynomials may be determined from the *recursion formula*

$$(n+1)P_{n+1}(x) - (2n+1)xP_n(x) + nP_{n-1}(x) = 0 \quad (n = 1, 2, \dots)$$

or the *Rodrigues formula*

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n$$

## Laguerre Polynomials

Laguerre polynomials, denoted  $L_n(x)$ , are solutions of the differential equation

$$xy'' + (1-x)y' + ny = 0$$

and are given by

$$L_n(x) = \sum_{j=0}^n \frac{(-1)^j}{j!} C_{(n,j)} x^j \quad (n = 0, 1, 2, \dots)$$

Thus,

$$L_0(x) = 1$$

$$L_1(x) = 1 - x$$

$$L_2(x) = 1 - 2x + \frac{1}{2}x^2$$

$$L_3(x) = 1 - 3x + \frac{3}{2}x^2 - \frac{1}{6}x^3$$

Additional Laguerre polynomials may be obtained from the recursion formula

$$(n+1)L_{n+1}(x) - (2n+1-x)L_n(x) + nL_{n-1}(x) = 0$$

## Hermite Polynomials

The Hermite polynomials, denoted  $H_n(x)$ , are given by

$$H_0 = 1, \quad H_n(x) = (-1)^n e^{x^2} \frac{d^n e^{-x^2}}{dx^n}, \quad (n = 1, 2, \dots)$$

and are solutions of the differential equation

$$y'' - 2xy' + 2ny = 0 \quad (n = 0, 1, 2, \dots)$$

The first few Hermite polynomials are

$$H_0 = 1$$

$$H_1(x) = 2x$$

$$H_2(x) = 4x^2 - 2$$

$$H_3(x) = 8x^3 - 12x$$

$$H_4(x) = 16x^4 - 48x^2 + 12$$

Additional Hermite polynomials may be obtained from the relation

$$H_{n+1}(x) = 2xH_n(x) - H'_n(x)$$

where prime denotes differentiation with respect to  $x$ .

## Orthogonality

A set of functions  $\{f_n(x)\}$  ( $n = 1, 2, \dots$ ) is orthogonal in an interval  $(a, b)$  with respect to a given weight function  $w(x)$  if

$$\int_a^b w(x) f_m(x) f_n(x) dx = 0 \quad \text{when } m \neq n$$

The following polynomials are orthogonal on the given interval for the given  $w(x)$  :

Legendre polynomials:	$P_n(x)$	$w(x) = 1$ $a = -1, b = 1$
Laguerre polynomials:	$L_n(x)$	$w(x) = \exp(-x)$ $a = 0, b = \infty$
Hermite polynomials:	$H_n(x)$	$w(x) = \exp(-x^2)$ $a = -\infty, b = \infty$

The Bessel functions of order  $n$ ,  $J_n(\lambda_1 x), J_n(\lambda_2 x), \dots$ , are orthogonal with respect to  $w(x) = x$  over the interval  $(0, c)$  provided that the  $\lambda_i$  are the positive roots of  $J_n(\lambda c) = 0$  :

$$\int_0^c x J_n(\lambda_j x) J_n(\lambda_k x) dx = 0 \quad (j \neq k)$$

where  $n$  is fixed and  $n \geq 0$ .

## Functions with $x^2/a^2 \pm y^2/b^2$

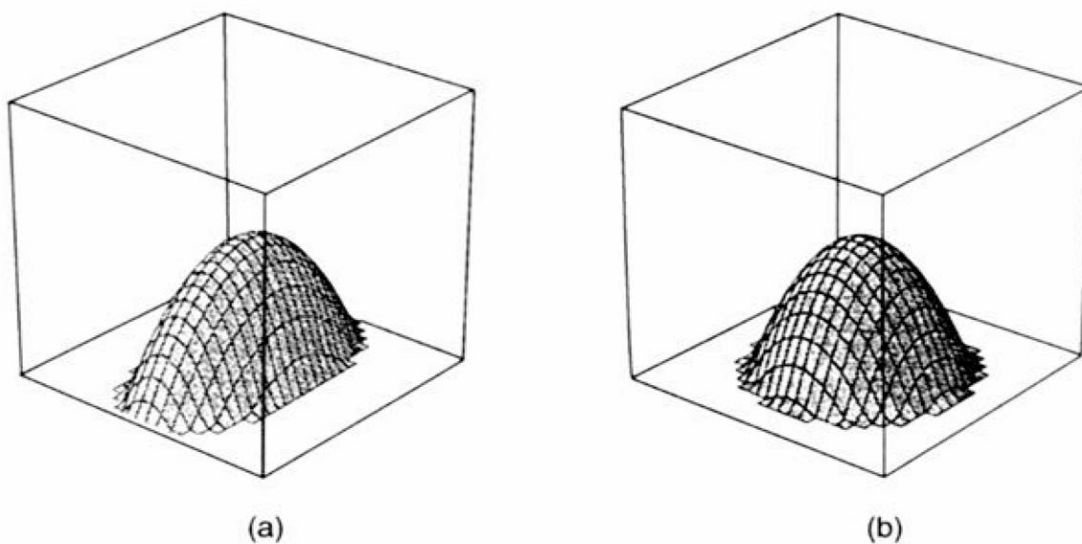
### Elliptic Paraboloid

(Fig. 196.7)

$$z = c(x^2/a^2 + y^2/b^2)$$

$$x^2/a^2 + y^2/b^2 - z/c = 0$$

**Figure 196.7** Elliptic paraboloid. (a)  $a = 0.5, b = 1.0, c = -1.0$ ; viewpoint = (5,-6,4). (b)  $a = 1.0, b = 1.0, c = -2.0$ ; viewpoint = (5,-6,4).



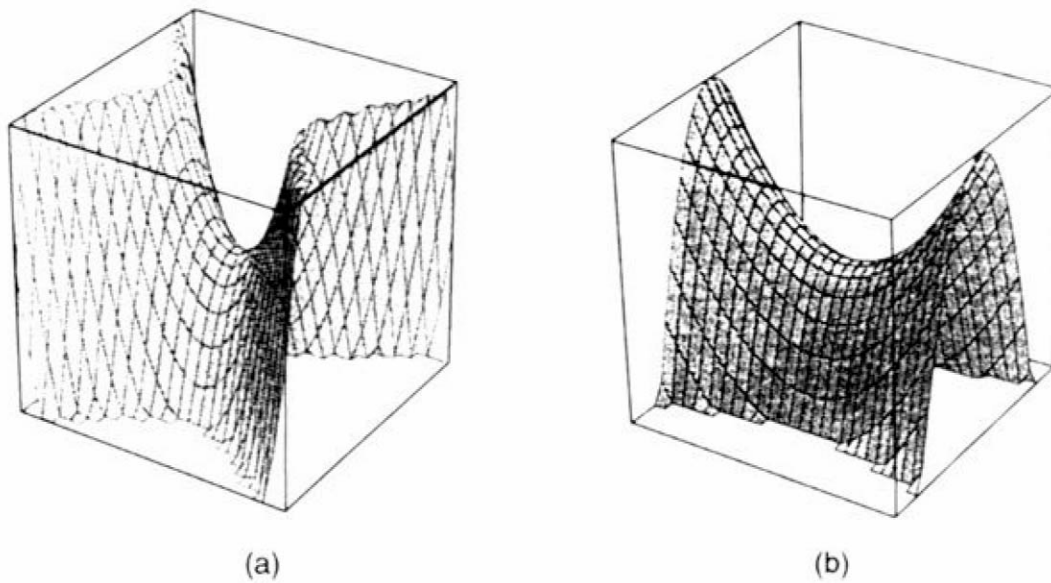
### Hyperbolic Paraboloid (Commonly Called *Saddle*)

(Fig. 196.8)

$$z = c(x^2/a^2 - y^2/b^2)$$

$$x^2/a^2 - y^2/b^2 - z/c = 0$$

**Figure 196.8** Hyperbolic paraboloid. (a)  $a = 0.50$ ,  $b = 0.5$ ,  $c = 1.0$ ; viewpoint = (4,-6,4). (b)  $a = 1.00$ ,  $b = 0.5$ ,  $c = 1.0$ ; viewpoint = (4,-6,4).



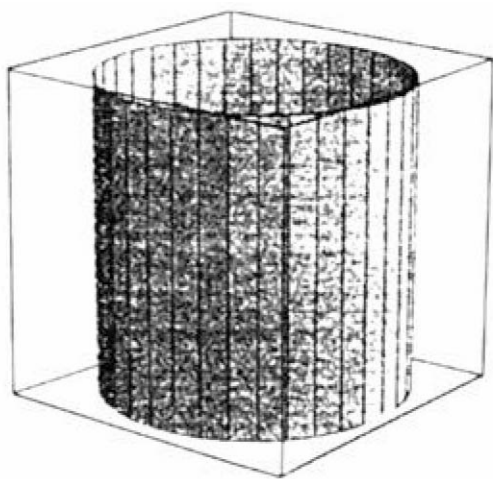
### Elliptic Cylinder

(Fig. 196.9)

$$1 = x^2/a^2 + y^2/b^2$$

$$x^2/a^2 + y^2/b^2 - 1 = 0$$

**Figure 196.9** Elliptic cylinder.  $a = 1.0$ ,  $b = 1.0$ ; viewpoint = (4,-5,2).

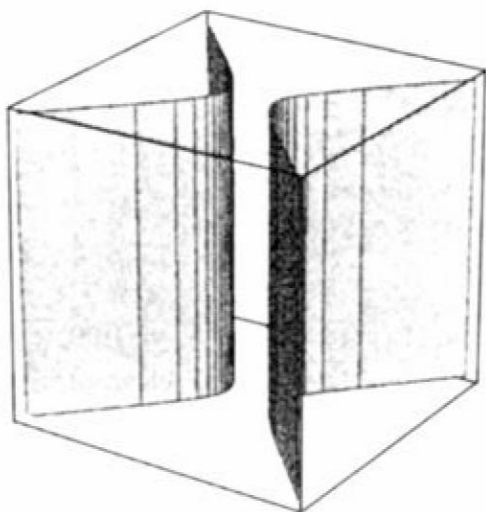


**Hyperbolic Cylinder**  
(Fig. 196.10)

$$1 = x^2/a^2 - y^2/b^2$$

$$x^2/a^2 - y^2/b^2 - 1 = 0$$

**Figure 196.10** Hyperbolic cylinder.  $a = 1.0$ ,  $b = 1.0$ ; viewpoint = (4,-6,3).



## Functions with $(x^2/a^2 + y^2/b^2 \pm c^2)^{1/2}$

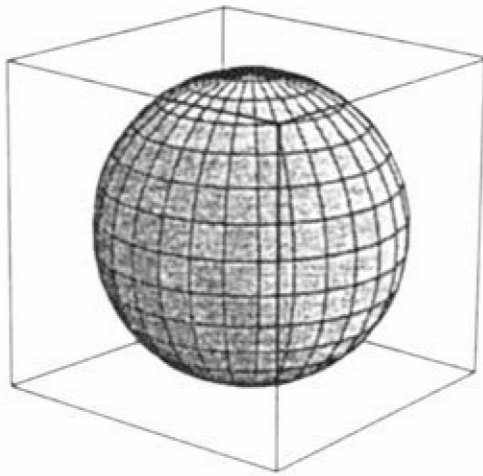
### Sphere

(Fig. 196.11)

$$z = (1 - x^2 - y^2)^{1/2}$$

$$x^2 + y^2 + z^2 - 1 = 0$$

**Figure 196.11** Sphere. Viewpoint = (4,-5,2).



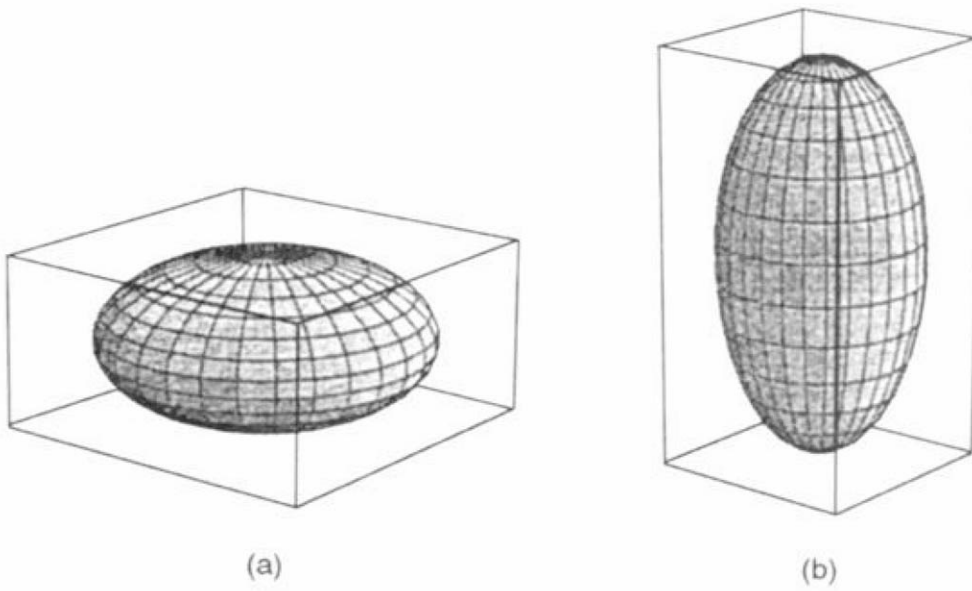
### Ellipsoid

(Fig. 196.12)

$$z = c(1 - x^2/a^2 - y^2/b^2)^{1/2}$$

$$x^2/a^2 + y^2/b^2 + z^2/c^2 - 1 = 0$$

**Figure 196.12** Ellipsoid. (a)  $a = 1.00$ ,  $b = 1.00$ ,  $c = 0.5$ ; viewpoint = (4,-5,2). (b)  $a = 0.50$ ,  $b = 0.50$ ,  $c = 1.0$ ; viewpoint = (4,-5,2).



*Special cases:*

$a = b > c$  gives oblate spheroid

$a = b < c$  gives prolate spheroid

## Cone

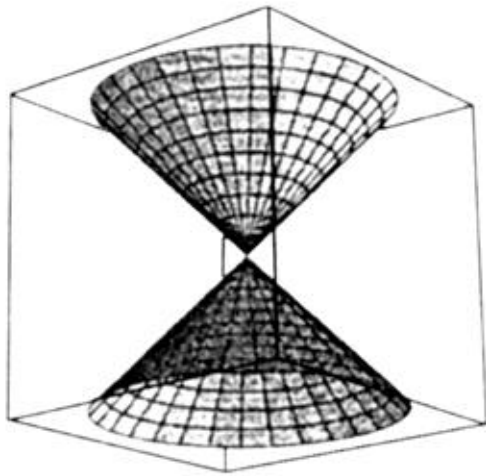
(Fig. 196.13)

$$z = (x^2 + y^2)^{1/2}$$

$$x^2 + y^2 - z^2 = 0$$



**Figure 196.13** Cone. Viewpoint = (4,-5,2).

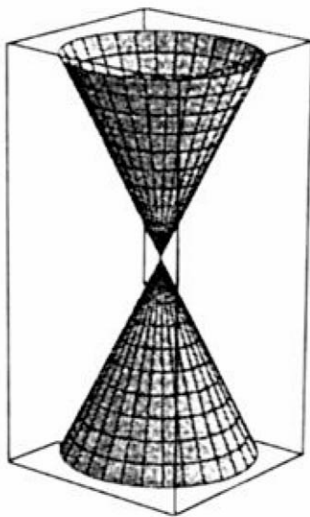


**Elliptic Cone (Circular Cone if  $a = b$ )**  
(Fig. 196.14)

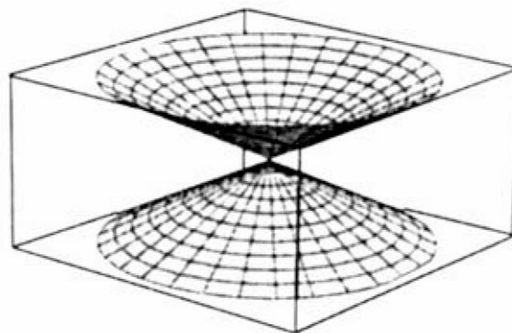
$$z = c(x^2/a^2 + y^2/b^2)^{1/2}$$

$$x^2/a^2 + y^2/b^2 - z^2/c^2 = 0$$

**Figure 196.14** Elliptic cone. (a)  $a = 0.5$ ,  $b = 0.5$ ,  $c = 1.00$ ; viewpoint = (4,-5,2). (b)  $a = 1.0$ ,  $b = 1.0$ ,  $c = 0.50$ ; viewpoint = (4,-5,2).



(a)



(b)

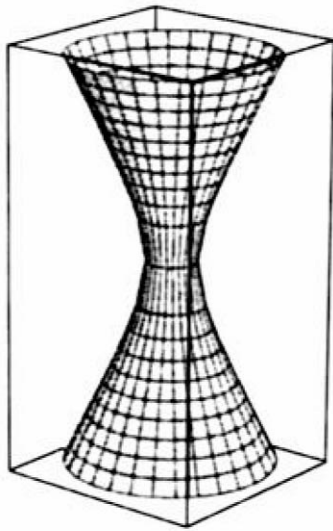
## Hyperboloid of One Sheet

(Fig. 196.15)

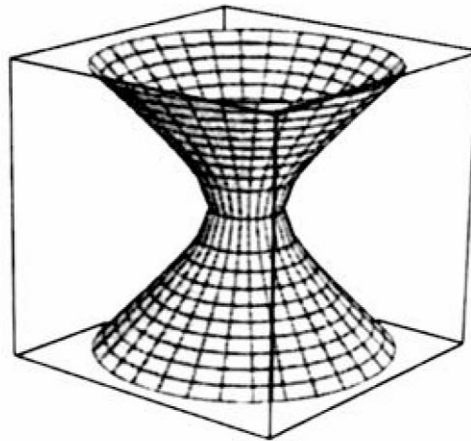
$$z = c(x^2/a^2 + y^2/b^2 - 1)^{1/2}$$

$$x^2/a^2 + y^2/b^2 - z^2/c^2 - 1 = 0$$

**Figure 196.15** Hyperboloid of one sheet. (a)  $a = 0.1, b = 0.1, c = 0.2; \pm z = c\sqrt{15}$ ; viewpoint = (4,-5,2).  
(b)  $a = 0.2, b = 0.2, c = 0.2; \pm z = c\sqrt{15}$ ; viewpoint = (4,-5,2).



(a)



(b)

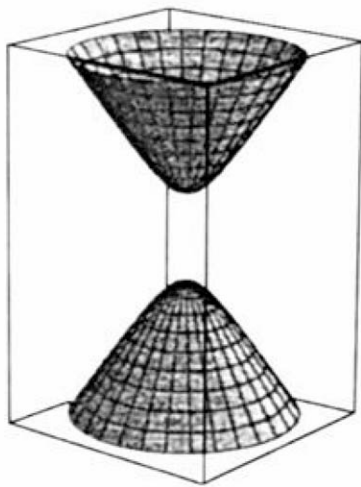
## Hyperboloid of Two Sheets

(Fig. 196.16)

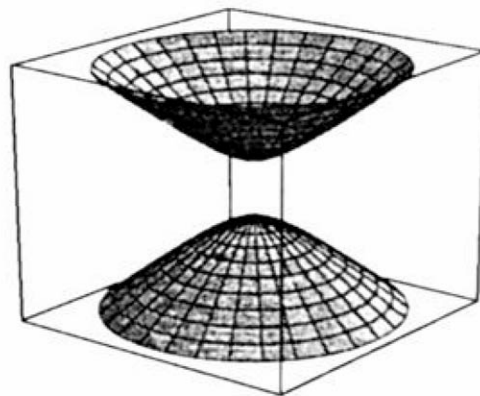
$$z = c(x^2/a^2 + y^2/b^2 + 1)^{1/2}$$

$$x^2/a^2 + y^2/b^2 - z^2/c^2 + 1 = 0$$

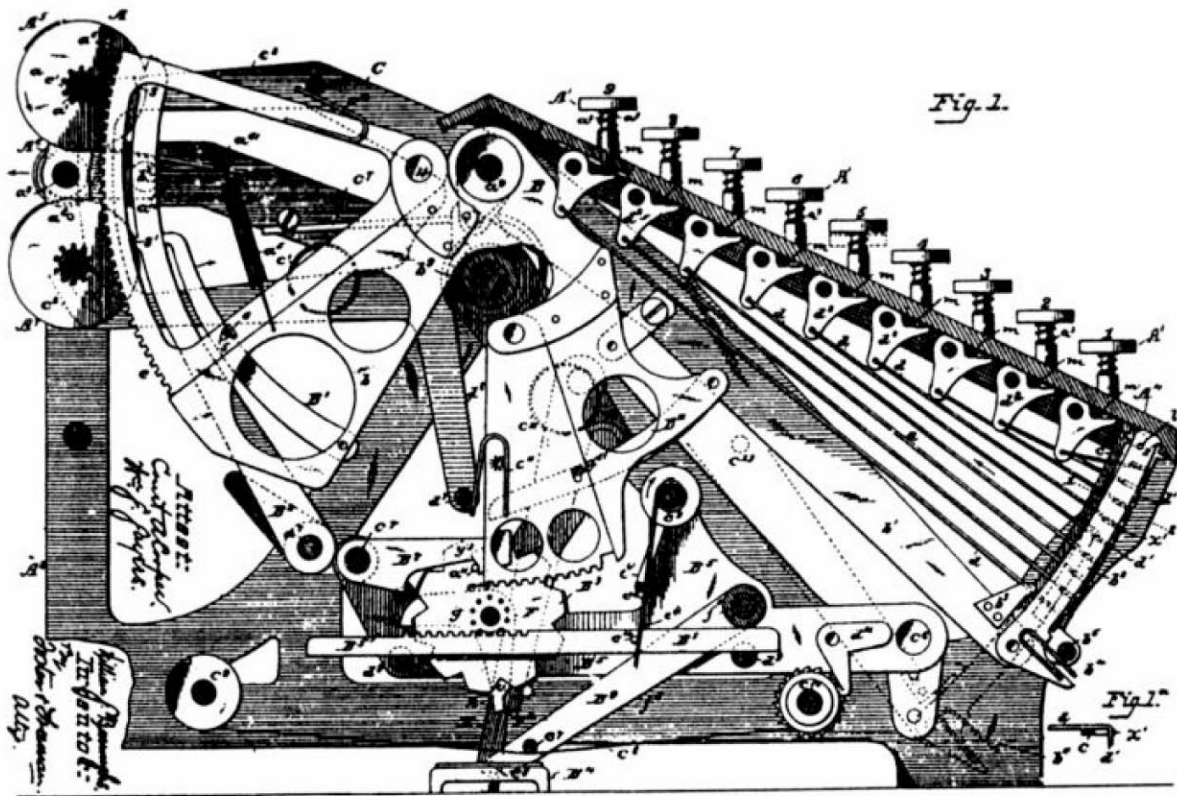
**Figure 196.16** Hyperboloid of two sheets. (a)  $a = 0.125$ ,  $b = 0.125$ ,  $c = 0.2$ ;  $\pm z = c\sqrt{17}$ ; viewpoint =  $(4, -5, 2)$ . (b)  $a = 0.25$ ,  $b = 0.25$ ,  $c = 0.2$ ;  $\pm z = c\sqrt{17}$ ; viewpoint =  $(4, -5, 2)$ .



(a)



(b)



## CALCULATING MACHINE

*William S. Burroughs*

*Patented August 21, 1888*

*#388,116*

An excerpt:

My invention relates to that class of apparatus used for mechanically assisting arithmetic calculations; and my invention consists in the combination, with one or more registers, of a series of independent keys and intervening connections construed, arranged, and operating, as fully specified hereinafter, so as to indicate upon the register the sum of any series of numbers by the proper manipulation of the keys, and also so as to print or permanently record the final result.

Burroughs had invented the adding machine. He and some St. Louis investors formed the American Arithmometer Company to market the machine. Improved versions were developed through the 1890's and Burroughs died in Alabama in 1898 after more than 1000 machines had been sold. In 1905, the company was renamed the Burroughs Adding Machine Company and moved its headquarters to Detroit, Michigan, where it became a force in the computer industry in 1950's and 1960's. Burroughs Corporation merged with Sperry in 1986 to become UniSys.  
(©1994, Dewray Products, Inc. Used with permission.)

Cain, G. "Linear Algebra Matrices"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

- 197.1 Basic Definitions
- 197.2 Algebra of Matrices
- 197.3 Systems of Equations
- 197.4 Vector Spaces
- 197.5 Rank and Nullity
- 197.6 Orthogonality and Length
- 197.7 Determinants
- 197.8 Eigenvalues and Eigenvectors

**George Cain**

*Georgia Institute of Technology*

## 197.1 Basic Definitions

---

A *matrix*  $\mathbf{A}$  is a rectangular array of numbers (real or complex):

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

The *size* of the matrix is said to be  $n \times m$ . The  $1 \times m$  matrices  $[a_{i1} \quad a_{i2} \quad \cdots \quad a_{im}]$  are called *rows* of  $\mathbf{A}$ , and the  $n \times 1$  matrices

$$\begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{bmatrix}$$

are called *columns* of  $\mathbf{A}$ . An  $n \times m$  matrix thus consists of  $n$  rows and  $m$  columns;  $a_{ij}$  denotes the *element*, or *entry*, of  $\mathbf{A}$  in the  $i$ th row and  $j$ th column. A matrix consisting of just one row is called a *row vector*, whereas a matrix of just one column is called a *column vector*. The elements of a vector are frequently called *components* of the vector. When the size of the matrix is clear from the

context, we sometimes write  $\mathbf{A} = (a_{ij})$ .

A matrix with the same number of rows as columns is a *square* matrix, and the number of rows and columns is the *order* of the matrix. The diagonal of an  $n \times n$  square matrix  $\mathbf{A}$  from  $a_{11}$  to  $a_{nn}$  is called the *main*, or *principal*, *diagonal*. The word *diagonal* with no modifier usually means the main diagonal. The *transpose* of a matrix  $\mathbf{A}$  is the matrix that results from interchanging the rows and columns of  $\mathbf{A}$ . It is usually denoted by  $\mathbf{A}^T$ . A matrix  $\mathbf{A}$  such that  $\mathbf{A} = \mathbf{A}^T$  is said to be *symmetric*. The *conjugate transpose* of  $\mathbf{A}$  is the matrix that results from replacing each element of  $\mathbf{A}^T$  by its complex conjugate, and is usually denoted by  $\mathbf{A}^H$ . A matrix such that  $\mathbf{A} = \mathbf{A}^H$  is said to be *Hermitian*.

A square matrix  $\mathbf{A} = (a_{ij})$  is *lower triangular* if  $a_{ij} = 0$  for  $j > i$  and is *upper triangular* if  $a_{ij} = 0$  for  $j < i$ . A matrix that is both upper and lower triangular is a *diagonal* matrix. The  $n \times n$  *identity matrix* is the  $n \times n$  diagonal matrix in which each element of the main diagonal is 1. It is traditionally denoted  $\mathbf{I}_n$ , or simply  $\mathbf{I}$  when the order is clear from the context.

## 197.2 Algebra of Matrices

---

The sum and difference of two matrices  $\mathbf{A}$  and  $\mathbf{B}$  are defined whenever  $\mathbf{A}$  and  $\mathbf{B}$  have the same size. In that case  $\mathbf{C} = \mathbf{A} \pm \mathbf{B}$  is defined by  $\mathbf{C} = (c_{ij}) = (a_{ij} \pm b_{ij})$ . The product  $t\mathbf{A}$  of a scalar  $t$  (real or complex number) and a matrix  $\mathbf{A}$  is defined by  $t\mathbf{A} = (ta_{ij})$ . If  $\mathbf{A}$  is an  $n \times m$  matrix and  $\mathbf{B}$  is an  $m \times p$  matrix, the product  $\mathbf{C} = \mathbf{AB}$  is defined to be the  $n \times p$  matrix  $\mathbf{C} = (c_{ij})$  given by  $c_{ij} = \sum_{k=1}^m a_{ik}b_{kj}$ . Note that the product of an  $n \times m$  matrix and an  $m \times p$  matrix is an  $n \times p$  matrix, and the product is defined only when the number of columns of the first factor is the same as the number of rows of the second factor. Matrix multiplication is, in general, associative:  $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$ . It also distributes over addition (and subtraction):

$$\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC} \quad \text{and} \quad (\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$$

It is, however, not in general true that  $\mathbf{AB} = \mathbf{BA}$ , even in case both products are defined. It is clear that  $(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$  and  $(\mathbf{A} + \mathbf{B})^H = \mathbf{A}^H + \mathbf{B}^H$ . It is also true, but not so obvious perhaps, that  $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$  and  $(\mathbf{AB})^H = \mathbf{B}^H\mathbf{A}^H$ .

The  $n \times n$  identity matrix  $\mathbf{I}$  has the property that  $\mathbf{IA} = \mathbf{AI} = \mathbf{A}$  for every  $n \times n$  matrix  $\mathbf{A}$ . If  $\mathbf{A}$  is square, and if there is a matrix  $\mathbf{B}$  such that  $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$ , then  $\mathbf{B}$  is called the *inverse* of  $\mathbf{A}$  and is denoted  $\mathbf{A}^{-1}$ . This terminology and notation are justified by the fact that a matrix can have at most one inverse. A matrix having an inverse is said to be *invertible*, or *nonsingular*, while a matrix not having an inverse is said to be *noninvertible*, or *singular*. The product of two invertible matrices is invertible and, in fact,  $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ . The sum of two invertible matrices is, obviously, not necessarily invertible.

## 197.3 Systems of Equations

---

The system of  $n$  linear equations in  $m$  unknowns

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1m}x_m &= b_1 \\
a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2m}x_m &= b_2 \\
&\vdots \\
a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \cdots + a_{nm}x_m &= b_n
\end{aligned}$$

may be written  $\mathbf{Ax} = \mathbf{b}$ , where  $\mathbf{A} = (a_{ij})$ ,  $\mathbf{x} = [x_1 \ x_2 \ \cdots \ x_m]^T$ , and  $\mathbf{b} = [b_1 \ b_2 \ \cdots \ b_n]^T$ . Thus  $\mathbf{A}$  is an  $n \times m$  matrix, and  $\mathbf{x}$  and  $\mathbf{b}$  are column vectors of the appropriate sizes.

The matrix  $\mathbf{A}$  is called the *coefficient matrix* of the system. Let us first suppose the coefficient matrix is square; that is, there are an equal number of equations and unknowns. If  $\mathbf{A}$  is upper triangular, it is quite easy to find all solutions of the system. The  $i$ th equation will contain only the unknowns  $x_i, x_{i+1}, \dots, x_n$ , and one simply solves the equations in reverse order: the last equation is solved for  $x_n$ ; the result is substituted into the  $(n-1)$ st equation, which is then solved for  $x_{n-1}$ ; these values of  $x_n$  and  $x_{n-1}$  are substituted in the  $(n-2)$ th equation, which is solved for  $x_{n-2}$ , and so on. This procedure is known as *back substitution*.

The strategy for solving an arbitrary system is to find an upper-triangular system equivalent with it and solve this upper-triangular system using back substitution. First suppose the element  $a_{11} \neq 0$ . We may rearrange the equations to ensure this, unless, of course the first column of  $\mathbf{A}$  is all 0s. In this case proceed to the next step, to be described later. For each  $i \geq 2$  let  $m_{i1} = a_{i1}/a_{11}$ . Now replace the  $i$ th equation by the result of multiplying the first equation by  $m_{i1}$  and subtracting the new equation from the  $i$ th equation. Thus,

$$a_{i1}x_1 + a_{i2}x_2 + a_{i3}x_3 + \cdots + a_{im}x_m = b_i$$

is replaced by

$$0 \cdot x_1 + (a_{i2} + m_{i1}a_{12})x_2 + (a_{i3} + m_{i1}a_{13})x_3 + \cdots + (a_{im} + m_{i1}a_{1m})x_m = b_i + m_{i1}b_1$$

After this is done for all  $i = 2, 3, \dots, n$ , there results the equivalent system

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1 \\
0 \cdot x_1 + a'_{22}x_2 + a'_{23}x_3 + \cdots + a'_{2n}x_n &= b'_2 \\
0 \cdot x_1 + a'_{32}x_2 + a'_{33}x_3 + \cdots + a'_{3n}x_n &= b'_3 \\
&\vdots \\
0 \cdot x_1 + a'_{n2}x_2 + a'_{n3}x_3 + \cdots + a'_{nn}x_n &= b'_n
\end{aligned}$$

in which all entries in the first column below  $a_{11}$  are 0. (Note that if all entries in the first column were 0 to begin with, then  $a_{11} = 0$  also.) This procedure is now repeated for the  $(n-1) \times (n-1)$  system



$$\begin{aligned}
a'_{22}x_2 + a'_{23}x_3 + \cdots + a'_{2n}x_n &= b'_2 \\
a'_{32}x_2 + a'_{33}x_3 + \cdots + a'_{3n}x_n &= b'_3 \\
&\vdots \\
a'_{n2}x_2 + a'_{n3}x_3 + \cdots + a'_{nn}x_n &= b'_n
\end{aligned}$$

to obtain an equivalent system in which all entries of the coefficient matrix below  $a'_{22}$  are 0. Continuing, we obtain an upper-triangular system  $\mathbf{U}\mathbf{x} = \mathbf{c}$  equivalent with the original system. This procedure is known as *Gaussian elimination*. The numbers  $m_{ij}$  are known as the *multipliers*.

Essentially the same procedure may be used in case the coefficient matrix is not square. If the coefficient matrix is not square, we may make it square by appending either rows or columns of 0s as needed. Appending rows of 0s and appending 0s to make  $\mathbf{b}$  have the appropriate size is equivalent to appending equations  $0 = 0$  to the system. Clearly the new system has precisely the same solutions as the original system. Appending columns of 0s and adjusting the size of  $\mathbf{x}$  appropriately yields a new system with additional unknowns, each appearing only with coefficient 0, thus not affecting the solutions of the original system. In either case we may assume the coefficient matrix is square, and apply the Gauss elimination procedure.

Suppose the matrix  $\mathbf{A}$  is invertible. Then if there were no row interchanges in carrying out the above Gauss elimination procedure, we have the *LU factorization* of the matrix  $\mathbf{A}$ :

$$\mathbf{A} = \mathbf{L}\mathbf{U}$$

where  $\mathbf{U}$  is the upper-triangular matrix produced by elimination and  $\mathbf{L}$  is the lower-triangular matrix given by

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ m_{21} & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \\ m_{n1} & m_{n2} & \cdots & & 1 \end{bmatrix}$$

A *permutation*  $\mathbf{P}_{ij}$  matrix is an  $n \times n$  matrix such that  $\mathbf{P}_{ij}\mathbf{A}$  is the matrix that results from exchanging row  $i$  and  $j$  of the matrix  $\mathbf{A}$ . The matrix  $\mathbf{P}_{ij}$  is the matrix that results from exchanging rows  $i$  and  $j$  of the identity matrix. A product  $\mathbf{P}$  of such matrices  $\mathbf{P}_{ij}$  is called a *permutation* matrix. If row interchanges are required in the Gauss elimination procedure, then we have the factorization

$$\mathbf{P}\mathbf{A} = \mathbf{L}\mathbf{U}$$

where  $\mathbf{P}$  is the permutation matrix giving the required row exchanges.

## 197.4 Vector Spaces

The collection of all column vectors with  $n$  real components is *Euclidean  $n$ -space*, and is denoted  $\mathbf{R}^n$ . The collection of column vectors with  $n$  complex components is denoted  $\mathbf{C}^n$ . We shall use *vector space* to mean either  $\mathbf{R}^n$  or  $\mathbf{C}^n$ . In discussing the space  $\mathbf{R}^n$ , the word *scalar* will mean a real number, and in discussing the space  $\mathbf{C}^n$ , it will mean a complex number. A subset  $\mathbf{S}$  of a vector space is a *subspace* such that if  $\mathbf{u}$  and  $\mathbf{v}$  are vectors in  $\mathbf{S}$ , and if  $c$  is any scalar, then  $\mathbf{u} + \mathbf{v}$  and  $c\mathbf{u}$  are in  $\mathbf{S}$ . We shall sometimes use the word *space* to mean a subspace. If  $B = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$  is a collection of vectors in a vector space, then the set  $\mathbf{S}$  consisting of all vectors  $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \dots + c_m\mathbf{v}_m$  for all scalars  $c_1, c_2, \dots, c_m$  is a subspace, called the *span* of  $B$ . A collection  $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m\}$  of vectors  $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \dots + c_m\mathbf{v}_m$  is a *linear combination* of  $B$ . If  $\mathbf{S}$  is a subspace and  $B = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m\}$  is a subset of  $\mathbf{S}$  such that  $\mathbf{S}$  is the span of  $B$ , then  $B$  is said to *span*  $\mathbf{S}$ .

A collection  $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m\}$  of  $n$ -vectors is *linearly dependent* if there exist scalars  $c_1, c_2, \dots, c_m$ , not all zero, such that  $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \dots + c_m\mathbf{v}_m = \mathbf{0}$ . A collection of vectors that is not linearly dependent is said to be *linearly independent*. The modifier *linearly* is frequently omitted, and we speak simply of dependent and independent collections. A linearly independent collection of vectors in a space  $\mathbf{S}$  that spans  $\mathbf{S}$  is a *basis* of  $\mathbf{S}$ . Every basis of a space  $\mathbf{S}$  contains the same number of vectors; this number is the *dimension* of  $\mathbf{S}$ . The dimension of the space consisting of only the zero vector is 0. The collection  $B = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ , where  $\mathbf{e}_1 = [1, 0, 0, \dots, 0]^T$ ,  $\mathbf{e}_2 = [0, 1, 0, \dots, 0]^T$ , and so forth ( $\mathbf{e}_i$  has 1 as its  $i$ th component and zero for all other components) is a basis for the spaces  $\mathbf{R}^n$  and  $\mathbf{C}^n$ . This is the *standard basis* for these spaces. The dimension of these spaces is thus  $n$ . In a space  $\mathbf{S}$  of dimension  $n$ , no collection of fewer than  $n$  vectors can span  $\mathbf{S}$ , and no collection of more than  $n$  vectors in  $\mathbf{S}$  can be independent.

## 197.5 Rank and Nullity

The *column space* of an  $n \times m$  matrix  $\mathbf{A}$  is the subspace of  $\mathbf{R}^n$  or  $\mathbf{C}^n$  spanned by the columns of  $\mathbf{A}$ . The *row space* is the subspace of  $\mathbf{R}^m$  or  $\mathbf{C}^m$  spanned by the rows of  $\mathbf{A}$ . Note that for any vector  $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_m]^T$ ,

$$\mathbf{Ax} = x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{bmatrix} + \dots + x_m \begin{bmatrix} a_{1m} \\ a_{2m} \\ \vdots \\ a_{nm} \end{bmatrix}$$

so that the column space is the collection of all vectors  $\mathbf{Ax}$ , and thus the system  $\mathbf{Ax} = \mathbf{b}$  has a solution if and only if  $\mathbf{b}$  is a member of the column space of  $\mathbf{A}$ .

The dimension of the column space is the *rank* of  $\mathbf{A}$ . The row space has the same dimension as the column space. The set of all solutions of the system  $\mathbf{Ax} = \mathbf{0}$  is a subspace called the *null space* of  $\mathbf{A}$ , and the dimension of this null space is the *nullity* of  $\mathbf{A}$ . A fundamental result in matrix theory is the fact that, for an  $n \times m$  matrix  $\mathbf{A}$ ,

$$\text{rank } \mathbf{A} + \text{nullity } \mathbf{A} = m$$

The difference of any two solutions of the linear system  $\mathbf{Ax} = \mathbf{b}$  is a member of the null space of  $\mathbf{A}$ . Thus this system has at most one solution if and only if the nullity of  $\mathbf{A}$  is zero. If the system is square (that is, if  $\mathbf{A}$  is  $n \times n$ ), then there will be a solution for every right-hand side  $\mathbf{b}$  if and only if the collection of columns of  $\mathbf{A}$  is linearly independent, which is the same as saying the rank of  $\mathbf{A}$  is  $n$ . In this case the nullity must be zero. Thus, for any  $\mathbf{b}$ , the square system  $\mathbf{Ax} = \mathbf{b}$  has exactly one solution if and only if  $\text{rank } \mathbf{A} = n$ . In other words the  $n \times n$  matrix  $\mathbf{A}$  is invertible if and only if  $\text{rank } \mathbf{A} = n$ .

## 197.6 Orthogonality and Length

The *inner product* of two vectors  $\mathbf{x}$  and  $\mathbf{y}$  is the scalar  $\mathbf{x}^H \mathbf{y}$ . The *length*, or *norm*,  $\|\mathbf{x}\|$ , of the vector  $\mathbf{x}$  is given by  $\|\mathbf{x}\| = \sqrt{\mathbf{x}^H \mathbf{x}}$ . A *unit vector* is a vector of norm 1. Two vectors  $\mathbf{x}$  and  $\mathbf{y}$  are *orthogonal* if  $\mathbf{x}^H \mathbf{y} = 0$ . A collection of vectors  $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m\}$  in a space  $\mathbf{S}$  is said to be an *orthonormal* collection if  $\mathbf{v}_i^H \mathbf{v}_j = 0$  for  $i \neq j$  and  $\mathbf{v}_i^H \mathbf{v}_i = 1$ . An orthonormal collection is necessarily linearly independent. If  $\mathbf{S}$  is a subspace (of  $\mathbb{R}^n$  or  $\mathbb{C}^n$ ) spanned by the orthonormal collection  $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m\}$ , then the *projection* of a vector  $\mathbf{x}$  onto  $\mathbf{S}$  is the vector

$$\text{proj}(\mathbf{x}; \mathbf{S}) = (\mathbf{x}^H \mathbf{v}_1) \mathbf{v}_1 + (\mathbf{x}^H \mathbf{v}_2) \mathbf{v}_2 + \dots + (\mathbf{x}^H \mathbf{v}_m) \mathbf{v}_m$$

The projection of  $\mathbf{x}$  onto  $\mathbf{S}$  minimizes the function  $f(\mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$  for  $\mathbf{y} \in \mathbf{S}$ . In other words the projection of  $\mathbf{x}$  onto  $\mathbf{S}$  is the vector in  $\mathbf{S}$  that is "closest" to  $\mathbf{x}$ .

If  $\mathbf{b}$  is a vector and  $\mathbf{A}$  is an  $n \times m$  matrix, then a vector  $\mathbf{x}$  minimizes  $\|\mathbf{b} - \mathbf{Ax}\|^2$  if and only if it is a solution of  $\mathbf{A}^H \mathbf{Ax} = \mathbf{A}^H \mathbf{b}$ . This system of equations is called the *system of normal equations* for the least-squares problem of minimizing  $\|\mathbf{b} - \mathbf{Ax}\|^2$ .

If  $\mathbf{A}$  is an  $n \times m$  matrix and  $\text{rank } \mathbf{A} = k$ , then there is a  $n \times k$  matrix  $\mathbf{Q}$  whose columns form an orthonormal basis for the column space of  $\mathbf{A}$  and a  $k \times m$  upper-triangular matrix  $\mathbf{R}$  of rank  $k$  such that

$$\mathbf{A} = \mathbf{QR}$$

This is called the *QR factorization* of  $\mathbf{A}$ . It now follows that  $\mathbf{x}$  minimizes  $\|\mathbf{b} - \mathbf{Ax}\|^2$  if and only if it is a solution of the upper-triangular system  $\mathbf{Rx} = \mathbf{Q}^H \mathbf{b}$ .

If  $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m\}$  is a basis for a space  $\mathbf{S}$ , the following procedure produces an orthonormal basis  $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m\}$  for  $\mathbf{S}$ .

Set  $\mathbf{v}_1 = \mathbf{w}_1 / \|\mathbf{w}_1\|$ .

Let  $\tilde{\mathbf{v}}_2 = \mathbf{w}_2 - \text{proj}(\mathbf{w}_2; \mathbf{S}_1)$ , where  $\mathbf{S}_1$  is the span of  $\{\mathbf{v}_1\}$ ; set  $\mathbf{v}_2 = \tilde{\mathbf{v}}_2 / \|\tilde{\mathbf{v}}_2\|$ .

Next, let  $\tilde{\mathbf{v}}_3 = \mathbf{w}_3 - \text{proj}(\mathbf{w}_3; \mathbf{S}_2)$ , where  $\mathbf{S}_2$  is the span of  $\{\mathbf{v}_1, \mathbf{v}_2\}$ ; set  $\mathbf{v}_3 = \tilde{\mathbf{v}}_3 / \|\tilde{\mathbf{v}}_3\|$ .

And so on:  $\tilde{\mathbf{v}}_i = \mathbf{w}_i - \text{proj}(\mathbf{w}_i; \mathbf{S}_{i-1})$ , where  $\mathbf{S}_{i-1}$  is the span of  $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{i-1}\}$ ; set  $\mathbf{v}_i = \tilde{\mathbf{v}}_i / \|\tilde{\mathbf{v}}_i\|$ . This is the *Gram-Schmidt procedure*.

If the collection of columns of a square matrix is an orthonormal collection, the matrix is called a *unitary matrix*. In case the matrix is a real matrix, it is usually called an *orthogonal matrix*. A unitary matrix  $\mathbf{U}$  is invertible, and  $\mathbf{U}^{-1} = \mathbf{U}^H$ . (In the real case an orthogonal matrix  $\mathbf{Q}$  is invertible, and  $\mathbf{Q}^{-1} = \mathbf{Q}^T$ .)

## 197.7 Determinants

---

The *determinant* of a square matrix is defined inductively. First, suppose the determinant  $\det \mathbf{A}$  has been defined for all square matrices of order  $< n$ . Then

$$\det \mathbf{A} = a_{11} \mathbf{C}_{11} + a_{12} \mathbf{C}_{12} + \cdots + a_{1n} \mathbf{C}_{1n}$$

where the numbers  $\mathbf{C}_{ij}$  are *cofactors* of the matrix  $\mathbf{A}$ :

$$\mathbf{C}_{ij} = (-1)^{i+j} \det \mathbf{M}_{ij}$$

where  $\mathbf{M}_{ij}$  is the  $(n-1) \times (n-1)$  matrix obtained by deleting the  $i$ th row and  $j$ th column of  $\mathbf{A}$ . Now  $\det \mathbf{A}$  is defined to be the only entry of a matrix of order 1. Thus, for a matrix of order 2, we have

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$$

There are many interesting but not obvious properties of determinants. It is true that

$$\det \mathbf{A} = a_{i1} \mathbf{C}_{i1} + a_{i2} \mathbf{C}_{i2} + \cdots + a_{in} \mathbf{C}_{in}$$

for any  $1 \leq i \leq n$ . It is also true that  $\det \mathbf{A} = \det \mathbf{A}^T$ , so that we have

$$\det \mathbf{A} = a_{1j} \mathbf{C}_{1j} + a_{2j} \mathbf{C}_{2j} + \cdots + a_{nj} \mathbf{C}_{nj}$$

for any  $1 \leq j \leq n$ .

If  $\mathbf{A}$  and  $\mathbf{B}$  are matrices of the same order, then  $\det \mathbf{AB} = (\det \mathbf{A})(\det \mathbf{B})$ , and the determinant of any identity matrix is 1. Perhaps the most important property of the determinant is the fact that a matrix is invertible if and only if its determinant is not zero.

## 197.8 Eigenvalues and Eigenvectors

---

If  $\mathbf{A}$  is a square matrix, and  $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$  for a scalar  $\lambda$  and a nonzero  $\mathbf{v}$ , then  $\lambda$  is an *eigenvalue* of  $\mathbf{A}$  and  $\mathbf{v}$  is an *eigenvector* of  $\mathbf{A}$  that *corresponds* to  $\lambda$ . Any nonzero linear combination of eigenvectors corresponding to the same eigenvalue  $\lambda$  is also an eigenvector corresponding to  $\lambda$ . The collection of all eigenvectors corresponding to a given eigenvalue  $\lambda$  is thus a subspace, called an *eigenspace* of  $\mathbf{A}$ . A collection of eigenvectors corresponding to different eigenvalues is

necessarily linear-independent. It follows that a matrix of order  $n$  can have at most  $n$  distinct eigenvectors. In fact, the eigenvalues of  $\mathbf{A}$  are the roots of the  $n$ th degree polynomial equation

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0$$

called the *characteristic equation* of  $\mathbf{A}$ . (Eigenvalues and eigenvectors are frequently called *characteristic values* and *characteristic vectors*.)

If the  $n$ th order matrix  $\mathbf{A}$  has an independent collection of  $n$  eigenvectors, then  $\mathbf{A}$  is said to have a *full set* of eigenvectors. In this case there is a set of eigenvectors of  $\mathbf{A}$  that is a basis for  $\mathbb{R}^n$  or, in the complex case,  $\mathbb{C}^n$ . In case there are  $n$  distinct eigenvalues of  $\mathbf{A}$ , then, of course,  $\mathbf{A}$  has a full set of eigenvectors. If there are fewer than  $n$  distinct eigenvalues, then  $\mathbf{A}$  may or may not have a full set of eigenvectors. If there is a full set of eigenvectors, then

$$\mathbf{D} = \mathbf{S}^{-1} \mathbf{A} \mathbf{S} \quad \text{or} \quad \mathbf{A} = \mathbf{S} \mathbf{D} \mathbf{S}^{-1}$$

where  $\mathbf{D}$  is a diagonal matrix with the eigenvalues of  $\mathbf{A}$  on the diagonal, and  $\mathbf{S}$  is a matrix whose columns are the full set of eigenvectors. If  $\mathbf{A}$  is symmetric, there are  $n$  real distinct eigenvalues of  $\mathbf{A}$  and the corresponding eigenvectors are orthogonal. There is thus an orthonormal collection of eigenvectors that span  $\mathbb{R}^n$ , and we have

$$\mathbf{A} = \mathbf{Q} \mathbf{D} \mathbf{Q}^T \quad \text{and} \quad \mathbf{D} = \mathbf{Q}^T \mathbf{A} \mathbf{Q}$$

where  $\mathbf{Q}$  is a real orthogonal matrix and  $\mathbf{D}$  is diagonal. For the complex case, if  $\mathbf{A}$  is Hermitian, we have

$$\mathbf{A} = \mathbf{U} \mathbf{D} \mathbf{U}^H \quad \text{and} \quad \mathbf{D} = \mathbf{U}^H \mathbf{A} \mathbf{U}$$

where  $\mathbf{U}$  is a unitary matrix and  $\mathbf{D}$  is a *real* diagonal matrix. (A Hermitian matrix also has  $n$  distinct real eigenvalues.)

## References

Daniel, J. W., and Noble, B. 1988. *Applied Linear Algebra*. Prentice Hall, Englewood Cliffs, NJ.  
Strang, G. 1993. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, Wellesley, MA.

Cain, G. "Vector Algebra and Calculus"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Vector Algebra and Calculus

---

- 198.1 Basic Definitions
- 198.2 Coordinate Systems
- 198.3 Vector Functions
- 198.4 Gradient, Curl, and Divergence
- 198.5 Integration
- 198.6 Integral Theorems

**George Cain**

*Georgia Institute of Technology*

## 198.1 Basic Definitions

---

A vector is a directed line segment, with two vectors being equal if they have the same length and the same direction. More precisely, a *vector* is an equivalence class of directed line segments, where two directed segments are equivalent if they have the same length and the same direction. The *length* of a vector is the common length of its directed segments, and the *angle between* vectors is the angle between any of their segments. The length of a vector  $\mathbf{u}$  is denoted  $|\mathbf{u}|$ . There is defined a distinguished vector having zero length, which is usually denoted  $\mathbf{0}$ . It is frequently useful to visualize a directed segment as an arrow; we then speak of the nose and the tail of the segment. The *sum*  $\mathbf{u} + \mathbf{v}$  of two vectors  $\mathbf{u}$  and  $\mathbf{v}$  is defined by taking directed segments from  $\mathbf{u}$  and  $\mathbf{v}$  and placing the tail of the segment representing  $\mathbf{v}$  at the nose of the segment representing  $\mathbf{u}$  and defining  $\mathbf{u} + \mathbf{v}$  to be the vector determined by the segment from the tail of the  $\mathbf{u}$  representative to the nose of the  $\mathbf{v}$  representative. It is easy to see that  $\mathbf{u} + \mathbf{v}$  is well defined and that  $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$ . Subtraction is the inverse operation of addition. Thus the *difference*  $\mathbf{u} - \mathbf{v}$  of two vectors is defined to be the vector that when added to  $\mathbf{v}$  gives  $\mathbf{u}$ . In other words, if we take a segment from  $\mathbf{u}$  and a segment from  $\mathbf{v}$  and place their tails together, the difference is the segment from the nose of  $\mathbf{v}$  to the nose of  $\mathbf{u}$ . The zero vector behaves as one might expect:  $\mathbf{u} + \mathbf{0} = \mathbf{u}$ , and  $\mathbf{u} - \mathbf{u} = \mathbf{0}$ . Addition is associative:  $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$ .

To distinguish them from vectors, the real numbers are called *scalars*. The product  $t\mathbf{u}$  of a scalar  $t$  and a vector  $\mathbf{u}$  is defined to be the vector having length  $|t||\mathbf{u}|$  and direction the same as  $\mathbf{u}$  if  $t > 0$ , the opposite direction if  $t < 0$ . If  $t = 0$ , then  $t\mathbf{u}$  is defined to be the zero vector. Note that  $t(\mathbf{u} + \mathbf{v}) = t\mathbf{u} + t\mathbf{v}$ , and  $(t + s)\mathbf{u} = t\mathbf{u} + s\mathbf{u}$ . From this it follows that  $\mathbf{u} - \mathbf{v} = \mathbf{u} + (-1)\mathbf{v}$ .

The *scalar product*  $\mathbf{u} \cdot \mathbf{v}$  of two vectors is  $|\mathbf{u}||\mathbf{v}| \cos \theta$ , where  $\theta$  is the angle between  $\mathbf{u}$  and  $\mathbf{v}$ .

The scalar product is frequently called the *dot product*. The scalar product distributes over addition :

$$\mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}$$

and it is clear that  $(t\mathbf{u}) \cdot \mathbf{v} = t(\mathbf{u} \cdot \mathbf{v})$  . The *vector product*  $\mathbf{u} \times \mathbf{v}$  of two vectors is defined to be the vector perpendicular to both  $\mathbf{u}$  and  $\mathbf{v}$  and having length  $|\mathbf{u}||\mathbf{v}| \sin \theta$  , where  $\theta$  is the angle between  $\mathbf{u}$  and  $\mathbf{v}$ . The direction of  $\mathbf{u} \times \mathbf{v}$  is the direction a right-hand threaded bolt advances if the vector  $\mathbf{u}$  is rotated to  $\mathbf{v}$ . The vector product is frequently called the *cross product*. The vector product is both associative and distributive, but not commutative:  $\mathbf{u} \times \mathbf{v} = -\mathbf{v} \times \mathbf{u}$  .

## 198.2 Coordinate Systems

---

Suppose we have a right-handed Cartesian coordinate system in space. For each vector  $\mathbf{u}$ , we associate a point in space by placing the tail of a representative of  $\mathbf{u}$  at the origin and associating with  $\mathbf{u}$  the point at the nose of the segment. Conversely, associated with each point in space is the vector determined by the directed segment from the origin to that point. There is thus a one-to-one correspondence between the points in space and all vectors. The origin corresponds to the zero vector. The coordinates of the point associated with a vector  $\mathbf{u}$  are called *coordinates* of  $\mathbf{u}$ . One frequently refers to the vector  $\mathbf{u}$  and writes  $\mathbf{u} = (x, y, z)$  , which is, strictly speaking, incorrect, because the left side of this equation is a vector and the right side gives the coordinates of a point in space. What is meant is that  $(x, y, z)$  are the coordinates of the point associated with  $\mathbf{u}$  under the correspondence described. In terms of coordinates, for  $\mathbf{u} = (u_1, u_2, u_3)$  and  $\mathbf{v} = (v_1, v_2, v_3)$  , we have

$$\mathbf{u} + \mathbf{v} = (u_1 + v_1, u_2 + v_2, u_3 + v_3)$$

$$t\mathbf{u} = (tu_1, tu_2, tu_3)$$

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + u_3 v_3$$

$$\mathbf{u} \times \mathbf{v} = (u_2 v_3 - v_2 u_3, u_3 v_1 - v_3 u_1, u_1 v_2 - v_1 u_2)$$

The *coordinate vectors*  $\mathbf{i}$ ,  $\mathbf{j}$ , and  $\mathbf{k}$  are the unit vectors  $\mathbf{i} = (1, 0, 0)$  ,  $\mathbf{j} = (0, 1, 0)$  , and  $\mathbf{k} = (0, 0, 1)$  . Any vector  $\mathbf{u} = (u_1, u_2, u_3)$  is thus a linear combination of these coordinate vectors:  $\mathbf{u} = u_1 \mathbf{i} + u_2 \mathbf{j} + u_3 \mathbf{k}$  . A convenient form for the vector product is the formal determinant

$$\mathbf{u} \times \mathbf{v} = \det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{bmatrix}$$



## 198.3 Vector Functions

---

A *vector function*  $\mathbf{F}$  of one variable is a rule that associates a vector  $\mathbf{F}(t)$  with each real number  $t$  in some set, called the *domain* of  $\mathbf{F}$ . The expression  $\lim_{t \rightarrow t_0} \mathbf{F}(t) = \mathbf{a}$  means that for any  $\varepsilon > 0$ , there is a  $\delta > 0$  such that  $|\mathbf{F}(t) - \mathbf{a}| < \varepsilon$  whenever  $0 < |t - t_0| < \delta$ . If  $\mathbf{F}(t) = [x(t), y(t), z(t)]$  and  $\mathbf{a} = (a_1, a_2, a_3)$ , then  $\lim_{t \rightarrow t_0} \mathbf{F}(t) = \mathbf{a}$  if and only if

$$\lim_{t \rightarrow t_0} x(t) = a_1$$

$$\lim_{t \rightarrow t_0} y(t) = a_2$$

$$\lim_{t \rightarrow t_0} z(t) = a_3$$

A vector function  $\mathbf{F}$  is *continuous* at  $t_0$  if  $\lim_{t \rightarrow t_0} \mathbf{F}(t) = \mathbf{F}(t_0)$ . The vector function  $\mathbf{F}$  is continuous at  $t_0$  if and only if each of the coordinates  $x(t)$ ,  $y(t)$ , and  $z(t)$  is continuous at  $t_0$ .

The function  $\mathbf{F}$  is *differentiable* at  $t_0$  if the limit

$$\lim_{h \rightarrow 0} \frac{1}{h} [\mathbf{F}(t+h) - \mathbf{F}(t)]$$

exists. This limit is called the *derivative* of  $\mathbf{F}$  at  $t_0$  and is usually written  $\mathbf{F}'(t_0)$ , or  $(d\mathbf{F}/dt)(t_0)$ .

The vector function  $\mathbf{F}$  is differentiable at  $t_0$  if and only if each of its coordinate functions is differentiable at  $t_0$ . Moreover,  $(d\mathbf{F}/dt)(t_0) = [(dx/dt)(t_0), (dy/dt)(t_0), (dz/dt)(t_0)]$ . The usual rules for derivatives of real valued functions all hold for vector functions. Thus if  $\mathbf{F}$  and  $\mathbf{G}$  are vector functions and  $s$  is a scalar function, then

$$\frac{d}{dt}(\mathbf{F} + \mathbf{G}) = \frac{d\mathbf{F}}{dt} + \frac{d\mathbf{G}}{dt}$$

$$\frac{d}{dt}(s\mathbf{F}) = s \frac{d\mathbf{F}}{dt} + \frac{ds}{dt} \mathbf{F}$$

$$\frac{d}{dt}(\mathbf{F} \cdot \mathbf{G}) = \mathbf{F} \cdot \frac{d\mathbf{G}}{dt} + \frac{d\mathbf{F}}{dt} \cdot \mathbf{G}$$

$$\frac{d}{dt}(\mathbf{F} \times \mathbf{G}) = \mathbf{F} \times \frac{d\mathbf{G}}{dt} + \frac{d\mathbf{F}}{dt} \times \mathbf{G}$$

If  $\mathbf{R}$  is a vector function defined for  $t$  in some interval, then, as  $t$  varies, with the tail of  $\mathbf{R}$  at the origin, the nose traces out some object  $C$  in space. For nice functions  $\mathbf{R}$ , the object  $C$  is a *curve*. If  $\mathbf{R}(t) = [x(t), y(t), z(t)]$ , then the equations

$$x = x(t)$$

$$y = y(t)$$

$$z = z(t)$$

are called *parametric equations* of  $C$ . At points where  $\mathbf{R}$  is differentiable, the derivative  $d\mathbf{R}/dt$  is a vector *tangent* to the curve. The unit vector  $\mathbf{T} = (d\mathbf{R}/dt)/|d\mathbf{R}/dt|$  is called the *unit tangent vector*. If  $\mathbf{R}$  is differentiable and if the length of the arc of curve described by  $\mathbf{R}$  between  $\mathbf{R}(a)$  and  $\mathbf{R}(t)$  is given by  $s(t)$ , then

$$\frac{ds}{dt} = \left| \frac{d\mathbf{R}}{dt} \right|$$

Thus the length  $L$  of the arc from  $\mathbf{R}(t_0)$  to  $\mathbf{R}(t_1)$  is

$$L = \int_{t_0}^{t_1} \frac{ds}{dt} dt = \int_{t_0}^{t_1} \left| \frac{d\mathbf{R}}{dt} \right| dt$$

The vector  $d\mathbf{T}/ds = (d\mathbf{T}/dt)/(ds/dt)$  is perpendicular to the unit tangent  $\mathbf{T}$ , and the number  $\kappa = |d\mathbf{T}/ds|$  is the *curvature* of  $C$ . The unit vector  $\mathbf{N} = (1/\kappa)(d\mathbf{T}/ds)$  is the *principal normal*. The vector  $\mathbf{B} = \mathbf{T} \times \mathbf{N}$  is the *binormal*, and  $d\mathbf{B}/ds = -\tau\mathbf{N}$ . The number  $\tau$  is the *torsion*. Note that  $C$  is a plane curve if and only if  $\tau$  is zero for all  $t$ .

A *vector function*  $\mathbf{F}$  of two variables is a rule that assigns a vector  $\mathbf{F}(s, t)$  to each point  $(s, t)$  in some subset of the plane, called the *domain* of  $\mathbf{F}$ . If  $\mathbf{R}(s, t)$  is defined for all  $(s, t)$  in some region  $D$  of the plane, then as the point  $(s, t)$  varies over  $D$ , with its tail at the origin, the nose of  $\mathbf{R}(s, t)$  traces out an object in space. For a nice function  $\mathbf{R}$ , this object is a *surface*,  $S$ . The partial derivatives  $(\partial\mathbf{R}/\partial s)(s, t)$  and  $(\partial\mathbf{R}/\partial t)(s, t)$  are tangent to the surface at  $\mathbf{R}(s, t)$ , and the vector  $(\partial\mathbf{R}/\partial s) \times (\partial\mathbf{R}/\partial t)$  is thus *normal* to the surface. Of course,  $(\partial\mathbf{R}/\partial t) \times (\partial\mathbf{R}/\partial s) = -(\partial\mathbf{R}/\partial s) \times (\partial\mathbf{R}/\partial t)$  is also normal to the surface and points in the direction opposite that of  $(\partial\mathbf{R}/\partial s) \times (\partial\mathbf{R}/\partial t)$ . By electing one of these normals, we are choosing an *orientation* of the surface. A surface can be oriented only if it has two sides, and the process of orientation consists of choosing which side is "positive" and which is "negative."

## 198.4 Gradient, Curl, and Divergence

If  $f(x, y, z)$  is a scalar field defined in some region  $D$ , the *gradient* of  $\mathbf{f}$  is the vector function

$$\text{grad } f = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} + \frac{\partial f}{\partial z}\mathbf{k}$$

If  $\mathbf{F}(x, y, z) = F_1(x, y, z)\mathbf{i} + F_2(x, y, z)\mathbf{j} + F_3(x, y, z)\mathbf{k}$  is a vector field defined in some region  $D$ , then the *divergence* of  $\mathbf{F}$  is the scalar function

$$\operatorname{div} \mathbf{F} = \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z}$$

The curl is the vector function

$$\operatorname{curl} \mathbf{F} = \left( \frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z} \right) \mathbf{i} + \left( \frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x} \right) \mathbf{j} + \left( \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) \mathbf{k}$$

In terms of the vector operator *del*,  $\nabla = \mathbf{i}(\partial/\partial x) + \mathbf{j}(\partial/\partial y) + \mathbf{k}(\partial/\partial z)$ , we can write

$$\operatorname{grad} f = \nabla f$$

$$\operatorname{div} \mathbf{F} = \nabla \cdot \mathbf{F}$$

$$\operatorname{curl} \mathbf{F} = \nabla \times \mathbf{F}$$

The *Laplacian operator* is  $\operatorname{div}(\operatorname{grad}) = \nabla \cdot \nabla = \nabla^2 = (\partial^2/\partial x^2) + (\partial^2/\partial y^2) + (\partial^2/\partial z^2)$ .

## 198.5 Integration

---

Suppose  $C$  is a curve from the point  $(x_0, y_0, z_0)$  to the point  $(x_1, y_1, z_1)$  and is described by the vector function  $\mathbf{R}(t)$  for  $t_0 \leq t \leq t_1$ . If  $f$  is a scalar function (sometimes called a *scalar field*) defined on  $C$ , then the integral of  $f$  over  $C$  is

$$\int_C f(x, y, z) ds = \int_{t_0}^{t_1} f[\mathbf{R}(t)] \left| \frac{d\mathbf{R}}{dt} \right| dt$$

If  $\mathbf{F}$  is a vector function (sometimes called a *vector field*) defined on  $C$ , then the integral of  $\mathbf{F}$  over  $C$  is

$$\int_C \mathbf{F}(x, y, z) \cdot d\mathbf{R} = \int_{t_0}^{t_1} \mathbf{F}[\mathbf{R}(t)] \cdot \frac{d\mathbf{R}}{dt} dt$$

These integrals are called *line integrals*.

In case there is a scalar function  $f$  such that  $\mathbf{F} = \operatorname{grad} f$ , then the line integral

$$\int_C \mathbf{F}(x, y, z) \cdot d\mathbf{R} = f[\mathbf{R}(t_1)] - f[\mathbf{R}(t_0)]$$

The value of the integral thus depends only on the end points of the curve  $C$  and not on the curve  $C$

itself. The integral is said to be *path-independent*. The function  $f$  is called a *potential function* for the vector field  $\mathbf{F}$ , and  $\mathbf{F}$  is said to be a *conservative field*. A vector field  $\mathbf{F}$  with domain  $D$  is conservative if and only if the integral of  $\mathbf{F}$  around every closed curve in  $D$  is zero. If the domain  $D$  is simply connected (that is, every closed curve in  $D$  can be continuously deformed in  $D$  to a point), then  $\mathbf{F}$  is conservative if and only if  $\text{curl } \mathbf{F} = 0$  in  $D$ .

Suppose  $S$  is a surface described by  $\mathbf{R}(s, t)$  for  $(s, t)$  in a region  $D$  of the plane. If  $f$  is a scalar function defined on  $D$ , then the integral of  $f$  over  $S$  is given by

$$\int_S f(x, y, z) dS = \int_D \int f[\mathbf{R}(s, t)] \left| \frac{\partial \mathbf{R}}{\partial s} \times \frac{\partial \mathbf{R}}{\partial t} \right| ds dt$$

If  $\mathbf{F}$  is a vector function defined on  $S$ , and if an orientation for  $S$  is chosen, then the integral of  $\mathbf{F}$  over  $S$ , sometimes called the **flux** of  $\mathbf{F}$  through  $S$ , is

$$\int_S \mathbf{F}(x, y, z) \cdot d\mathbf{S} = \int_D \int \mathbf{F}[\mathbf{R}(s, t)] \cdot \left( \frac{\partial \mathbf{R}}{\partial s} \times \frac{\partial \mathbf{R}}{\partial t} \right) ds dt$$

## 198.6 Integral Theorems

---

Suppose  $\mathbf{F}$  is a vector field with a closed domain  $D$  bounded by the surface  $S$  oriented so that the normal points out from  $D$ . Then the *divergence theorem* states that

$$\int_D \int \int \text{div } \mathbf{F} dV = \int_S \int \mathbf{F} \cdot d\mathbf{S}$$

If  $S$  is an orientable surface bounded by a closed curve  $C$ , the orientation of the closed curve  $C$  is chosen to be consistent with the orientation of the surface  $S$ . Then we have *Stokes's theorem*:

$$\int_S (\text{curl } \mathbf{F}) \cdot d\mathbf{S} = \oint_C \mathbf{F} \cdot d\mathbf{s}$$

## References

- Davis, H. F. and Snider, A. D. 1991. *Introduction to Vector Analysis*, 6th ed. Wm. C. Brown, Dubuque, IA.
- Wylie, C. R. 1975. *Advanced Engineering Mathematics*, 4th ed. McGraw-Hill, New York.

## Further Information

More advanced topics leading into the theory and applications of tensors may be found in J. G. Simmonds, *A Brief on Tensor Analysis* (1982, Springer-Verlag, New York).

Cain, G. "Complex Variables"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

[199.1 Basic Definitions and Arithmetic](#)[199.2 Complex Functions](#)[199.3 Analytic Functions](#)[199.4 Integration](#)[199.5 Series](#)[199.6 Singularities](#)[199.7 Conformal Mapping](#)**George Cain***Georgia Institute of Technology*

---

## 199.1 Basic Definitions and Arithmetic

---

A *complex number* is an ordered pair  $z = (x, y)$  of real numbers  $x$  and  $y$ . The sum, difference, product, and quotient of two complex numbers  $z = (x, y)$  and  $w = (u, v)$  are defined by

$$z \pm w = (x \pm u, y \pm v)$$

$$zw = (xu - yv, xv + yu)$$

$$\frac{w}{z} = \left( \frac{xu + yv}{x^2 + y^2}, \frac{xv - yu}{x^2 + y^2} \right)$$

The *real part* of the complex number  $z = (x, y)$  is  $x$ , and the *imaginary part* of  $z$  is  $y$ . If  $z_1 = (a, 0)$  and  $z_2 = (b, 0)$ , then  $z_1 \pm z_2 = (a \pm b, 0)$ ,  $z_1 z_2 = (ab, 0)$ , and  $z_1 / z_2 = (a/b, 0)$ . Thus, for complex numbers with 0 imaginary part, complex arithmetic coincides with the usual arithmetic for real numbers, and we see that the complex numbers are an extension of the real numbers. In this case we write simply  $a$  for the complex number  $(a, 0)$ , and so on.

For the complex number  $i = (0, 1)$  we have that  $i^2 = (-1, 0) \equiv -1$ . Now note that for any  $z = (x, y)$ , it is true that

$$z = (x, y) = (x, 0) + (0, y) = x + iy$$

and the algebra of complex numbers reduces to the usual algebra for real numbers, together with

the fact that  $i^2 = -1$  .

The *conjugate* of a complex number  $z = x + iy$  is defined to be  $\bar{z} = x - iy$  . The *modulus* of  $z$  is the real number  $|z| = \sqrt{z\bar{z}} = \sqrt{x^2 + y^2}$  . We have the following properties of the conjugate and the modulus:

$$\overline{(z \pm w)} = \bar{z} \pm \bar{w}$$

$$\overline{zw} = \bar{z}\bar{w} \quad \text{and} \quad \overline{\left(\frac{w}{z}\right)} = \frac{\bar{w}}{\bar{z}}$$

$$|z + w| \leq |z| + |w|$$

$$|zw| = |z||w| \quad \text{and} \quad \left|\frac{w}{z}\right| = \frac{|w|}{|z|}$$

Considering  $z = (x, y)$  to be the rectangular coordinates of a point in the plane establishes a one-to-one correspondence between the set of all complex numbers and the points in the plane. We thus speak of points in the plane as being complex numbers. The usual Euclidean distance  $d(z, w)$  between the points  $z$  and  $w$  is then given by  $d(z, w) = |z - w|$  . Let  $(r, \theta)$  be polar coordinates of the point  $z = (x, y)$  . (We always assume  $r \geq 0$  .) Then  $r$  is, of course, simply  $|z|$ , the modulus of  $z$ . The number  $\theta$  is called an *argument* of  $z$  and is denoted  $\arg z$ . Thus,  $z = r(\cos \theta + i \sin \theta)$  . If  $w = s(\cos \vartheta + i \sin \vartheta)$  , then

$$zw = rs[\cos(\theta + \vartheta) + i \sin(\theta + \vartheta)]$$

$$\frac{z}{w} = \frac{r}{s}[\cos(\theta - \vartheta) + i \sin(\theta - \vartheta)]$$

$$z^n = r^n[\cos n\theta + i \sin n\theta]$$

Note that a complex number has an infinite number of arguments, and  $\arg z$  denotes any one of them. Any interval  $(a, b]$  of reals of length  $2\pi$  contains exactly one argument of  $z$ . The particular value of  $\arg z$  in the interval  $(-\pi, \pi]$  is called the *principal value of the argument* and is denoted  $\text{Arg } z$ .

## 199.2 Complex Functions

---

Let  $\mathbf{C}$  denote the set of all complex numbers (the *complex plane*). A set  $\mathbf{S} \subset \mathbf{C}$  is said to be a *neighborhood* of the point  $z_0$  if there is an  $r > 0$  such that  $z \in \mathbf{S}$  whenever  $|z_0 - z| < r$  . The point  $z_0$  is an *interior point* of  $\mathbf{S}$  in case  $\mathbf{S}$  is a neighborhood of  $z_0$ . A point  $z_0$  is called a *limit point* of a

set  $\mathbf{A}$  if every neighborhood of  $z_0$  meets  $\mathbf{A}$ . A set that is a neighborhood of each of its points is called an *open set*, and a set that contains all its limit points is called a *closed set*.

A function  $f$  from a subset of the complex numbers into the complex numbers is called a *complex function*. The statement that the *limit* of  $f$  at  $z_0$  is  $w_0$  means that for each  $\varepsilon > 0$ , there is a  $\delta > 0$  such that  $|f(z) - w_0| < \varepsilon$  whenever  $z$  is in the domain of  $f$  and  $|z - z_0| < \delta$ . We write

$$\lim_{z \rightarrow z_0} f(z) = w_0$$

Note that it is not required that  $z_0$  be in the domain of  $f$ .

The function  $f$  is *continuous* at the point  $z_0$  in its domain if  $\lim_{z \rightarrow z_0} f(z) = f(z_0)$ . If  $z_0$  is an interior point of the domain of  $f$ , then  $f$  is said to be *differentiable* at  $z_0$  if the limit

$$f'(z_0) = \lim_{z \rightarrow z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

exists. The number  $f'(z_0)$  is the *derivative* of  $f$  at  $z_0$ .

A complex function of the form  $p(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$  is called a *polynomial*. A *rational function*  $r$  is the quotient of two polynomials:  $r(z) = p(z)/q(z)$ . A polynomial is continuous and differentiable at every complex number, and a rational function is continuous and differentiable everywhere except at those points  $z_0$  for which  $q(z_0) = 0$ .

## 199.3 Analytic Functions

---

A complex function  $f$  is *analytic* on an open set  $\mathbf{U}$  if it has a derivative at each point of  $\mathbf{U}$ . The function  $f$  is said to be analytic at the point  $z_0$  if it is analytic on some open set containing  $z_0$ . A function analytic on all of the complex plane is an *entire* function. If  $f(z) = u(x, y) + iv(x, y)$  is analytic in an open set  $\mathbf{U}$ , then the real and imaginary parts of  $f(z)$  satisfy the Cauchy-Riemann equations:

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \text{and} \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$$

and both  $u$  and  $v$  satisfy *Laplace's equation* in  $\mathbf{U}$ :

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

A function that satisfies Laplace's equation in a domain  $\mathbf{U}$  is said to be *harmonic* in  $\mathbf{U}$ . Thus the real and imaginary parts of a function analytic in an open set are harmonic on the same set.



## 199.4 Integration

---

An integral of a complex function  $f$  is defined on a curve (or *contour*)  $C$  in the complex plane and is simply the line integral of  $f$  on  $C$ . Thus if  $z = \gamma(t)$ ,  $a \leq t \leq b$  defines a contour  $C$  from the point  $z_0 = \gamma(a)$  to the point  $z_1 = \gamma(b)$ , the *contour integral* of  $f$  along  $C$  is given by

$$\int_C f(z) dz = \int_a^b f(\gamma(t)) \gamma'(t) dt$$

All of the usual nice linearity properties of integrals hold for complex integrals. Thus,

$$\int_C [c_1 f(z) + c_2 g(z)] dz = c_1 \int_C f(z) dz + c_2 \int_C g(z) dz$$

for any complex numbers  $c_1$  and  $c_2$ . If  $C_1$  and  $C_2$  are contours, then  $\int_{C_1+C_2} f(z) dz$  is defined by

$$\int_{C_1+C_2} f(z) dz = \int_{C_1} f(z) dz + \int_{C_2} f(z) dz$$

A contour given by  $z = \gamma(t)$ ,  $a \leq t \leq b$  is called a *simple* contour if  $\gamma(t) \neq \gamma(s)$  for  $a < t, s < b$ , and  $t \neq s$ . A *closed* contour is a contour for which  $\gamma(a) = \gamma(b)$ . If  $C$  is given by  $z = \gamma(t)$ ,  $a \leq t \leq b$ , then  $-C$  is the contour given by  $z = \gamma(a + b - t)$ ,  $a \leq t \leq b$ . It is clear that

$$\int_{-C} f(z) dz = - \int_C f(z) dz$$

Suppose  $C$  is a simple closed contour and  $f$  is a function analytic at all points on and inside  $C$ . Then

$$\int_C f(z) dz = 0$$

This is the celebrated *Cauchy-Goursat theorem*.

An open connected set  $U$  such that every simple closed contour in  $U$  encloses only points of  $U$  is said to be *simply connected*. If  $z_0$  and  $z_1$  are points in a simply connected open set  $U$ , and if  $C_1$  and  $C_2$  are contours in  $U$  from  $z_0$  and  $z_1$ , then for any  $f$  analytic in  $U$ , it is true that

$$\int_{C_1} f(z) dz = \int_{C_2} f(z) dz$$

A simple closed contour  $C$  given by  $z = \gamma(t)$ ,  $a \leq t \leq b$  is *positively oriented* if, for  $z_0$  inside  $C$ , the argument  $\text{Arg}[\gamma(t) - z_0]$  increases as  $t$  goes from  $a$  to  $b$ . If  $f$  is analytic in simply connected region  $U$ , and if  $C$  is a positively oriented simple closed curve in  $U$ , then for every  $z_0$  in  $U$  that is not on  $C$ , we have

$$f(z_0) = \frac{1}{2\pi i} \int_C \frac{f(z)}{z - z_0} dz$$

This is the celebrated *Cauchy integral formula*. If  $f$  is analytic at a point  $z_0$ , then it has derivatives of all orders, and

$$f^{(n)}(z_0) = \frac{n!}{2\pi i} \int_C \frac{f(z)}{(z - z_0)^{n+1}} dz$$

If the function  $f$  is continuous on an open connected set  $\mathbf{U}$  and if

$$\int_C f(z) dz = 0$$

for every closed contour  $C$  in  $\mathbf{U}$ , then  $f$  is analytic on  $\mathbf{U}$ . This is *Morera's theorem*.

If  $f$  is analytic on an open set  $\mathbf{U}$ , the point  $z_0$  is in  $\mathbf{U}$ , and  $C_R$  is a circle in  $\mathbf{U}$  of radius  $r$  centered at  $z_0$  on which  $|f(z)| \leq M$ , then

$$|f^{(n)}(z_0)| \leq \frac{n!M}{r^n}$$

It follows from this that a bounded entire function must be constant.

## 199.5 Series

---

Let  $f$  be a function analytic in an open set  $\mathbf{U}$ , and let  $z_0$  be a point of  $\mathbf{U}$ . Then there is a unique power series  $\sum_{k=0}^{\infty} a_k(z - z_0)^k$  such that

$$f(z) = \sum_{k=0}^{\infty} a_k(z - z_0)^k$$

for all  $z$  in some neighborhood of  $z_0$ . This series is the *Taylor series* of  $f$  at  $z_0$  and converges to  $f(z)$  for all  $z$  inside a circle  $C$ , on which is found the singularity of  $f$  closest to  $z_0$ . The radius of  $C$  is called the *radius of convergence* of the series. This series may be differentiated term by term to obtain the Taylor series for the  $n$ th derivative of  $f$  at  $z_0$ :

$$f^{(n)}(z) = \sum_{k=n}^{\infty} k(k-1) \cdots (k-n+1) a_k(z - z_0)^{k-n}$$

The radius of convergence of this series is the same as that of the Taylor series for  $f$ . It follows easily that, for each  $n = 0, 1, 2, \dots$ , we have

$$a_n = \frac{1}{n!} f^{(n)}(z_0)$$

Let  $C_0$  and  $C_1$  be two concentric circles centered at a point  $z_0$ , with  $C_0$  having a smaller radius than  $C_1$ . Suppose the function  $f$  is analytic on an open set containing the two circles and the annular region between them. Then, for each point in the annular region bounded by the circles, we have

$$f(z) = \sum_{k=0}^{\infty} a_k (z - z_0)^k + \sum_{k=1}^{\infty} b_k (z - z_0)^{-k}$$

where

$$a_k = \frac{1}{2\pi i} \int_C \frac{f(z)}{(z - z_0)^{k+1}} dz$$

$$b_k = \frac{1}{2\pi i} \int_C \frac{f(z)}{(z - z_0)^{-k+1}} dz$$

for any positively oriented simply closed contour  $C$  around the annular region. This is a *Laurent series*. It is sometimes written

$$f(z) = \sum_{k=-\infty}^{\infty} c_k (z - z_0)^k$$

with

$$c_k = \frac{1}{2\pi i} \int_C \frac{f(z)}{(z - z_0)^{k+1}} dz$$

## 199.6 Singularities

---

Suppose that  $z_0$  is a singular point of the function  $f$  and  $f$  is analytic in a neighborhood of  $z_0$ . (Except, of course, at  $z_0$ .) Then  $z_0$  is called an *isolated singular point*, or an *isolated singularity*, of  $f$ . In case  $z_0$  is an isolated singular point of  $f$ , the Laurent series

$$f(z) = \sum_{k=-\infty}^{\infty} c_k (z - z_0)^k$$

represents  $f$  for all  $z$  such that  $0 < |z - z_0| < r$ , for some  $r > 0$ . If  $c_k = 0$  for all  $k \leq -1$ , then  $z_0$

is said to be a *removable* singularity. (The value of  $f$  at  $z_0$  can, in this case, be redefined to be  $c_0$ , and the function so defined is analytic at  $z_0$ .) If there is a negative integer  $N < -1$  such that  $c_k = 0$  for all  $k \leq N$ , then  $z_0$  is a *pole* of order  $p$ , where  $p$  is the largest positive integer for which  $c_{-p} \neq 0$ . A pole of order one is called a *simple pole*. An isolated singularity that is neither a removable singularity nor a pole is called an *essential* singularity. Thus,  $z_0$  is an essential singularity if, for every negative  $N$ , there is a  $k < N$  such that  $c_k \neq 0$ .

If  $z_0$  is an isolated singularity of the function  $f$ , then the coefficient  $c_{-1}$  in the Laurent series given is the *residue* of  $f$  at  $z_0$ . If  $z_0$  is a simple pole, then the residue of  $f$  at  $z_0$  is given by

$$\text{Res}(z_0) = \lim_{z \rightarrow z_0} (z - z_0) f(z)$$

If  $z_0$  is a pole of order  $p$ , then

$$\text{Res}(z_0) = \lim_{z \rightarrow z_0} \frac{1}{(p-1)!} \frac{d^{p-1}}{dz^{p-1}} [(z - z_0)^p f(z)]$$

The importance of the idea of the residue comes from *Cauchy's residue theorem*, which says that if  $C$  is a positively oriented simple closed contour and the function  $f$  is analytic on and inside  $C$  except at the points  $z_1, z_2, \dots, z_n$ , then

$$\int_C f(z) dz = 2\pi i \sum_{k=1}^n \text{Res}(z_k)$$

## 199.7 Conformal Mapping

We can regard a one-to-one complex function  $f$  on a region  $\mathbf{D}$  as defining a *mapping* from  $\mathbf{D}$  into the plane. In case  $f$  is analytic, this mapping is *conformal*; that is, angles are preserved. More importantly, such a mapping preserves the harmonic property of a function. Thus, if  $f$  is analytic and  $w = f(z) = u(x, y) + iv(x, y)$  maps a domain  $\mathbf{D}_z$  onto a domain  $\mathbf{D}_w$ , and if

$$\frac{\partial^2 \Phi(u, v)}{\partial u^2} + \frac{\partial^2 \Phi(u, v)}{\partial v^2} = 0 \text{ in } \mathbf{D}_w$$

then

$$\frac{\partial^2 \Theta(x, y)}{\partial x^2} + \frac{\partial^2 \Theta(x, y)}{\partial y^2} = 0 \text{ in } \mathbf{D}_z$$

where  $\Theta(x, y) = \Phi[u(x, y), v(x, y)]$ . Conformal mappings are useful in solving boundary value problems involving Laplace's equation in a region of the plane. The given region is mapped conformally to another region in which the solution is either known or easy to find.

## References

- Ahlfors, L. V. 1979. *Complex Analysis*, 3rd ed. McGraw-Hill, New York.
- Churchill R. V. and Brown J. W. 1990. *Complex Variables and Applications*, 5th ed. McGraw-Hill, New York.
- Saff, E. B. and Snider, A. D. 1976. *Fundamentals of Complex Analysis for Mathematics, Science, and Engineering*. Prentice Hall, Englewood Cliffs, NJ.

## Further Information

Extensive compilations of conformal mappings are found in *Dictionary of Conformal Representations*, by H. Kober (1957, Dover Publications, New York.), and *Complex Variables and Applications*, by R. V. Churchill and J. W. Brown (1990, McGraw-Hill, New York.)

Ames, W. F. "Difference Equations"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## 200.1 First-Order Equations

## 200.2 Second-Order Equations

## 200.3 Linear Equations with Constant Coefficients

200.4 Generating Function ( $z$  Transform)**William F. Ames***Georgia Institute of Technology*

Difference equations are equations involving *discrete variables*. They appear as natural descriptions of natural phenomena and in the study of discretization methods for differential equations, which have continuous variables.

Let  $y_n = y(nh)$ , where  $n$  is an integer and  $h$  a real number. (One can think of measurements taken at equal intervals,  $h, 2h, 3h, \dots$ , and  $y_n$  describes these.) A typical equation is that describing the famous Fibonacci sequence— $y_{n+2} - y_{n+1} - y_n = 0$ . Another example is the equation  $y_{n+2} - 2zy_{n+1} + y_n = 0$ ,  $z \in C$ , which describes the Chebyshev polynomials.

**200.1 First-Order Equations**

The general first-order equation  $y_{n+1} = f(y_n)$ ,  $y_0 = y(0)$  is easily solved, for as many terms as are needed, by *iteration*. Then  $y_1 = f(y_0)$ ;  $y_2 = f(y_1)$ ,  $\dots$ . An example is the logistic equation  $y_{n+1} = ay_n(1 - y_n) = f(y_n)$ . The logistic equation has two fixed (critical or equilibrium) points where  $y_{n+1} = y_n$ . They are 0 and  $\bar{y} = (a - 1)/a$ . This has physical meaning only for  $a > 1$ . For  $1 < a < 3$  the equilibrium  $\bar{y}$  is asymptotically stable, and for  $a > 3$  there are two points  $y_1$  and  $y_2$ , called a *cycle of period two*, in which  $y_2 = f(y_1)$  and  $y_1 = f(y_2)$ . This study leads into chaos, which is outside our interest. By iteration, with  $y_0 = 1/2$ , we have  $y_1 = (a/2)(1/2) = a/2^2$ ,  $y_2 = a(a/2^2)(1 - a/2^2) = (a^2/2^2)(1 - a/2^2)$ ,  $\dots$ .

With a constant, the equation  $y_{n+1} = ay_n$  is solved by making the assumption  $y_n = A\lambda^n$  and finding  $\lambda$  so that the equation holds. Thus  $A\lambda^{n+1} = aA\lambda^n$ , and hence  $\lambda = 0$  or  $\lambda = a$  and  $A$  is arbitrary. Discarding the trivial solution 0 we find  $y_n = Aa^{n+1}$  is the desired solution. By using a method called the *variation of constants*, the equation  $y_{n+1} - ay_n = g_n$  has the solution  $y_n = y_0a^n + \sum_{j=0}^{n-1} g_ja^{n-j-1}$ , with  $y_0$  arbitrary.

In various applications we find the first-order equation of *Riccati type*  $y_n y_{n-1} + ay_n + by_{n-1} + c = 0$  where  $a$ ,  $b$ , and  $c$  are real constants. This equation can be transformed to a linear second-order equation by setting  $y_n = z_n/z_{n-1} - a$  to obtain

$z_{n+1} + (b+a)z_n + (c-ab)z_{n-1} = 0$  , which is solvable as described in the next section.

## 200.2 Second-Order Equations

The second-order linear equation with constant coefficients  $y_{n+2} + ay_{n+1} + by_n = f_n$  is solved by first solving the homogeneous equation (with right-hand side zero) and adding to that solution any solution of the inhomogeneous equation. The *homogeneous equation*  $y_{n+2} + ay_{n+1} + by_n = 0$  is solved by assuming  $y_n = \lambda^n$ , whereupon  $\lambda^{n+2} + a\lambda^{n+1} + b\lambda^n = 0$  or  $\lambda = 0$  (rejected) or  $\lambda^2 + a\lambda + b = 0$ . The roots of this quadratic are  $\lambda_1 = \frac{1}{2}(-a + \sqrt{a^2 - 4b})$ ,  $\lambda_2 = -\frac{1}{2}(a + \sqrt{a^2 - 4b})$  and the solution of the homogeneous equation is  $y_n = c_1\lambda_1^n + c_2\lambda_2^n$ . As an example consider the Fibonacci equation  $y_{n+2} - y_{n+1} - y_n = 0$ . The roots of  $\lambda^2 - \lambda - 1 = 0$  are  $\lambda_1 = \frac{1}{2}(1 + \sqrt{5})$ ,  $\lambda_2 = \frac{1}{2}(1 - \sqrt{5})$ , and the solution  $y_n = c_1[(1 + \sqrt{5})/2]^n + c_2[(1 - \sqrt{5})/2]^n$  is known as the *Fibonacci sequence*.

Many of the orthogonal polynomials of differential equations and numerical analysis satisfy a second-order difference equation (recurrence relation) involving a discrete variable, say  $n$ , and a continuous variable, say  $z$ . One such is the *Chebyshev equation*  $y_{n+2} - 2zy_{n+1} + y_n = 0$  with the initial conditions  $y_0 = 1$ ,  $y_1 = z$  (*first-kind* Chebyshev polynomials) and  $y_{-1} = 0$ ,  $y_0 = 1$  (*second-kind* Chebyshev polynomials). They are denoted  $T_n(z)$  and  $V_n(z)$ , respectively. By iteration we find

$$\begin{aligned} T_0(z) &= 1, & T_1(z) &= z, & T_2(z) &= 2z^2 - 1, \\ T_3(z) &= 4z^3 - 3z, & T_4(z) &= 8z^4 - 8z^2 + 1 \\ V_0(z) &= 0, & V_1(z) &= 1, & V_2(z) &= 2z, \\ V_3(z) &= 4z^2 - 1, & V_4(z) &= 8z^3 - 4z \end{aligned}$$

and the general solution is  $y_n(z) = c_1T_n(z) + c_2V_{n-1}(z)$ .

## 200.3 Linear Equations with Constant Coefficients

The general  $k$ th-order linear equation with constant coefficients is  $\sum_{i=0}^k p_i y_{n+k-i} = g_n$ ,  $p_0 = 1$ . The solution to the corresponding homogeneous equation (obtained by setting  $g_n = 0$ ) is as follows. (a)  $y_n = \sum_{i=1}^k c_i \lambda_i^n$  if the  $\lambda_i$  are the distinct roots of the characteristic polynomial  $p(\lambda) = \sum_{i=0}^k p_i \lambda^{k-i} = 0$ . (b) If  $m_s$  is the multiplicity of the root  $\lambda_s$ , then the functions  $y_{n,s} = u_s(n)\lambda_s^n$ , where  $u_s(n)$  are polynomials in  $n$  whose degree does not exceed  $m_s - 1$ , are solutions of the equation. Then the general solution of the homogeneous equation is  $y_n = \sum_{i=1}^d a_i u_i(n)\lambda_i^n = \sum_{i=1}^d a_i \sum_{j=0}^{m_i-1} c_j n^j \lambda_i^n$ . To this solution one adds any particular solution to obtain the general solution of the general equation.

### Example 200.1

A model equation for the price  $p_n$  of a product, at the  $n$ th time, is



$p_n + \frac{b}{a}(1 + \rho)p_{n-1} - \frac{b}{a}\rho p_{n-2} + (s_0 - d_0)/a = 0$  . The equilibrium price is obtained by setting  $p_n = p_{n-1} = p_{n-2} = p_e$  , and one finds  $p_e = (d_0 - s_0)/(a + b)$  . The homogeneous equation has the characteristic polynomial  $\lambda^2 + (b/a)(1 + \rho)\lambda - (b/a)\rho = 0$  . With  $\lambda_1$  and  $\lambda_2$  as the roots the general solution of the full equation is  $p_n = c_1\lambda_1^n + c_2\lambda_2^n + p_e$  , since  $p_e$  is a solution of the full equation. This is one method for finding the solution of the nonhomogeneous equation.

## 200.4 Generating Function (z Transform)

An elegant way of solving linear difference equations with constant coefficients, among other applications, is by use of *generating functions* or, as an alternative, the  $z$  transform. The generating function of a sequence  $\{y_n\}$  ,  $n = 0, 1, 2, \dots$  , is the function  $f(x)$  given by the formal series  $f(x) = \sum_{n=0}^{\infty} y_n x^n$  . The  $z$  transform of the same sequence is  $z(x) = \sum_{n=0}^{\infty} y_n x^{-n}$  . Clearly,  $z(x) = f(1/x)$  . A table of some important sequences is given in [Table 200.1](#)

**Table 200.1** Important Sequences

$y_n$	$f(x)$	Convergence Domain
1	$(1 - x)^{-1}$	$ x  < 1$
$n$	$x(1 - x)^{-2}$	$ x  < 1$
$n^m$	$x p_m(x)(1 - x)^{-m-1}$	$ x  < 1$
$k^n$	$(1 - kx)^{-1}$	$ x  < k^{-1}$
$e^{an}$	$(1 - e^a x)^{-1}$	$ x  < e^{-a}$
$k^n \cos an$	$\frac{1 - kx \cos a}{1 - 2kx \cos a + k^2 x^2}$	$ x  < k^{-1}$
$k^n \sin an$	$\frac{1 - kx \sin a}{1 - 2kx \cos a + k^2 x^2}$	$ x  < k^{-1}$
$\binom{n}{m}$	$x^m (1 - x)^{-m-1}$	$ x  < 1$
$\binom{k}{n}$	$(1 + x)^k$	$ x  < 1$

\*The term  $P_m(z)$  is a polynomial of degree  $m$  satisfying  $P_{m+1}(z) = (mz + 1)p_m(z) + z(1 - z)p'_m(z)$ ,  $p_1 = 1$

To solve the linear difference equation  $\sum_{i=0}^k p_i y_{n+k-i} = 0$ ,  $p_0 = 1$  we associate with it the two formal series  $P = p_0 + p_1 x + \cdots + p_k x^k$  and  $Y = y_0 + y_1 x + y_2 x^2 + \cdots$ . If  $p(x)$  is the characteristic polynomial then  $P(x) = x^k p(1/x) = \bar{p}(x)$ . The product of the two series is  $Q = YP = q_0 + q_1 x + \cdots + q_{k-1} x^{k-1} + q_k x^k + \cdots$  where  $q_n = \sum_{i=0}^n p_i y_{n-i}$ . Because  $p_{k+1} = p_{k+2} = \cdots = 0$ , it is obvious that  $q_{k+1} = q_{k+2} = \cdots = 0$ —that is,  $Q$  is a polynomial (formal series with finite number of terms). Then  $Y = P^{-1}Q = q(x)/\bar{p}(x) = q(x)/x^k p(1/x)$ , where  $p$  is the characteristic polynomial and  $q(x) = \sum_{i=0}^k q_i x^i$ . The roots of  $\bar{p}(x)$  are  $x_i^{-1}$  where the  $x_i$  are the roots of  $p(x)$ .

**Theorem 1.** If the roots of  $p(x)$  are less than one in absolute value, then  $Y(x)$  converges for  $|x| < 1$ .

**Theorem 2.** If  $p(x)$  has no roots greater than one in absolute value and those on the unit circle are simple roots, then the coefficients  $y_n$  of  $Y$  are bounded. Now  $q_k = g_0$ ,  $q_{n+k} = g_n$ , and  $Q(x) = Q_1(x) + x^k Q_2(x)$ . Hence  $\sum_{i=1}^{\infty} y_i x^i = [Q_1(x) + x^k Q_2(x)]/\bar{p}(x)$ .

### Example 200.2

Consider the equation  $y_{n+1} + y_n = -(n+1)$ ,  $y_0 = 1$ . Here  $Q_1 = 1$ ,  $Q_2 = -\sum_{n=0}^{\infty} (n+1)x^n = -1/(1-x)^2$ .

$$G(x) = \frac{1 - x/(1-x)^2}{1+x} = \frac{5}{4} \frac{1}{1+x} - \frac{1}{4} \frac{1}{1-x} - \frac{1}{2} \frac{x}{(1-x)^2}.$$

Using the table term by term, we find  $\sum_{n=0}^{\infty} y_n x^n = \sum_{n=0}^{\infty} \left[ \frac{5}{4}(-1)^n - \frac{1}{4} - \frac{1}{2}n \right] x^n$ , so  $y_n = \frac{5}{4}(-1)^n - \frac{1}{4} - \frac{1}{2}n$ .

### References

- Fort, T. 1948. *Finite Differences and Difference Equations in the Real Domain*. Oxford University Press, London.
- Jordan, C. 1950. *Calculus of Finite Differences*. Chelsea, New York.
- Jury, E. I. 1964. *Theory and Applications of the Z Transform Method*. John Wiley & Sons, New York.
- Lakshmikantham V. and Trigrante, D. 1988. *Theory of Difference Equations*. Academic Press, Boston, MA.
- Levy, H. and Lessman, F. 1961. *Finite Difference Equations*. Macmillan, New York.
- Miller, K. S. 1968. *Linear Difference Equations*. Benjamin, New York.
- Wilf, W. S. 1994. *Generatingfunctionology*, 2nd ed. Academic Press, Boston, MA.

Ames, W. F. "Differential Equations"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

**201.1 Ordinary Differential Equations**

First-Order Equations • Second-Order Equations • Second-Order Inhomogeneous Equations • Series Solution

**201.2 Partial Differential Equations**

Methods of Solution

**William F. Ames***Georgia Institute of Technology*

Any equation involving derivatives is called a *differential equation*. If there is only one independent variable the equation is termed a *total differential equation* or an *ordinary differential equation*. If there is more than one independent variable the equation is called a *partial differential equation*. If the highest-order derivative is the  $n$ th then the equation is said to be  $n$ th order. If there is no function of the dependent variable and its derivatives other than the linear one, the equation is said to be *linear*. Otherwise, it is *nonlinear*. Thus  $(d^3y/dx^3) + a(dy/dx) + by = 0$  is a *linear* third-order ordinary (total) differential equation. If we replace  $by$  with  $by^3$ , the equation becomes nonlinear. An example of a second-order linear partial differential equation is the famous wave equation  $(\partial^2 u / \partial x^2) - a^2(\partial^2 u / \partial t^2) = f(x)$ . There are two independent variables  $x$  and  $t$  and  $a^2 > 0$  (of course). If we replace  $f(x)$  by  $f(u)$  (say  $u^3$  or  $\sin u$ ) the equation is nonlinear. Another example of a nonlinear third-order partial differential equation is  $u_t + uu_x = au_{xxx}$ . This chapter uses the common subscript notation to indicate the partial derivatives.

Now we briefly indicate some methods of solution and the solution of some commonly occurring equations.

---

**201.1 Ordinary Differential Equations**

---

**First-Order Equations**

The *general* first-order equation is  $f(x, y, y') = 0$ . Equations capable of being written in either of the forms  $y' = f(x)g(y)$  or  $f(x)g(y)y' + F(x)G(y) = 0$  are *separable* equations. Their solution is obtained by using  $y' = dy/dx$  and writing the equations in differential form as  $dy/g(y) = f(x)dx$  or  $g(y)[dy/G(y)] = -F(x)[dx/f(x)]$  and integrating. An example is the famous *logistic* equation of inhibited growth  $(dy/dt) = ay(1 - y)$ . The integral of  $dy/y(1 - y) = a dt$  is  $y = 1/[1 + (y_0^{-1} - 1)e^{-at}]$  for  $t \geq 0$  and  $y(0) = y_0$  (the initial state called

the *initial condition*).

Equations may not have unique solutions. An example is  $y' = 2y^{1/2}$  with the initial condition  $y(0) = 0$ . One solution by separation is  $y = x^2$ . But there are an *infinity* of others—namely,  $y_a(x) = 0$  for  $-\infty < x \leq a$ , and  $(x - a)^2$  for  $a \leq x < \infty$ .

If the equation  $P(x, y) dy + Q(x, y) dx$  is reducible to

$$\frac{dy}{dx} = f\left(\frac{y}{x}\right) \quad \text{or} \quad \frac{dy}{dx} = f\left(\frac{a_1x + b_1y + c_1}{a_2x + b_2y + c_2}\right)$$

the equation is called *homogeneous* (nearly homogeneous). The first form reduces to the separable equation  $u + x(du/dx) = f(u)$  with the substitution  $y/x = u$ . The nearly homogeneous equation is handled by setting  $x = X + \alpha$ ,  $y = Y + \beta$ , and choosing **a** and **b** so that  $a_1\alpha + b_1\beta + c_1 = 0$

and  $a_2\alpha + b_2\beta + c_2 = 0$ . If  $\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} \neq 0$  this is always possible; the equation becomes

$dY/dX = [a_1 + b_1(Y/X)]/[a_2 + b_2(Y/X)]$  and the substitution  $Y = Xu$  gives a separable equation. If  $\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = 0$  then  $a_2x + b_2y = k(a_1x + b_1y)$  and the equation becomes

$du/dx = a_1 + b_1(u + c_1)/(ku + c_2)$ , with  $u = a_1x + b_1y$ . Lastly, any equation of the form  $dy/dx = f(ax + by + c)$  transforms into the separable equation  $du/dx = a + bf(u)$  using the change of variable  $u = ax + by + c$ .

The general first-order linear equation is expressible in the form  $y' + f(x)y = g(x)$ . It has the *general solution* (a solution with an arbitrary constant  $c$ )

$$y(x) = \exp\left[-\int f(x) dx\right] \left\{c + \int \exp[f(x) dx]g(x) dx\right\}$$

Two noteworthy examples of first-order equations are as follows:

1. An often-occurring nonlinear equation is the *Bernoulli equation*,  $y' + p(x)y = g(x)y^\alpha$ , with **a** real,  $\alpha \neq 0$ ,  $\alpha \neq 1$ . The transformation  $z = y^{1-\alpha}$  converts the equation to the linear first-order equation  $z' + (1 - \alpha)p(x)z = (1 - \alpha)g(x)$ .
2. The famous *Riccati equation*,  $y' = p(x)y^2 + q(x)y + r(x)$ , cannot in general be solved by integration. But some useful transformations are helpful. The substitution  $y = y_1 + u$  leads to the equation  $u' - (2py_1 + q)u = pu^2$ , which is a Bernoulli equation for  $u$ . The substitution  $y = y_1 + v^{-1}$  leads to the equation  $v' + (2py_1 + q)v + p = 0$ , which is a linear first-order equation for  $v$ . Once either of these equations has been solved, the general solution of the Riccati equation is  $y = y_1 + u$  or  $y = y_1 + v^{-1}$ .

## Second-Order Equations

The simplest of the second-order equations is  $y'' + ay' + by = 0$  ( $a, b$  real), with the initial conditions  $y(x_0) = y_0$ ,  $y'(x_0) = y'_0$  or the boundary conditions  $y(x_0) = y_0$ ,  $y(x_1) = y_1$ . The

general solution of the equation is given as follows.

1.  $a^2 - 4b > 0$ ,  $\lambda_1 = \frac{1}{2}(-a + \sqrt{a^2 - 4b})$ ,  $\lambda_2 = \frac{1}{2}(-a - \sqrt{a^2 - 4b})$   
 $y = c_1 \exp(\lambda_1 x) + c_2 \exp(\lambda_2 x)$
2.  $a^2 - 4b = 0$ ,  $\lambda_1 = \lambda_2 = -\frac{a}{2}$ ,  $y = (c_1 + c_2 x) \exp(\lambda_1 x)$
3.  $a^2 - 4b < 0$ ,  $\lambda_1 = \frac{1}{2}(-a + i\sqrt{4b - a^2})$ ,  $\lambda_2 = \frac{1}{2}(-a - i\sqrt{4b - a^2})$ ,  $i^2 = -1$   
 With  $p = -a/2$  and  $q = \frac{1}{2}\sqrt{4b - a^2}$ ,

$$y = c_1 \exp[(p + iq)x] + c_2 \exp[(p - iq)x] = \exp(px)[A \sin qx + B \cos qx]$$

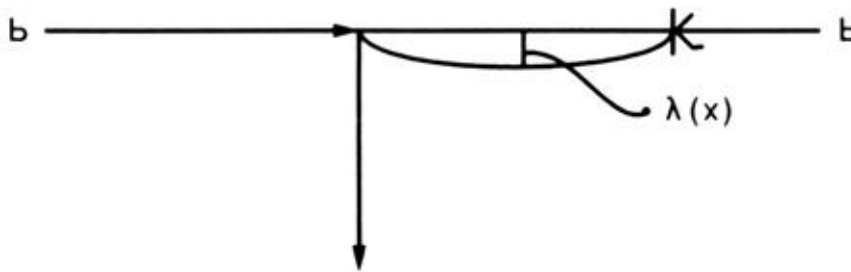
The initial conditions or boundary conditions are used to evaluate the arbitrary constants  $c_1$  and  $c_2$  (or  $A$  and  $B$ ).

Note that a linear problem with specified data may not have a solution. This is especially serious if numerical methods are employed without serious thought.

For example, consider  $y'' + y = 0$  with the boundary condition  $y(0) = 1$  and  $y(\pi) = 1$ . The general solution is  $y = c_1 \sin x + c_2 \cos x$ . The first condition  $y(0) = 1$  gives  $c_2 = 1$ , and the second condition requires  $y(\pi) = c_1 \sin \pi + \cos \pi$  or " $1 = -1$ " which is a *contradiction*.

**Example 201.1—The Euler Strut.** When a strut of uniform construction is subjected to a compressive load  $P$  it exhibits no transverse displacement until  $P$  exceeds some critical value  $P_1$ . When this load is exceeded, buckling occurs and large deflections are produced as a result of small load changes. Let the rod of length  $\ell$  be placed as shown in Fig. 201.1.

**Figure 201.1**



From the linear theory of elasticity (Timoshenko), the transverse displacement  $y(x)$  satisfies the linear second-order equation  $y'' = -Py/EI$ , where  $E$  is the modulus of elasticity and  $I$  is the moment of inertia of the strut. The boundary conditions are  $y(0) = 0$  and  $y(a) = 0$ . With  $k^2 = P/EI$  the general solution is  $y = c_1 \sin kx + c_2 \cos kx$ . The condition  $y(0) = 0$  gives  $c_2 = 0$ . The second condition gives  $c_1 \sin ka = 0$ . Since  $c_1 = 0$  gives only the trivial solution  $y = 0$  we must have  $\sin ka = 0$ . This occurs for  $ka = n\pi$ ,  $n = 0, 1, 2, \dots$  (these are called *eigenvalues*). The first nontrivial solution occurs for  $n = 1$ —that is,  $k = \pi/a$ —whereupon  $y_1 = c_1 \sin(\pi x/a)$ , with arbitrary  $c_1$ . Since  $P = EI k^2$  the critical compressive load is  $P_1 = EI\pi^2/a^2$ . This is the buckling load. The weakness of the linear theory is its failure to model

the situation when buckling occurs.

**Example 201.234Some Solvable Nonlinear Equations.** Many physical phenomena are modeled using nonlinear second-order equations. Some general cases are given here.

1.  $y'' = f(y)$  , first integral  $(y')^2 = 2 \int f(y) dy + c$  .
2.  $f(x, y', y'')$  . Set  $p = y'$  and obtain a first-order equation  $f(x, p, dp/dx) = 0$  . Use first-order methods.
3.  $f(y, y', y'') = 0$  . Set  $p = y'$  and then  $y'' = p(dp/dy)$  so that a first-order equation  $f[y, p, p(dp/dy)] = 0$  for  $p$  as a function of  $y$  is obtained.
4. The *Riccati transformation*  $du/dx = yu$  leads to the Riccati chain of equations, which linearize by raising the order. Thus,

	Equation in $y$	Equation in $u$
1.	$y' + y^2 = f(x)$	$u'' = f(x)u$
2.	$y'' + 3yy' + y^3 = f(x)$	$u''' = f(x)u$
3.	$y''' + 6y^2y' + 3(y')^2 + 4yy'' = f(x)$	$u^{(iv)} = f(x)u$

This method can be generalized to  $u' = a(x)yu$  or  $u' = a(x)f(u)y$  .

## Second-Order Inhomogeneous Equations

The general solution of  $a_0(x)y'' + a_1(x)y' + a_2(x)y = f(x)$  is  $y = y_H(x) + y_P(x)$  , where  $y_H(x)$  is the general solution of the homogeneous equation (with the right-hand side zero) and  $y_P$  is the particular integral of the equation. Construction of particular integrals can sometimes be done by the *method of undetermined coefficients*. See Table 201.1. This applies only to the linear constant coefficient case in which the function  $f(x)$  is a linear combination of a polynomial, exponentials, sines and cosines, and some products of these functions. This method has as its basis the observation that repeated differentiation of such functions gives rise to similar functions.

**Table 201.1** Method of Undetermined Coefficients—Equation  $L(y) = f(x)$  (Constant Coefficients)

	Terms in $f(x)$	Terms To Be Included in $y_P(x)$
1.	Polynomial of degree $n$	(i) If $L(y)$ contains $y$ , try $y_P = a_0x^n + a_1x^{n-1} + \cdots + a_n$ . (ii) If $L(y)$ does not contain $y$ and lowest-order derivative is $y^{(r)}$ , try $y_P = a_0x^{n+r} + \cdots + a_nx^r$ .
2.	$\sin qx, \cos qx$	(i) $\sin qx$ and/or $\cos qx$ are not in $y_H$ ; $y_P = B \sin qx + C \cos qx$ . (ii) $y_H$ contains terms of form $x^r \sin qx$ and/or $x^r \cos qx$ for $r = 0, 1, \dots, m$ ; include in $y_P$ terms of the form $a_0x^{m+1} \sin qx + a_1x^{m+1} \cos qx$ .
3.	$e^{ax}$	(i) $y_H$ does not contain $e^{ax}$ ; include $Ae^{ax}$ in $y_P$ . (ii) $y_H$ contains $e^{ax}, xe^{ax}, \dots, x^ne^{ax}$ ; include in $y_P$ terms of

4.  $e^{px} \sin qx, e^{px} \cos qx$
- the form  $Ax^{n+1}e^{ax}$ .
- (i)  $y_H$  does not contain these terms; in  $y_P$  include  $Ae^{px} \sin qx + Be^{px} \cos qx$ .
- (ii)  $y_H$  contains  $x^r e^{px} \sin qx$  and/or  $x^r e^{px} \cos qx$ ;  $r = 0, 1, \dots, m$  include in  $y_P$   $Ax^{m+1}e^{px} \sin qx + Bx^{m+1}e^{px} \cos qx$ .
- 

**Example 201.3.** Consider the equation  $y'' + 3y' + 2y = \sin 2x$ . The characteristic equation of the homogeneous equation  $\lambda^2 + 3\lambda + 2 = 0$  has the two roots  $\lambda_1 = -1$  and  $\lambda_2 = -2$ . Consequently,  $y_H = c_1 e^{-x} + c_2 e^{-2x}$ . Since  $\sin 2x$  is not linearly dependent on the exponentials and since  $\sin 2x$  repeats after two differentiations, we assume a particular solution with undetermined coefficients of the form  $y_P(x) = B \sin 2x + C \cos 2x$ . Substituting into the original equation gives  $-(2B + 6C) \sin 2x + (6B - 2C) \cos 2x = \sin 2x$ . Consequently,  $-(2B + 6C) = 1$  and  $6B - 2C = 0$  to satisfy the equation. These two equations in two unknowns have the solution  $B = -\frac{1}{20}$  and  $C = -\frac{3}{20}$ . Hence  $y_P = -\frac{1}{20}(\sin 2x + 3 \cos 2x)$  and  $y = c_1 e^{-x} + c_2 e^{-2x} - \frac{1}{20}(\sin 2x + 3 \cos 2x)$ .

A general method for finding  $y_P(x)$  called *variation of parameters* uses as its starting point  $y_H(x)$ . This method applies to *all* linear differential equations irrespective of whether they have constant coefficients. But it assumes  $y_H(x)$  is known. We illustrate the idea for  $a(x)y'' + b(x)y' + c(x)y = f(x)$ . If the solution of the homogeneous equation is  $y_H(x) = c_1 \phi_1(x) + c_2 \phi_2(x)$ , then vary the parameters  $c_1$  and  $c_2$  to seek  $y_P(x)$  as  $y_P(x) = u_1(x)\phi_1(x) + u_2(x)\phi_2(x)$ . Then  $y'_P = u_1\phi'_1 + u_2\phi'_2 + u'_1\phi_1 + u'_2\phi_2$  and choose  $u'_1\phi_1 + u'_2\phi_2 = 0$ . Calculating  $y''_P$  and setting in the original equation gives  $a(x)u'_1\phi'_1 + a(x)u'_2\phi'_2 = f$ . Solving the last two equations for  $u'_1$  and  $u'_2$  gives  $u'_1 = -\phi_2 f/wa$ ,  $u'_2 = \phi_1 f/wa$ , where  $w = \phi_1\phi'_2 - \phi'_1\phi_2 \neq 0$ . Integrating the general solution gives  $y = c_1\phi_1(x) + c_2\phi_2(x) - \left\{ \int [\phi_2 f(x)]/wa \right\} \phi_1(x) + \left[ \int (\phi_1 f/wa) dx \right] \phi_2(x)$ .

**Example 201.4.** Consider the equations  $y'' - 4y = \sin x/(1 + x^2)$  and  $y_H = c_1 e^{2x} + c_2 e^{-2x}$ . With  $\phi_1 = e^{2x}$  and  $\phi_2 = e^{-2x}$ ,  $w = 4$ , so the general solution is

$$y = c_1 e^{2x} + c_2 e^{-2x} - \frac{e^{-2x}}{4} \int \frac{e^{2x} \sin x}{1 + x^2} dx + \frac{e^{2x}}{4} \int \frac{e^{-2x} \sin x}{1 + x^2} dx$$

The method of variation of parameters can be generalized as described in the references.

Higher-order systems of linear equations with constant coefficients are treated in a similar manner. Details can be found in the references.

## Series Solution

The solution of differential equations can only be obtained in closed form in special cases. For all others, series or approximate or numerical solutions are necessary. In the simplest case, for an initial value problem, the solution can be developed as a Taylor series expansion about the point



where the initial data are specified. The method fails in the *singular case*<sup>3/4</sup> that is, a point where the coefficient of the highest-order derivative is zero. The general method of approach is called the *Frobenius method*.

To understand the nonsingular case consider the equation  $y'' + xy = x^2$  with  $y(2) = 1$  and  $y'(2) = 2$  (an initial value problem). We seek a series solution of the form  $y(x) = a_0 + a_1(x - 2) + a_2(x - 2)^2 + \cdots$ . To proceed, set  $1 = y(2) = a_0$ , which evaluates  $a_0$ . Next,  $y'(x) = a_1 + 2a_2(x - 2) + \cdots$ , so  $2 = y'(2) = a_1$  or  $a_1 = 2$ . Next  $y''(x) = 2a_2 + 6a_3(x - 2) + \cdots$  and from the equation,  $y'' = x^2 - xy$ , so  $y''(2) = 4 - 2y(2) = 4 - 2 = 2$ . Hence  $2 = 2a_2$  or  $a_2 = 1$ . Thus, to third-order  $y(x) = 1 + 2(x - 2) + (x - 2)^2 + R_2(x)$ , where the remainder  $R_2(x) = [(x - 2)^3/3!] y'''(\xi)$ , where  $2 < \xi < x$  can be bounded for each  $x$  by finding the maximum of  $y'''(x) = 2x - y - xy'$ . The third term of the series follows by evaluating  $y'''(2) = 4 - 1 - 2 \cdot 2 = -1$ , so  $6a_3 = -1$  or  $a_3 = -1/6$ .

By now the nonsingular process should be familiar. The algorithm for constructing a series solution about a nonsingular (ordinary) point  $x_0$  of the equation  $P(x)y'' + Q(x)y' + R(x)y = f(x)$  (note that  $P(x_0) \neq 0$ ) is as follows:

1. Substitute into the differential equation the expressions

$$y(x) = \sum_{n=0}^{\infty} a_n(x - x_0)^n, \quad y'(x) = \sum_{n=1}^{\infty} n a_n(x - x_0)^{n-1},$$

$$y''(x) = \sum_{n=2}^{\infty} n(n-1)a_n(x - x_0)^{n-2}$$

2. Expand  $P(x)$ ,  $Q(x)$ ,  $R(x)$ , and  $f(x)$  about the point  $x_0$  in a power series in  $(x - x_0)$  and substitute these series into the equation.
3. Gather all terms involving the same power of  $(x - x_0)$  to arrive at an identity of the form  $\sum_{n=0}^{\infty} A_n(x - x_0)^n \equiv 0$ .
4. Equate to zero each coefficient  $A_n$  of step 3.
5. Use the expressions of step 4 to determine  $a_2, a_3, \dots$  in terms of  $a_0, a_1$  (we need two arbitrary constants) to arrive at the general solution.
6. With the given initial conditions, determine  $a_0$  and  $a_1$ .

If the equation has a regular singular point—that is, a point  $x_0$  at which  $P(x)$  vanishes and a series expansion is sought about that point—a solution is sought of the form  $y(x) = (x - x_0)^r \sum_{n=0}^{\infty} a_n(x - x_0)^n$ ,  $a_0 \neq 0$  and the index  $r$  and coefficients  $a_n$  must be determined from the equation by an algorithm analogous to that already described. The description of this Frobenius method is left for the references.

## 201.2 Partial Differential Equations

The study of partial differential equations is of continuing interest in applications. It is a vast subject, so the focus in this chapter will be on the most commonly occurring equations in the engineering literature—the second-order equations in two variables. Most of these are of the three basic types: elliptic, hyperbolic, and parabolic.

*Elliptic equations* are often called *potential equations* since they occur in potential problems where the potential may be temperature, voltage, and so forth. They also give rise to the steady solutions of parabolic equations. They require boundary conditions for the complete determination of their solution.

*Hyperbolic equations* are often called *wave equations* since they arise in the propagation of waves. For the development of their solutions, initial and boundary conditions are required. In principle they are solvable by the method of characteristics.

*Parabolic equations* are usually called *diffusion equations* because they occur in the transfer (diffusion) of heat and chemicals. These equations require initial conditions (for example, the initial temperature) and boundary conditions for the determination of their solutions.

Partial differential equations (PDEs) of the second order in two independent variables  $(x, y)$  are of the form  $a(x, y)u_{xx} + b(x, y)u_{xy} + c(x, y)u_{yy} = E(x, y, u, u_x, u_y)$ . If  $E = E(x, y)$  the equation is linear; if  $E$  depends also on  $u$ ,  $u_x$ , and  $u_y$ , it is said to be *quasilinear*, and if  $E$  depends only on  $x$ ,  $y$ , and  $u$ , it is *semilinear*. Such equations are classified as follows: If  $b^2 - 4ac$  is less than, equal to, or greater than zero at some point  $(x, y)$ , then the equation is elliptic, parabolic, or hyperbolic, respectively, at that point. A PDE of this form can be transformed into canonical (standard) forms by use of new variables. These standard forms are most useful in analysis and numerical computations.

For hyperbolic equations the standard form is  $u_{\xi\eta} = \phi(u, u_\eta, u_\xi, \eta, \xi)$ , where  $\xi_x/\xi_y = (-b + \sqrt{b^2 - 4ac})/2a$ , and  $\eta_x/\eta_y = (-b - \sqrt{b^2 - 4ac})/2a$ . The right-hand sides of these equations determine the so-called characteristics  $(dy/dx)|_+ = (-b + \sqrt{b^2 - 4ac})/2a$ ,  $(dy/dx)|_- = (-b - \sqrt{b^2 - 4ac})/2a$ .

**Example 201.5.** Consider the equation  $y^2 u_{xx} - x^2 u_{yy} = 0$ .  $\xi_x/\xi_y = -x/y$ ,  $\eta_x/\eta_y = x/y$ , so  $\xi = y^2 - x^2$  and  $\eta = y^2 + x^2$ . In these new variables the equation becomes  $u_{\xi\eta} = (\xi u_\eta - \eta u_\xi)/2(\xi^2 - \eta^2)$ .

For parabolic equations the standard form is  $u_{\xi\xi} = \phi(u, u_\eta, u_\xi, \eta, \xi)$  or  $u_{\eta\eta} = \phi(u, u_\eta, u_\xi, \xi, \eta)$ , depending upon how the variables are defined. In this case  $\xi_x/\xi_y = -b/2a$  if  $a \neq 0$ , and  $\xi_x/\xi_y = -b/2c$  if  $c \neq 0$ . Only  $\mathbf{x}$  must be determined (there is only one characteristic) and  $\mathbf{h}$  can be chosen as any function that is linearly independent of  $\mathbf{x}$ .

**Example 201.6.** Consider the equation  $y^2 u_{xx} - 2xy u_{xy} + x^2 u_{yy} + u_y = 0$ . Clearly,  $b^2 - 4ac = 0$ . Neither  $a$  nor  $c$  is zero so either path can be chosen. With  $\xi_x/\xi_y = -b/2a = x/y$ , there results  $\xi = x^2 + y^2$ . With  $\eta = x$ , the equation becomes  $u_{\eta\eta} = [2(\xi + \eta)u_\xi + u_\eta]/(\xi - \eta^2)$ .

For *elliptic equations* the standard form is  $u_{\alpha\alpha} + u_{\beta\beta} = \phi(u, u_\alpha, u_\beta, \alpha, \beta)$ , where  $\mathbf{x}$  and  $\mathbf{h}$  are determined by solving the  $\mathbf{x}$  and  $\mathbf{h}$  equations of the hyperbolic system (they are complex) and taking  $\alpha = (\eta + \xi)/2$ ,  $\beta = (\eta - \xi)/2i$  ( $i^2 = -1$ ). Since  $\mathbf{x}$  and  $\mathbf{h}$  are complex conjugates, both  $\mathbf{a}$  and  $\mathbf{b}$  are real.

**Example 201.7.** Consider the equation  $y^2 u_{xx} + x^2 u_{yy} = 0$ . Clearly,  $b^2 - 4ac < 0$ , so the equation is elliptic. Then  $\xi_x/\xi_y = -ix/y$ ,  $\eta_x/\eta_y = ix/y$ , so  $\alpha = (\eta + \xi)/2 = y^2$  and  $\beta = (\eta - \xi)/2i = x^2$ . The standard form is  $u_{\alpha\alpha} + u_{\beta\beta} = -(u_\alpha/2\alpha + u_\beta/2\beta)$ .

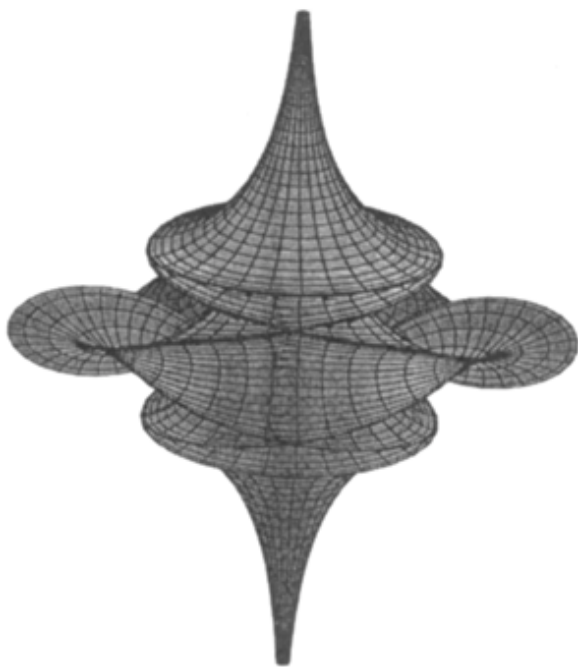


Figure 1

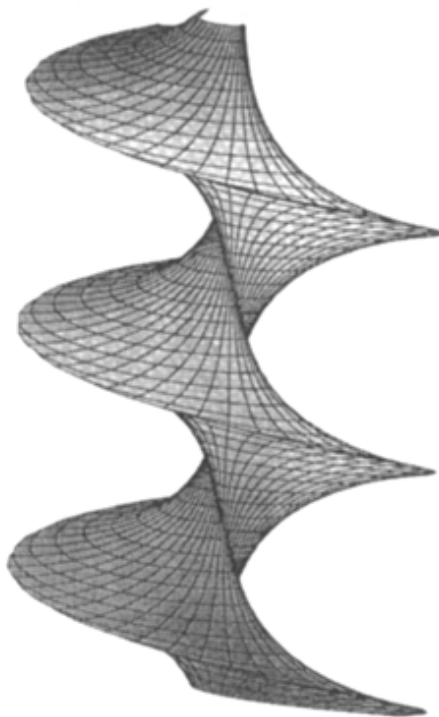


Figure 2

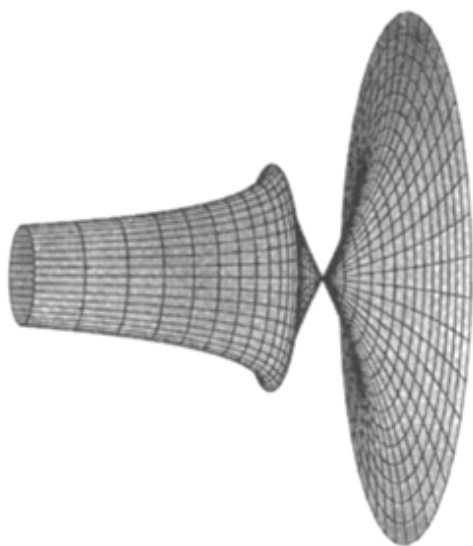


Figure 3

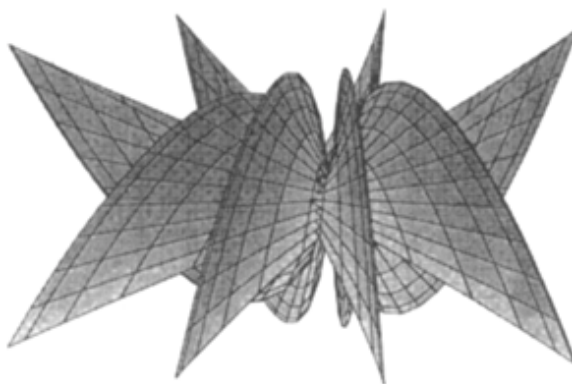


Figure 4

The mathematical equations used to generate these three-dimensional figures are worth a thousand words. The figures shown illustrate some of the nonlinear ideas of engineering, applied physics, and chemistry. [Figure 1](#) represents a breather soliton surface for the sine-Gordon equation  $w_{uv} = \sin w$  generated by a Backlund transformation. A single-soliton surface for the sine-Gordon equation  $w_{uv} = \sin w$  is illustrated in [Fig. 2](#). [Figure 3](#) represents a single-soliton surface for the Tzitzeica-Dodd-Bullough equation associated with an integrable anisotropic gas dynamics system. [Figure 4](#) represents a single-soliton Bianchi surface.

The solutions to the equations were developed by W. K. Schief and C. Rogers at the Center for Dynamical Systems and Nonlinear Studies at the Georgia Institute of Technology and the University of New South Wales in Sydney, Australia. All of these three-dimensional projections were generated using the MAPLE software package. (Figures courtesy of Schief and Rogers.)

## Methods of Solution

### Separation of Variables

Perhaps the most elementary method for solving linear PDEs with homogeneous boundary conditions is the method of *separation of variables*. To illustrate, consider  $u_t - u_{xx} = 0$ ,  $u(x, 0) = f(x)$  (the initial condition) and  $u(0, t) = u(1, t) = 0$  for  $t > 0$  (the boundary conditions). A solution is assumed in "separated form"  $u(x, t) = X(x)T(t)$ . Upon substituting into the equation we find  $\dot{T}/T = X''/X$  (where  $\dot{T} = dT/dt$  and  $X'' = d^2X/dx^2$ ). Since  $T = T(t)$  and  $X = X(x)$ , the ratio must be constant, and for finiteness in  $t$  the constant must be negative, say  $-\lambda^2$ . The solutions of the separated equations  $X'' + \lambda^2 X = 0$  with the boundary conditions  $X(0) = 0$ ,  $X(1) = 0$ , and  $\dot{T} = -\lambda^2 T$  are  $X = A \sin \lambda x + B \cos \lambda x$  and  $T = C e^{-\lambda^2 t}$ , where  $A$ ,  $B$ , and  $C$  are arbitrary constants. To satisfy the boundary condition  $X(0) = 0$ ,  $B = 0$ . An infinite number of values of  $\lambda$  (eigenvalues), say  $\lambda_n = n\pi$  ( $n = 1, 2, 3, \dots$ ), permit all the eigenfunctions  $X_n = b_n \sin \lambda_n x$  to satisfy the other boundary condition  $X(1) = 0$ . The solution of the equation and boundary conditions (not the initial condition) is, by superposition,  $u(x, t) = \sum_{n=1}^{\infty} b_n e^{-n^2 \pi^2 t} \sin n\pi x$  (a Fourier sine series), where the  $b_n$  are arbitrary. These values are obtained from the initial condition using the orthogonality properties of the trigonometric function (e.g.,  $\int_{-\pi}^{\pi} \sin mx \sin nx dx$  is 0 for  $m \neq n$  and is  $\pi$  for  $m = n \neq 0$ ) to be  $b_n = 2 \int_0^1 f(r) \sin n\pi r dr$ . Then the solution of the problem is  $u(x, t) = \sum_{n=1}^{\infty} [2 \int_0^1 f(r) \sin n\pi r dr] e^{-n^2 \pi^2 t} \sin n\pi x$ , which is a Fourier sine series.

If  $f(x)$  is a piecewise smooth or a piecewise continuous function defined for  $a \leq x \leq b$ , then its Fourier series within  $a \leq x \leq b$  as its fundamental interval (it is extended periodically outside that interval) is

$$f(x) \sim \frac{1}{2} a_0 + \sum_{n=1}^{\infty} a_n \cos [2n\pi x / (b - a)] + b_n \sin [2n\pi x / (b - a)]$$

where

$$a_n = \left[ \frac{2}{(b-a)} \right] \int_a^b f(x) \cos[2n\pi x/(b-a)] dx, \quad n = 0, 1, \dots$$

$$b_n = \left[ \frac{2}{(b-a)} \right] \int_a^b f(x) \sin[2n\pi x/(b-a)] dx \quad n = 1, 2, \dots$$

The Fourier sine series has  $a_n \equiv 0$ , and the Fourier cosine series has  $b_n \equiv 0$ . The symbol  $\sim$  means that the series converges to  $f(x)$  at points of continuity, and at the (allowable) points of finite discontinuity the series converges to the *average value* of the discontinuous values.

**Caution:** This method *only* applies to linear equations with homogeneous boundary conditions. Linear equations with variable coefficients use other orthogonal functions, such as the Bessel functions, Laguerre functions, Chebyshev functions, and so forth.

Some inhomogeneous boundary value problems can be transformed into homogeneous ones. Consider the problem  $u_t - u_{xx} = 0$ ,  $0 \leq x \leq 1$ ,  $0 \leq t < \infty$  with initial condition  $u(x, 0) = f(x)$ , and boundary conditions  $u(0, t) = g(t)$ ,  $u(1, t) = h(t)$ . To homogenize the boundary conditions set  $u(x, t) = w(x, t) + x[h(t) - g(t)] + g(t)$  and then solve  $w_t - w_{xx} = [\dot{g}(t) - \dot{h}(t)]x - \dot{g}(t)$  with the initial condition  $w(x, 0) = f(x) - x[h(0) - g(0)] + g(0)$  and  $w(0, t) = w(1, t) = 0$ .

## Operational Methods

A number of integral transforms are useful for solving a variety of linear problems. To apply the Laplace transform to the problem  $u_t - u_{xx} = \delta(x)\delta(t)$ ,  $-\infty < x < \infty$ ,  $0 \leq t$  with the initial condition  $u(x, 0^-) = 0$ , where  $\delta$  is the Dirac delta function, we multiply by  $e^{-st}$  and integrate with respect to  $t$  from 0 to  $\infty$ . With the Laplace transform of  $u(x, t)$  denoted by  $U(x, s)$ —that is,  $U(x, s) = \int_0^\infty e^{-st} u(x, t) dt$ —we have  $sU - U_{xx} = \delta(x)$ , which has the solution

$$U(x, s) = A(s)e^{-x\sqrt{s}} + B(s)e^{x\sqrt{s}} \quad \text{for } x > 0$$

$$U(x, s) = C(s)e^{-x\sqrt{s}} + D(s)e^{x\sqrt{s}} \quad \text{for } x < 0$$

Clearly,  $B(s) = C(s) = 0$  for bounded solutions as  $|x| \rightarrow \infty$ . Then, from the boundary condition,  $U(0^+, s) - U(0^-, s) = 0$  and integration of  $sU - U_{xx} = \delta(x)$  from  $0^-$  to  $0^+$  gives  $U_x(0^+, s) - U_x(0^-, s) = -1$ , so  $A = D = 1/2\sqrt{s}$ . Hence,  $U(x, s) = (1/2\sqrt{s})e^{-\sqrt{s}|x|}$  and the inverse is  $u(x, t) = (1/2\pi i) \int_\Gamma e^{st} U(x, s) ds$ , where  $\Gamma$  is a Bromwich path, a vertical line taken to the right of all singularities of  $U$  on the sphere.

## Similarity (Invariance)

This very useful approach is related to dimensional analysis; both have their foundations in group theory. The three important transformations that play a basic role in Newtonian mechanics are translation, scaling, and rotations. Using two independent variables  $x$  and  $t$  and one dependent variable  $u = u(x, t)$ , the *translation group* is  $\bar{x} = x + \alpha a$ ,  $\bar{t} = t + \beta a$ ,  $\bar{u} = u + \gamma a$ ; the *scaling*

group is  $\bar{x} = a^\alpha x$ ,  $\bar{t} = a^\beta t$ , and  $\bar{u} = a^\gamma u$ ; the *rotation group* is  $\bar{x} = x \cos a + t \sin a$ ,  $\bar{t} = t \cos a - x \sin a$ ,  $\bar{u} = u$ , with a nonnegative real number  $a$ . Important in what follows are the *invariants* of these groups. For the translation group there are two  $\eta = x - \lambda t$ ,  $\lambda = \alpha/\beta$ ,  $f(\eta) = u - \varepsilon t$ ,  $\varepsilon = \gamma/\beta$  or  $f(\eta) = u - \theta x$ ,  $\theta = \gamma/\alpha$ ; for the scaling group the invariants are  $\eta = x/t^{\alpha/\beta}$  (or  $t/x^{\beta/\alpha}$ ) and  $f(\eta) = u/t^{\gamma/\beta}$  (or  $u/x^{\gamma/\alpha}$ ); for the rotation group the invariants are  $\eta = x^2 + t^2$  and  $u = f(\eta) = f(x^2 + t^2)$ .

If a PDE and its data (initial and boundary conditions) are left invariant by a transformation group, then similar (invariant) solutions are sought using the invariants. For example, if an equation is left invariant under scaling, then solutions are sought of the form  $u(x, t) = t^{\gamma/\beta} f(\eta)$ ,  $\eta = xt^{-\alpha/\beta}$  or  $u(x, t) = x^{\gamma/\alpha} f(tx^{-\beta/\alpha})$ ; invariance under translation gives solutions of the form  $u(x, t) = f(x - \lambda t)$ ; and invariance under rotation gives rise to solutions of the form  $u(x, t) = f(x^2 + t^2)$ .

Examples of invariance include the following:

1. The equation  $u_{xx} + u_{yy} = 0$  is invariant under rotation, so we search for solutions of the form  $u = f(x^2 + y^2)$ . Substitution gives the ODE  $f' + \eta f'' = 0$  or  $(\eta f')' = 0$ . The solution is  $u(x, t) = c \ln \eta = c \ln(x^2 + t^2)$ , which is the (so-called) fundamental solution of Laplace's equation.
2. The nonlinear diffusion equation  $u_t = (u^n u_x)_x$  ( $n > 0$ ),  $0 \leq x$ ,  $0 \leq t$ ,  $u(0, t) = ct^n$  is invariant under scaling with the similar form  $u(x, t) = t^n f(\eta)$ ,  $\eta = xt^{-(n+1)/2}$ . Substituting into the PDE gives the equation  $(f^n f')' + ((n+1)/2)\eta f' - nf = 0$ , with  $f(0) = c$  and  $f(\infty) = 0$ . Note that the equation is an ODE.
3. The wave equation  $u_{xx} - u_{tt} = 0$  is invariant under translation. Hence, solutions exist of the form  $u = f(x - \lambda t)$ . Substitution gives  $f''(1 - \lambda^2) = 0$ . Hence,  $\lambda = \pm 1$  or  $f$  is linear. Rejecting the trivial linear solution we see that  $u = f(x - t) + g(x + t)$ , which is the general (d'Alembert) solution of the wave equation; the quantities  $x - t = \alpha$ ,  $x + t = \beta$  are the characteristics of the next section.

The construction of all transformations that leave a PDE invariant is a solved problem left for the references.

The study of "solitons" (solitary traveling waves with special properties) has benefited from symmetry considerations. For example, the nonlinear third-order (Korteweg–de Vries) equation  $u_t + uu_x - au_{xxx} = 0$  is invariant under translation. Solutions are sought of the form  $u = f(x - \lambda t)$ , and  $f$  satisfies the ODE, in  $\eta = x - \lambda t$ ,  $-\lambda f' + ff' - af''' = 0$ .

## Characteristics

Using the characteristics of the solution of the hyperbolic problem  $u_{tt} - u_{xx} = p(x, t)$ ,  $-\infty < x < \infty$ ,  $0 \leq t$ ,  $u(x, 0) = f(x)$ ,  $u_t(x, 0) = h(x)$  is

$$u(x, t) = \frac{1}{2} \int_0^t d\tau \int_{x-(t-\tau)}^{x+(t-\tau)} p(\xi, \tau) d\xi + \frac{1}{2} \int_{x-t}^{x+t} h(\xi) d\xi + \frac{1}{2}[f(x+t) + f(x-t)]$$

The solution of  $u_{tt} - u_{xx} = 0$ ,  $0 \leq x < \infty$ ,  $0 \leq t < \infty$ ,  $u(x, 0) = 0$ ,  $u_t(x, 0) = h(x)$ ,

$u(0, t) = 0$  ,  $t > 0$  is  $u(x, t) = \frac{1}{2} \int_{-x+t}^{x+t} h(\xi) d\xi$  .

The solution of  $u_{tt} - u_{xx} = 0$  ,  $0 \leq x < \infty$  ,  $0 \leq t < \infty$  ,  $u(x, 0) = 0$  ,  $u_t(x, 0) = 0$  ,  $u(0, t) = g(t)$  ,  $t > 0$  is

$$u(x, t) = \begin{cases} 0 & \text{if } t < x \\ g(t - x) & \text{if } t > x \end{cases}$$

From time to time, lower-order derivatives appear in the PDE in use. To remove these from the equation  $u_{tt} - u_{xx} + au_x + bu_t + cu = 0$  , where  $a$ ,  $b$ , and  $c$  are constants, set  $\xi = x + t$  ,

$\mu = t - x$  , whereupon  $u(x, t) = u[(\xi - \mu)/2, (\xi + \mu)/2] = U(\xi, \mu)$  , where

$U_{\xi\mu} + [(b + a)/4]U_{\xi} + [(b - a)/4]U_{\mu} + (c/4)U = 0$  . The transformation

$U(\xi, \mu) = W(\xi, \mu) \exp[-(b - a)\xi/4 - (b + a)\mu/4]$  reduces to satisfying  $W_{\xi\mu} + \lambda W = 0$  , where  $\lambda = (a^2 - b^2 + 4c)/16$  . If  $\lambda \neq 0$  , we lose the simple d'Alembert solution. But the equation for  $W$  is still easier to handle.

In linear problems discontinuities propagate along characteristics. In nonlinear problems the situation is usually different. The characteristics are often used as new coordinates in the numerical method of characteristics.

## Green's Function

Consider the diffusion problem  $u_t - u_{xx} = \delta(t) \delta(x - \xi)$  ,  $0 \leq x < \infty$  ,  $\xi > 0$  ,  $u(0, t) = 0$  ,

$u(x, 0) = 0$  [ $u(\infty, t) = u(\infty, 0) = 0$ ] , a problem that results from a unit source somewhere in the domain subject to a homogeneous (zero) boundary condition. The solution is called a *Green's function of the first kind*. For this problem there is  $G_1(x, \xi, t) = F(x - \xi, t) - F(x + \xi, t)$  , where  $F(x, t) = e^{-x^2/4t} / \sqrt{4\pi t}$  is the *fundamental* (invariant) *solution*. More generally, the solution of  $u_t - u_{xx} = \delta(x - \xi) \delta(t - \tau)$  ,  $\xi > 0$  ,  $\tau > 0$  , with the same conditions as before, is the Green's function of the first kind

$$G_1(x, \xi, t - \tau) = \frac{1}{\sqrt{4\pi(t - \tau)}} \left[ e^{-(x-\xi)^2/4(t-\tau)} - e^{-(x+\xi)^2/4(t-\tau)} \right]$$

for the semiinfinite interval.

The solution of  $u_t - u_{xx} = p(x, t)$  ,  $0 \leq x < \infty$  ,  $0 \leq t < \infty$  , with  $u(x, 0) = 0$  ,  $u(0, t) = 0$  ,  $t > 0$  is  $u(x, t) = \int_0^t d\tau \int_0^\infty p(\xi, \tau) G_1(x, \xi, t - \tau) d\xi$  , which is a superposition. Note that the Green's function and the desired solution must both satisfy a zero boundary condition at the origin for this solution to make sense.

The solution of  $u_t - u_{xx} = 0$  ,  $0 \leq x < \infty$  ,  $0 \leq t < \infty$  ,  $u(x, 0) = f(x)$  ,  $u(0, t) = 0$  ,  $t > 0$  is  $u(x, t) = \int_0^\infty f(\xi) G_1(x, \xi, t) d\xi$  .

The solution of  $u_t - u_{xx} = 0$  ,  $0 \leq x < \infty$  ,  $0 \leq t < \infty$  ,  $u(x, 0) = 0$  ,  $u(0, t) = g(t)$  ,  $t > 0$  (nonhomogeneous) is obtained by transforming to a new problem that has a homogeneous boundary condition. Thus, with  $w(x, t) = u(x, t) - g(t)$  the equation for  $w$  becomes

$w_t - w_{xx} = -\dot{g}(t) - g(0) \delta(t)$  and  $w(x, 0) = 0$  ,  $w(0, t) = 0$  . Using  $G_1$  , above, we finally obtain  $u(x, t) = (x/\sqrt{4\pi}) \int_0^t g(t - \tau) e^{-x^2/4t} / \tau^{3/2} d\tau$  .

The Green's function approach can also be employed for elliptic and hyperbolic problems.

### Equations in Other Spatial Variables

The spherically symmetric wave equation  $u_{rr} + 2u_r/r - u_{tt} = 0$  has the general solution

$$u(r, t) = [f(t - r) + g(t + r)]/r \quad .$$

The Poisson-Euler-Darboux equation, arising in gas dynamics,

$$u_{rs} + N(u_r + u_s)/(r + s) = 0$$

where  $N$  is a positive integer  $\geq 1$ , has the general solution

$$u(r, s) = k + \frac{\partial^{N-1}}{\partial r^{N-1}} \left[ \frac{f(r)}{(r + s)^N} \right] + \frac{\partial^{N-1}}{\partial s^{N-1}} \left[ \frac{g(s)}{(r + s)^N} \right]$$

Here,  $k$  is an arbitrary constant and  $f$  and  $g$  are arbitrary functions whose form is determined from the problem initial and boundary conditions.

### Conversion to Other Orthogonal Coordinate Systems

Let  $(x^1, x^2, x^3)$  be rectangular (Cartesian) coordinates and  $(u^1, u^2, u^3)$  be any orthogonal coordinate system related to the rectangular coordinates by  $x^i = x^i(u^1, u^2, u^3)$ ,  $i = 1, 2, 3$ . With  $(ds)^2 = (dx^1)^2 + (dx^2)^2 + (dx^3)^2 = g_{11}(du^1)^2 + g_{22}(du^2)^2 + g_{33}(du^3)^2$ , where  $g_{ii} = (\partial x^1 / \partial u^i)^2 + (\partial x^2 / \partial u^i)^2 + (\partial x^3 / \partial u^i)^2$ . In terms of these "metric" coefficients the basic operations of applied mathematics are expressible. Thus (with  $g = g_{11} g_{22} g_{33}$ )

$$dA = (g_{11} g_{22})^{1/2} du^1 du^2; \quad dV = (g_{11} g_{22} g_{33})^{1/2} du^1 du^2 du^3;$$

$$\text{grad } \phi = \frac{\vec{a}_1}{(g_{11})^{1/2}} \frac{\partial \phi}{\partial u^1} + \frac{\vec{a}_2}{(g_{22})^{1/2}} \frac{\partial \phi}{\partial u^2} + \frac{\vec{a}_3}{(g_{33})^{1/2}} \frac{\partial \phi}{\partial u^3}$$

( $\vec{a}_i$  are unit vectors in direction  $i$ );

$$\text{div } \vec{E} = g^{-1/2} \left\{ \frac{\partial}{\partial u^1} [(g_{22} g_{33})^{1/2} E_1] + \frac{\partial}{\partial u^2} [(g_{11} g_{33})^{1/2} E_2] + \frac{\partial}{\partial u^3} [(g_{11} g_{22})^{1/2} E_3] \right\}$$

[here  $\vec{E} = (E_1, E_2, E_3)$ ];



$$\begin{aligned} \text{curl } \vec{E} = g^{-1/2} \left\{ \vec{a}_1 (g_{11})^{1/2} \left( \frac{\partial}{\partial u^2} [(g_{33})^{1/2} E_3] - \frac{\partial}{\partial u^3} [(g_{22})^{1/2} E_2] \right) \right. \\ + \vec{a}_2 (g_{22})^{1/2} \left( \frac{\partial}{\partial u^3} [(g_{11})^{1/2} E_1] - \frac{\partial}{\partial u^1} [(g_{33})^{1/2} E_3] \right) \\ \left. + \vec{a}_3 (g_{33})^{1/2} \left( \frac{\partial}{\partial u^1} [(g_{22})^{1/2} E_2] - \frac{\partial}{\partial u^2} [(g_{11})^{1/2} E_1] \right) \right\} \end{aligned}$$

$$\text{div grad } \psi = \nabla^2 \psi = \text{Laplacian of } \psi = g^{-1/2} \sum_{i=1}^3 \frac{\partial}{\partial u^i} \left[ \frac{g^{1/2}}{g_{ii}} \frac{\partial \psi}{\partial u^i} \right]$$

Table 201.2 shows some coordinate systems.

**Table 201.2** Some Coordinate Systems

Coordinate System	Metric Coefficients	
Circular Cylindrical		
$x = r \cos \theta$	$u^1 = r$	$g_{11} = 1$
$y = r \sin \theta$	$u^2 = \theta$	$g_{22} = r^2$
$z = z$	$u^3 = z$	$g_{33} = 1$
Spherical		
$x = r \sin \psi \cos \theta$	$u^1 = r$	$g_{11} = 1$
$y = r \sin \psi \sin \theta$	$u^2 = \psi$	$g_{22} = r^2$
$z = r \cos \theta$	$u^3 = \theta$	$g_{33} = r^2 \sin^2 \psi$
Parabolic Coordinates		
$x = \mu \nu \cos \theta$	$u^1 = \mu$	$g_{11} = \mu^2 + \nu^2$
$y = \mu \nu \sin \theta$	$u^2 = \nu$	$g_{22} = \mu^2 + \nu^2$
$z = \frac{1}{2}(\mu^2 - \nu^2)$	$u^3 = \theta$	$g_{33} = \mu^2 \nu^2$

Other metric coefficients and so forth can be found in Moon and Spencer [1961].

## References

- Ames, W. F. 1965. *Nonlinear Partial Differential Equations in Science and Engineering, Volume 1*. Academic Press, Boston, MA.
- Ames, W. F. 1972. *Nonlinear Partial Differential Equations in Science and Engineering, Volume 2*. Academic Press, Boston, MA.
- Brauer, F. and Nohel, J. A. 1986. *Introduction to Differential Equations with Applications*. Harper & Row, New York.
- Jeffrey, A. 1990. *Linear Algebra and Ordinary Differential Equations*. Blackwell Scientific, Boston, MA.

Kevorkian, J. 1990. *Partial Differential Equations*. Wadsworth and Brooks/Cole, Belmont, CA.  
 Moon, P. and Spencer, D. E. 1961. *Field Theory Handbook*. Springer, Berlin.  
 Rogers, C. and Ames, W. F. 1989. *Nonlinear Boundary Value Problems in Science and Engineering*. Academic Press, Boston, MA.  
 Whitham, G. B. 1974. *Linear and Nonlinear Waves*. John Wiley & Sons, New York.  
 Zauderer, E. 1983. *Partial Differential Equations of Applied Mathematics*. John Wiley & Sons, New York.  
 Zwillinger, D. 1992. *Handbook of Differential Equations*. Academic Press, Boston, MA.

## Further Information

A collection of solutions for linear and nonlinear problems is found in E. Kamke, *Differential-gleichungen-Lösungsmethoden und Lösungen*, Akad. Verlagsges, Leipzig, 1956. Also see G. M. Murphy, *Ordinary Differential Equations and Their Solutions*, Van Nostrand, Princeton, NJ, 1960 and D. Zwillinger, *Handbook of Differential Equations*, Academic Press, Boston, MA, 1992.

For nonlinear problems see

Ames, W. F. 1968. *Nonlinear Ordinary Differential Equations in Transport Phenomena*. Academic Press, Boston, MA.  
 Cunningham, W. J. 1958. *Introduction to Nonlinear Analysis*. McGraw-Hill, New York.  
 Jordan, D. N. and Smith, P. 1977. *Nonlinear Ordinary Differential Equations*. Clarendon Press, Oxford, UK.  
 McLachlan, N. W. 1955. *Ordinary Non-Linear Differential Equations in Engineering and Physical Sciences*, 2nd ed. Oxford University Press, London.  
 Zwillinger, D. 1992.

Ames, W. F. "Integral Equations"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

[202.1 Classification and Notation](#)[202.2 Relation to Differential Equations](#)[202.3 Methods of Solution](#)

Convolution Equations • Abel Equation • Approximate Method (Picard's Method)

**William F. Ames**

Georgia Institute of Technology

---

**202.1 Classification and Notation**

---

Any equation in which the unknown function  $u(x)$  appears under the integral sign is called an *integral equation*. If  $f(x)$ ,  $K(x, t)$ ,  $a$ , and  $b$  are known then the integral equation for  $u$ ,  $\int_a^b K(x, t)u(t) dt = f(x)$  is called a *linear integral equation of the first kind of Fredholm type*.  $K(x, t)$  is called the *kernel function* of the equation. If  $b$  is replaced by  $x$  (the independent variable) the equation is an equation of *Volterra type of the first kind*.

An equation of the form  $u(x) = f(x) + \lambda \int_a^b K(x, t)u(t) dt$  is said to be a linear integral equation of *Fredholm type of the second kind*. If  $b$  is replaced by  $x$  it is of *Volterra type*. If  $f(x)$  is not present the equation is homogeneous.

The equation  $\phi(x)u(x) = f(x) + \lambda \int_a^b \text{or } x K(x, t)u(t) dt$  is the *third kind equation* of Fredholm or Volterra type. If the unknown function  $u$  appears in the equation in any way other than to the first power then the integral equation is said to be *nonlinear*. Thus,  $u(x) = f(x) + \int_a^b K(x, t) \sin u(t) dt$  is nonlinear. An integral equation is said to be *singular* when either or both of the limits of integration are infinite or if  $K(x, t)$  becomes infinite at one or more points of the integration interval.

**Example 202.1.** Consider the singular equations  $u(x) = x + \int_0^\infty \sin(xt)u(t) dt$  and  $f(x) = \int_0^x [u(t)/(x-t)^2] dt$ .

---

**202.2 Relation to Differential Equations**

---

The *Leibnitz rule*

$$(d/dx) \int_{a(x)}^{b(x)} F(x, t) dt = \int_{a(x)}^{b(x)} (\partial F / \partial x) dt + F[x, b(x)](db/dx) - F[x, a(x)](da/dx)$$
 is useful

for differentiation of an integral involving a parameter ( $x$  in this case). With this, one can establish the relation

$$I_n(x) = \int_a^x (x-t)^{n-1} f(t) dt = (n-1)! \underbrace{\int_a^x \cdots \int_a^x}_{n \text{ times}} f(x) \underbrace{dx \cdots dx}_{n \text{ times}}$$

This result will be used to establish the relation of the second-order initial value problem to a Volterra integral equation.

The second-order differential equation  $y''(x) + A(x)y'(x) + B(x)y = f(x)$  ,  $y(a) = y_0$  ,  $y'(a) = y'_0$  is equivalent to the integral equations

$$\begin{aligned} y(x) = & - \int_a^x \{A(t) + (x-t)[B(t) - A'(t)]\} y(t) dt \\ & + \int_a^x (x-t)f(t) dt + [A(a)y_0 + y'_0](x-a) + y_0 \end{aligned}$$

which is of the type  $y(x) = \int_a^x K(x,t)y(t) dt + F(x)$  , where

$K(x,t) = (t-x)[B(t) - A'(t)] - A(t)$  and  $F(x)$  includes the rest of the terms. Thus, this initial value problem is equivalent to a Volterra integral equation of the second kind.

**Example 202.2.** Consider the equation  $y'' + x^2y' + xy = x$  ,  $y(0) = 1$  ,  $y'(0) = 0$  . Here  $A(x) = x^2$  ,  $B(x) = x$  ,  $f(x) = x$  ,  $a = 0$  ,  $y_0 = 1$  ,  $y'_0 = 0$  . The integral equation is  $y(x) = \int_0^x t(x-2t)y(t) dt + (x^3/6) + 1$  .

The expression for  $I_n(x)$  can also be useful in converting boundary value problems to integral equations. For example, the problem  $y''(x) + \lambda y = 0$  ,  $y(0) = 0$  ,  $y(a) = 0$  is equivalent to the Fredholm equation  $y(x) = \lambda \int_0^a K(x,t)y(t) dt$  , where  $K(x,t) = (t/a)(a-x)$  when  $t < x$  and  $K(x,t) = (x/a)(a-t)$  when  $t > x$  .

In both cases the differential equation can be recovered from the integral equation by using the Leibnitz rule.

Nonlinear differential equations can also be transformed into integral equations. In fact this is one method used to establish properties of the equation and to develop approximate and numerical solutions. For example, the "forced pendulum" equation  $y''(x) + a^2 \sin y(x) = f(x)$  ,  $y(0) = y(1) = 0$  transforms into the nonlinear Fredholm equation

$$y(x) = \int_0^1 K(x,t)[a^2 \sin y(t) - f(t)] dt$$

with  $K(x,t) = x(1-y)$  for  $0 < x < t$  and  $K(x,t) = t(1-x)$  for  $t < x < 1$  .

## 202.3 Methods of Solution

---

Only the simplest integral equations can be solved exactly. Usually approximate or numerical methods are employed. The advantage here is that integration is a "smoothing operation," whereas differentiation is a "roughening operation." A few exact and approximate methods are given in the following sections. The numerical methods are found in **Chapter 209**.

### Convolution Equations

The special convolution equation  $y(x) = f(x) + \lambda \int_0^x K(x-t)y(t) dt$  is a special case of the Volterra equation of the second kind.  $K(x-t)$  is said to be a *convolution kernel*. The integral part is the convolution integral discussed in **Chapter 204**. The solution can be accomplished by transforming with the Laplace transform:  $L[y(x)] = L[f(x)] + \lambda L[y(x)]L[K(x)]$  or  $y(x) = L^{-1}\{L[f(x)]/(1 - \lambda L[K(x)])\}$ .

### Abel Equation

The Volterra equation  $f(x) = \int_0^x y(t)/(x-t)^\alpha dt$ ,  $0 < \alpha < 1$  is the (singular) Abel equation. Its solution is  $y(x) = (\sin \alpha\pi/\pi)(d/dx) \int_0^x F(t)/(x-t)^{1-\alpha} dt$ .

### Approximate Method (Picard's Method)

This method is one of successive approximations that is described for the equation  $y(x) = f(x) + \lambda \int_a^x K(x,t)y(t) dt$ . Beginning with an initial guess  $y_0(t)$  (often the value at the initial point  $a$ ) generate the next approximation with  $y_1(x) = f(x) + \lambda \int_a^x K(x,t)y_0(t) dt$  and continue with the general iteration

$$y_n(x) = f(x) + \lambda \int_a^x K(x,t)y_{n-1}(t) dt$$

Then, by iterating, one studies the convergence of this process, as is described in the literature.

**Example 202.3.** Let  $y(x) = 1 + \int_0^x xt[y(t)]^2 dt$ ,  $y(0) = 1$ . With  $y_0(t) = 1$  we find  $y_1(x) = 1 + \int_0^x xt dt = 1 + (x^3/2)$  and  $y_2(x) = 1 + \int_0^x xt[1 + (t^3/2)]^2 dt$ , and so forth.

### References

- Jerri, A. J. 1985. *Introduction to Integral Equations with Applications*. Marcel Dekker, New York.  
Tricomi, F. G. 1958. *Integral Equations*. Wiley-Interscience, New York.  
Yosida, K. 1960. *Lectures on Differential and Integral Equations*. Wiley-Interscience, New York.

Ames, W. F. "Approximation Methods"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

**203.1 Perturbation**

Regular Perturbation • Singular Perturbation • Boundary Layer Method

**203.2 Iterative Methods**

Taylor Series • Picard's Method

**William F. Ames***Georgia Institute of Technology*

The term *approximation methods* usually refers to an analytical process that generates a symbolic approximation rather than a numerical one. Thus,  $1 + x + x^2/2$  is an approximation of  $e^x$  for small  $x$ . This chapter introduces some techniques for approximating the solution of various operator equations.

**203.1 Perturbation****Regular Perturbation**

This procedure is applicable to *some* equations in which a small parameter,  $\varepsilon$ , appears. Use this procedure with care; the procedure involves expansion of the dependent variables and data in a power series in the small parameter. The following example illustrates the procedure.

**Example 203.1.** Consider the equation  $y'' + \varepsilon y' + y = 0$ ,  $y(0) = 1$ ,  $y'(0) = 0$ . Write  $y(x; \varepsilon) = y_0(x) + \varepsilon y_1(x) + \varepsilon^2 y_2(x) + \cdots$ , and the initial conditions (data) become

$$\begin{aligned} y_0(0) + \varepsilon y_1(0) + \varepsilon^2 y_2(0) + \cdots &= 1 \\ y'_0(0) + \varepsilon y'_1(0) + \varepsilon^2 y'_2(0) + \cdots &= 0 \end{aligned}$$

Equating like powers of  $\varepsilon$  in all three equations yields the sequence of equations

$$\begin{aligned} O(\varepsilon^0) : \quad y''_0 + y_0 &= 0, & y_0(0) &= 1, & y'_0(0) &= 0 \\ O(\varepsilon^1) : \quad y''_1 + y_1 &= -y'_0, & y_1(0) &= 0, & y'_1(0) &= 0 \\ & \vdots \end{aligned}$$



The solution for  $y_0$  is  $y_0 = \cos x$  and using this for  $y_1$  we find  $y_1(x) = \frac{1}{2}(\sin x - x \cos x)$ . So  $y(x; \varepsilon) = \cos x + \varepsilon(\sin x - x \cos x)/2 + O(\varepsilon^2)$ . Appearance of the term  $x \cos x$  indicates a *secular term* that becomes arbitrarily large as  $x \rightarrow \infty$ . Hence, this approximation is valid only for  $x \ll 1/\varepsilon$  and for small  $\varepsilon$ . If an approximation is desired over a larger range of  $x$  then the method of multiple scales is required.

## Singular Perturbation

The *method of multiple scales* is a singular method that is *sometimes* useful if the regular perturbation method fails. In this case the assumption is made that the solution depends on *two* (or more) different length (or time) scales. By trying various possibilities, one can determine those scales. The scales are treated as dependent variables when transforming the given ordinary differential equation into a partial differential equation, but then the scales are treated as independent variables when solving the equations.

**Example 203.2.** Consider the equation  $\varepsilon y'' + y' = 2$ ,  $y(0) = 0$ ,  $y(1) = 1$ . This is singular since (with  $\varepsilon = 0$ ) the resulting first-order equation cannot satisfy both boundary conditions. For the problem the proper length scales are  $u = x$  and  $v = x/\varepsilon$ . The second scale can be ascertained by substituting  $\varepsilon^n x$  for  $x$  and requiring  $\varepsilon y''$  and  $y'$  to be of the same order in the transformed equation. Then

$$\frac{d}{dx} = \frac{\partial}{\partial u} \frac{du}{dx} + \frac{\partial}{\partial v} \frac{dv}{dx} = \frac{\partial}{\partial u} + \frac{1}{\varepsilon} \frac{\partial}{\partial v}$$

and the equation becomes

$$\varepsilon \left( \frac{\partial}{\partial u} + \frac{1}{\varepsilon} \frac{\partial}{\partial v} \right)^2 y + \left( \frac{\partial}{\partial u} + \frac{1}{\varepsilon} \frac{\partial}{\partial v} \right) y = 2.$$

With  $y(x; \varepsilon) = y_0(u, v) + \varepsilon y_1(u, v) + \varepsilon^2 y_2(u, v) + \cdots$  we have terms

$$O(\varepsilon^{-1}) : \frac{\partial^2 y_0}{\partial v^2} + \frac{\partial y_0}{\partial v} = 0 \quad (\text{actually ODEs with parameter } u)$$

$$O(\varepsilon^0) : \frac{\partial^2 y_1}{\partial v^2} + \frac{\partial y_1}{\partial v} = 2 - 2 \frac{\partial^2 y_0}{\partial u \partial v} - \frac{\partial y_0}{\partial u}$$

$$O(\varepsilon^1) : \frac{\partial^2 y_2}{\partial v^2} + \frac{\partial y_2}{\partial v} = -2 \frac{\partial^2 y_1}{\partial u \partial v} - \frac{\partial y_1}{\partial u} - \frac{\partial^2 y_0}{\partial u^2}$$

$\vdots$

Then  $y_0(u, v) = A(u) + B(u)e^{-v}$  and so the second equation becomes

$\partial^2 y_1 / \partial v^2 + \partial y_1 / \partial v = 2 - A'(u) + B'(u)e^{-v}$  , with the solution  
 $y_1(u, v) = [2 - A'(u)]v + vB'(u)e^{-v} + D(u) + E(u)e^{-v}$  . Here  $A, B, D$  and  $E$  are still arbitrary.  
 Now the solvability condition—"higher order terms must vanish no slower (as  $\varepsilon \rightarrow 0$ ) than the previous term" [[Kevorkian and Cole, 1981](#)]  
 is used. For  $y_1$  to vanish no slower than  $y_0$  we must have  $2 - A'(u) = 0$  and  $B'(u) = 0$  . If this were not true the terms in  $y_1$  would be larger than those in  $y_0$  ( $v \gg 1$ ) . Thus  $y_0(u, v) = (2u + A_0) + B_0 e^{-v}$  , or in the original variables  
 $y(x; \varepsilon) \approx (2x + A_0) + B_0 e^{-x/\varepsilon}$  and matching to both boundary conditions gives  
 $y(x; \varepsilon) \approx 2x - (1 - e^{-x/\varepsilon})$  .

## Boundary Layer Method

The boundary layer method is applicable to regions in which the solution is *rapidly varying*. See the references at the end of the chapter for detailed discussion.

## 203.2 Iterative Methods

---

### Taylor Series

If it is known that the solution of a differential equation has a power series in the independent variable ( $t$ ), then we may proceed from the initial data (the easiest problem) to compute the Taylor series by differentiation.

**Example 203.3.** Consider the equation  $(d^2 x / dt^2) = -x - x^2$  ,  $x(0) = 1$  ,  $x'(0) = 1$  . From the differential equation,  $x''(0) = -2$  , and, since  $x''' = -x' - 2xx'$  ,  $x'''(0) = -1 - 2 = -3$  , so the four term approximation for  $x(t) \approx 1 + t - (2t^2/2!) - (3t^3/3!) = 1 + t - t^2 - t^3/2$  . An estimate for the error at  $t = t_1$  , (see a discussion of series methods in any calculus text) is not greater than  $|d^4 x / dt^4|_{\max} [(t_1)^4 / 4!]$  ,  $0 \leq t \leq t_1$  .

### Picard's Method

If the vector differential equation  $\mathbf{x}' = f(t, \mathbf{x})$  ,  $\mathbf{x}(0)$  given, is to be approximated by Picard iteration, we begin with an initial guess  $\mathbf{x}_0 = \mathbf{x}(0)$  and calculate iteratively  $\mathbf{x}'_i = f(t, \mathbf{x}_{i-1})$  .

**Example 203.4.** Consider the equation  $x' = x + y^2$  ,  $y' = y - x^3$  ,  $x(0) = 1$  ,  $y(0) = 2$  . With  $x_0 = 1$  ,  $y_0 = 2$  ,  $x'_1 = 5$  ,  $y'_1 = 1$  , so  $x_1 = 5t + 1$  ,  $y_1 = t + 2$  , since  $x_i(0) = 1$  ,  $y_i(0) = 2$  for  $i \geq 0$  . To continue, use  $x'_{i+1} = x_i + y_i^2$  ,  $y'_{i+1} = y_i - x_i^3$  . A modification is the utilization of the first calculated term immediately in the second equation. Thus, the calculated value of  $x_1 = 5t + 1$  , when used in the second equation, gives  $y'_1 = y_0 - (5t + 1)^3 = 2 - (125t^3 + 75t^2 + 15t + 1)$  , so  $y_1 = 2t - (125t^4/4) - 25t^3 - (15t^2/2) - t + 2$  . Continue with the iteration  $x'_{i+1} = x_i + y_i^2$  ,  $y'_{i+1} = y_i - (x_{i+1})^3$  .

Another variation would be  $x'_{i+1} = x_{i+1} + (y_i)^2$  ,  $y'_{i+1} = y_{i+1} - (x_{i+1})^3$  .

## References

- Ames, W. F. 1965. *Nonlinear Partial Differential Equations in Engineering, Volume I*. Academic Press, Boston, MA.
- Ames, W. F. 1968. *Nonlinear Ordinary Differential Equations in Transport Processes*. Academic Press, Boston, MA.
- Ames, W. F. 1972. *Nonlinear Partial Differential Equations in Engineering, Volume II*. Academic Press, Boston, MA.
- Kevorkian, J. and Cole, J. D. 1981. *Perturbation Methods in Applied Mathematics*. Springer, New York.
- Miklin, S. G. and Smolitskiy, K. L. 1967. *Approximate Methods for Solutions of Differential and Integral Equations*. Elsevier, New York.
- Nayfeh, A. H. 1973. *Perturbation Methods*. John Wiley & Sons, New York.
- Zwillinger, D. 1992. *Handbook of Differential Equations*, 2nd ed. Academic Press, Boston, MA.

Ames, W. F. "Integral Transforms"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

**204.1 Laplace Transform**

Properties of the Laplace Transform

**204.2 Convolution Integral****204.3 Fourier Transform**

Properties of the Fourier Transform

**204.4 Fourier Cosine Transform**

Properties of the Fourier Cosine Transform

**William F. Ames***Georgia Institute of Technology*

All of the integral transforms are special cases of the equation  $g(s) = \int_a^b K(s, t)f(t) dt$ , in which  $g(s)$  is said to be the *transform* of  $f(t)$ , and  $K(s, t)$  is called the *kernel* of the transform. [Table 204.1](#) shows the more important kernels and the corresponding intervals  $(a, b)$ .

**Table 204.1** Kernels and Intervals of Various Integral Transforms

Name of Transform	$(a, b)$	$K(s, t)$
Laplace	$(0, \infty)$	$e^{-st}$
Fourier	$(-\infty, \infty)$	$\frac{1}{\sqrt{2\pi}}e^{-ist}$
Fourier cosine	$(0, \infty)$	$\sqrt{\frac{2}{\pi}} \cos st$
Fourier sine	$(0, \infty)$	$\sqrt{\frac{2}{\pi}} \sin st$
Mellin	$(0, \infty)$	$t^{s-1}$
Hankel	$(0, \infty)$	$tJ_v(st), \quad v \geq -\frac{1}{2}$

Details for the first three transforms listed in [Table 204.1](#) are given in this chapter. The details for the others are found in the literature.

**204.1 Laplace Transform**

The Laplace transform of  $f(t)$  is  $g(s) = \int_0^\infty e^{-st} f(t) dt$ . It may be thought of as transforming one class of functions into another. The advantage in the operation is that under certain circumstances it replaces complicated functions by simpler ones. The notation  $L[f(t)] = g(s)$  is called the *direct*

transform and  $L^{-1}[g(s)] = f(t)$  is called the *inverse transform*. Both the direct and inverse transforms are tabulated for many often-occurring functions. In general  $L^{-1}[g(s)] = (1/2\pi i) \int_{\alpha-i\infty}^{\alpha+i\infty} e^{st} g(s) ds$ , and to evaluate this integral requires a knowledge of complex variables, the theory of residues, and contour integration.

## Properties of the Laplace Transform

Let  $L[f(t)] = g(s)$ ,  $L^{-1}[g(s)] = f(t)$ .

1. The Laplace transform may be applied to a function  $f(t)$  if  $f(t)$  is continuous or piecewise continuous; if  $t^n |f(t)|$  is finite for all  $t$ ,  $t \rightarrow 0$ ,  $n < 1$ ; and if  $e^{-at} |f(t)|$  is finite as  $t \rightarrow \infty$  for some value of  $a$ ,  $a > 0$ .
2.  $L$  and  $L^{-1}$  are unique.
3.  $L[af(t) + bh(t)] = aL[f(t)] + bL[h(t)]$  (linearity).
4.  $L[e^{at} f(t)] = g(s - a)$  (shift theorem).
5.  $L[(-t)^k f(t)] = d^k g/ds^k$ ;  $k$  a positive integer.

**Example 204.1.**  $L[\sin at] = \int_0^\infty e^{-st} \sin at dt = a/(s^2 + a^2)$ ,  $s > 0$ . By property 5,

$$\int_0^\infty e^{-st} t \sin at dt = L[t \sin at] = \frac{2as}{s^2 + a^2}$$

6.

$$\begin{aligned} L[f'(t)] &= sL[f(t)] - f(0) \\ L[f''(t)] &= s^2 L[f(t)] - sf(0) - f'(0) \\ &\vdots \\ L[f^{(n)}(t)] &= s^n L[f(t)] - s^{n-1} f(0) - \dots - sf^{(n-2)}(0) - f^{(n-1)}(0) \end{aligned}$$

In this property it is apparent that the initial data are automatically brought into the computation.

**Example 204.2.** Solve  $y'' + y = e^t$ ,  $y(0) = 0$ ,  $y'(0) = 1$ . Now

$L[y''] = s^2 L[y] - sy(0) - y'(0) = s^2 L[y] - s - 1$ . Thus, using the linear property of the transform (property 3),  $s^2 L[y] + L[y] - s - 1 = L[e^t] = 1/(s - 1)$ . Therefore,  $L[y] = s^2 / [(s - 1)(s^2 + 1)]$ .

With the notations  $\Gamma(n + 1) = \int_0^\infty x^n e^{-x} dx$  (gamma function) and  $J_n(t)$  the Bessel function of the first kind of order  $n$ , a short table of Laplace transforms is given in [Table 204.2](#).

$$7. L\left[\int_a^t f(t) dt\right] = \frac{1}{s} L[f(t)] + \frac{1}{s} \int_a^0 f(t) dt.$$

**Example 204.3.** Find  $f(t)$  if  $L[f(t)] = (1/s^2) [1/(s^2 - a^2)]$ .  $L[1/a \sinh at] = 1/(s^2 - a^2)$ .  
Therefore,  $f(t) = \int_0^t [\int_0^t \frac{1}{a} \sinh at dt] dt = 1/a^2 [(\sinh at)/a - t]$ .

8.

$$L \left[ \frac{f(t)}{t} \right] = \int_s^\infty g(s) ds; \quad L \left[ \frac{f(t)}{t^k} \right] = \underbrace{\int_s^\infty \cdots \int_s^\infty}_{k \text{ integrals}} g(s) (ds)^k$$

**Example 204.4.**  $L[(\sin at)/t] = \int_s^\infty L[\sin at] ds = \int_s^\infty [a ds/(s^2 + a^2)] = \cot^{-1}(s/a)$ .

9. The *unit step function*  $u(t - a) = 0$  for  $t < a$  and 1 for  $t > a$ .  $L[u(t - a)] = e^{-as}/s$ .

10. The *unit impulse function* is  $\delta(a) = u'(t - a) = 1$  at  $t = a$  and 0 elsewhere.

$$L[u'(t - a)] = e^{-as}.$$

11.  $L^{-1}[e^{-as} g(s)] = f(t - a)u(t - a)$  (second shift theorem).

12. If  $f(t)$  is *periodic* of period  $b$  that is,  $f(t + b) = f(t)$  — then  $L[f(t)] = [1/(1 - e^{-bs})] \times \int_0^b e^{-st} f(t) dt$ .

**Example 204.5.** The equation  $\partial^2 y / (\partial t \partial x) + \partial y / \partial t + \partial y / \partial x = 0$  with  $(\partial y / \partial x)(0, x) = y(0, x) = 0$  and  $y(t, 0) + (\partial y / \partial t)(t, 0) = \delta(0)$  (see property 10) is solved by using the Laplace transform of  $y$  with respect to  $t$ . With  $g(s, x) = \int_0^\infty e^{-st} y(t, x) dt$ , the transformed equation becomes

$$s \frac{\partial g}{\partial x} - \frac{\partial y}{\partial x}(0, x) + sg - y(0, x) + \frac{\partial g}{\partial x} = 0$$

or

$$(s + 1) \frac{\partial g}{\partial x} + sg = \frac{\partial y}{\partial x}(0, x) + y(0, x) = 0$$

The second (boundary) condition gives  $g(s, 0) + sg(s, 0) - y(0, 0) = 1$  or  $g(s, 0) = 1/(1 + s)$ . A solution of the preceding ordinary differential equation consistent with this condition is  $g(s, x) = [1/(s + 1)] e^{-sx/(s+1)}$ . Inversion of this transform gives  $y(t, x) = e^{-(t+x)} I_0(2\sqrt{tx})$ , where  $I_0$  is the zero-order Bessel function of an imaginary argument.

**Table 204.2** Some Laplace Transforms

$f(t)$	$g(s)$	$f(t)$	$g(s)$
1	$\frac{1}{s}$	$e^{-at}(1 - at)$	$\frac{s}{(s + a)^2}$
$t^n, n \text{ is a + integer}$	$\frac{n!}{s^{n+1}}$	$\frac{t \sin at}{2a}$	$\frac{s}{(s^2 + a^2)^2}$

$t^n, n \neq a + \text{integer}$	$\frac{\Gamma(n+1)}{s^{n+1}}$	$\frac{1}{2a^2} \sin at \sinh at$	$\frac{s}{s^4 + 4a^4}$
$\cos at$	$\frac{s}{s^2 + a^2}$	$\cos at \cosh at$	$\frac{s^3}{s^4 + 4a^4}$
$\sin at$	$\frac{a}{s^2 + a^2}$	$\frac{1}{2a}(\sinh at + \sin at)$	$\frac{s^2}{s^4 - a^4}$
$\cosh at$	$\frac{s}{s^2 - a^2}$	$\frac{1}{2}(\cosh at + \cos at)$	$\frac{s^3}{s^4 - a^4}$
$\sinh at$	$\frac{a}{s^2 - a^2}$	$\frac{\sin at}{t}$	$\tan^{-1} \frac{a}{s}$
$e^{-at}$	$\frac{1}{s + a}$	$J_0(at)$	$\frac{1}{\sqrt{s^2 + a^2}}$
$e^{-bt} \cos at$	$\frac{s + b}{(s + b)^2 + a^2}$	$\frac{n}{a^n} \frac{J_n(at)}{t}$	$\frac{1}{(\sqrt{s^2 + a^2} + s)^n}$
$e^{-bt} \sin at$	$\frac{a}{(s + b)^2 + a^2}$	$J_0(2\sqrt{at})$	$\frac{1}{s} e^{-a/s}$

## 204.2 Convolution Integral

The *convolution integral* (*faltung*) of two functions  $f(t)$ ,  $r(t)$  is  
 $x(t) = f(t) * r(t) = \int_0^t f(\tau) r(t - \tau) d\tau$  .

**Example 204.6.**  $t * \sin t = \int_0^t \tau \sin(t - \tau) d\tau = t - \sin t$  .

$$13. L[f(t)]L[h(t)] = L[f(t) * h(t)] \quad .$$

## 204.3 Fourier Transform

The *Fourier transform* is given by  $F[f(t)] = (1/\sqrt{2\pi}) \int_{-\infty}^{\infty} f(t)e^{-ist} dt = g(s)$  and its *inverse* by  $F^{-1}[g(s)] = (1/\sqrt{2\pi}) \int_{-\infty}^{\infty} g(s)e^{ist} dt = f(t)$  . In brief, the condition for the Fourier transform to exist is that  $\int_{-\infty}^{\infty} |f(t)| dt < \infty$  , although certain functions may have a Fourier transform even if this is violated.

**Example 204.7.** The function  $f(t) = 1$  for  $-a \leq t \leq a$  and  $= 0$  elsewhere has

$$F[f(t)] = \int_{-a}^a e^{-ist} dt = \int_0^a e^{ist} dt + \int_0^a e^{-ist} dt = 2 \int_0^a \cos st dt = \frac{2 \sin sa}{s}$$

## Properties of the Fourier Transform

Let  $F[f(t)] = g(s)$  ;  $F^{-1}[g(s)] = f(t)$  .

$$1. F[f^{(n)}(t)] = (is)^n F[f(t)]$$



2.  $F[af(t) + bh(t)] = aF[f(t)] + bF[h(t)]$
3.  $F[f(-t)] = g(-s)$
4.  $F[f(at)] = 1/a g(s/a)$  ,  $a > 0$
5.  $F[e^{-iwt} f(t)] = g(s + w)$
6.  $F[f(t + t_1)] = e^{ist_1} g(s)$
7.  $F[f(t)] = G(is) + G(-is)$  if  $f(t) = f(-t)$  ( $f(t)$  even)  $F[f(t)] = G(is) - G(-is)$  if  $f(t) = -f(-t)$  ( $f$  odd)

where  $G(s) = L[f(t)]$  . This result allows the use of the Laplace transform tables to obtain the Fourier transforms.

**Example 204.8.** Find  $F[e^{-a|t|}]$  by property 7. The term  $e^{-a|t|}$  is even. So  $L[e^{-at}] = 1/(s + a)$  . Therefore,  $F[e^{-a|t|}] = 1/(is + a) + 1/(-is + a) = 2a/(s^2 + a^2)$  .

## 204.4 Fourier Cosine Transform

The *Fourier cosine transform* is given by  $F_c[f(t)] = g(s) = \sqrt{(2/\pi)} \int_0^\infty f(t) \cos st dt$  and its inverse by  $F_c^{-1}[g(s)] = f(t) = \sqrt{(2/\pi)} \int_0^\infty g(s) \cos st ds$  . The *Fourier sine transform*  $F_s$  is obtainable by replacing the cosine by the sine in the above integrals.

**Example 204.9.**  $F_c[f(t)]$ ,  $f(t) = 1$  for  $0 < t < a$  and 0 for  $a < t < \infty$  .  
 $F_c[f(t)] = \sqrt{(2/\pi)} \int_0^a \cos st dt = \sqrt{(2/\pi)} (\sin as)/s$  .

## Properties of the Fourier Cosine Transform

$F_c[f(t)] = g(s)$  .

1.  $F_c[af(t) + bh(t)] = aF_c[f(t)] + bF_c[h(t)]$
2.  $F_c[f(at)] = (1/a) g(s/a)$
3.  $F_c[f(at) \cos bt] = 1/2a [g((s + b)/a) + g((s - b)/a)]$  ,  $a, b > 0$
4.  $F_c[t^{2n} f(t)] = (-1)^n (d^{2n} g)/(ds^{2n})$
5.  $F_c[t^{2n+1} f(t)] = (-1)^n (d^{2n+1} g)/(ds^{2n+1}) F_s[f(t)]$

Table 204.3 presents some Fourier cosine transforms.

**Table 204.3**

$f(t)$	$\frac{g(s)}{\sqrt{2/\pi}}$
$\left. \begin{array}{ll} t & 0 < t < 1 \\ 2 - t & 1 < t < 2 \\ 0 & 2 < t < \infty \end{array} \right\}$	$\frac{1}{s^2} [2 \cos s - 1 - \cos 2s]$
$t^{-1/2}$	$\pi^{1/2} (2s)^{-1/2}$
$\left. \begin{array}{ll} 0 & 0 < t < a \\ (t - a)^{-1/2} & a < t < \infty \end{array} \right\}$	$\pi^{1/2} (2s)^{-1/2} [\cos as - \sin as]$
$(t^2 + a^2)^{-1}$	$\frac{1}{2} \pi a^{-1} e^{-as}$
$e^{-at}, \quad a > 0$	$\frac{a}{s^2 + a^2}$
$e^{-at^2}, \quad a > 0$	$\frac{1}{2} \pi^{1/2} a^{-1/2} e^{-s^2/4a}$
$\frac{\sin at}{t}, \quad a > 0$	$\begin{cases} \pi/2 & s < a \\ \pi/4 & s = a \\ 0 & s > a \end{cases}$

**Example 204.10.** The temperature  $\theta$  in the semiinfinite rod  $0 \leq x < \infty$  is determined by the differential equation  $\partial\theta/\partial t = k(\partial^2\theta/\partial x^2)$  and the condition  $\theta = 0$  when  $t = 0, x \geq 0$ ;  $\partial\theta/\partial x = -\mu = \text{constant}$  when  $x = 0, t > 0$ . By using the Fourier cosine transform, a solution may be found as  $\theta(x, t) = (2\mu/\pi) \int_0^\infty (\cos px/p) (1 - e^{-kp^2t}) dp$ .

## References

- Churchill, R. V. 1958. *Operational Mathematics*. McGraw-Hill, New York.
- Ditkin, B. A. and Proodnikov, A. P. 1965. *Handbook of Operational Mathematics* (in Russian). Nauka, Moscow.
- Doetsch, G. 1950–1956. *Handbuch der Laplace Transformation*, vols. I–IV (in German). Birkhauser, Basel.
- Nixon, F. E. 1960. *Handbook of Laplace Transforms*. Prentice Hall, Englewood Cliffs, NJ.
- Sneddon, I. 1951. *Fourier Transforms*. McGraw-Hill, New York.
- Widder, D. 1946. *The Laplace Transform*. Princeton University Press, Princeton, NJ.

## Further Information

The references citing G. Doetsch, *Handbuch der Laplace Transformation*, vols. I–IV, Birkhauser, Basel, 1950–1956 (in German) and B. A. Ditkin and A. P. Proodnikov, *Handbook of Operational Mathematics*, Moscow, 1965 (in Russian) are the most extensive tables known. The latter reference is 485 pages.

Deliu, A. "Chaos, Fractals, and Julia Sets"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Chaos, Fractals, and Julia Sets

---

- 205.1 Chaos
- 205.2 Fractals
- 205.3 Julia Sets

## Anca Deliu

*Georgia Institute of Technology*

The mathematical foundations of chaotic dynamics go back about 100 years to the work of the French mathematician Henri Poincaré, who studied the time evolution of three celestial bodies subject to mutual gravitational forces. He showed that very irregular orbits (called *chaotic* in today's terminology) could arise due to the nonlinearities present in the problem.

The explosive interest in chaos (or nonlinear science) in the last two decades is largely due to the availability of powerful computers with high-quality graphics. Researchers from fields other than mathematics have become aware that simple and familiar phenomena from physics and engineering, as well as economics, biology, and geology, can lead to unpredictable time evolutions, despite their deterministic nature.

As opposed to classical systems, which eventually settle into periodic motion or steady state, chaotic systems evolve apparently at random on a fractal attractor. The complexity of the system is related to the geometry of the attractor. The study of individual trajectories becomes irrelevant and is replaced by a statistical analysis.

Examples of attractors discovered in recent years are the Lorentz attractor, the Henon attractor, and the Smale horseshoe attractor [[Falconer, 1990](#)].

In addition to these are the classical Cantor and Julia sets which, one might say, are old examples for a new science [[Barnsley, 1988](#)].

## 205.1 Chaos

---

The goal of the study of dynamical systems is to determine the long-time evolution of the system. In practice one analyzes the evolution of an observable (a measurement)  $x$ , associated with the system, such as temperature, pressure, or population. According to whether time is discrete or continuous, we have discrete or continuous dynamical systems. Chaotic behavior can occur in continuous systems with phase spaces of dimension  $\geq 3$  and in discrete systems in dimension  $\geq 1$ .

Consider the difference equation

$$x_{n+1} = f(x_n) \quad (205.1)$$

where  $X \subset \mathbb{R}^n$  and  $f : X \rightarrow X$ , a mapping with some regularity. The pair  $(f, X)$  is a *discrete dynamical system* with *phase space*  $X$ . The orbit of  $x_0 \in X$  is the sequence of iterates  $x_k = f^k(x_0) : k \geq 0$ . The term  $x_k$  represents the value of the observable  $x$  at time  $k$ .

A point  $y$  is a *periodic point* of *period*  $p$  if  $f^p(y) = y$ . Then  $y$  has a *periodic orbit* consisting of finitely many points  $\{y, f(y), \dots, f^{p-1}(y)\}$ . A periodic point of period 0 is called a *fixed point*. Periodic points play an important role in understanding the dynamics of  $f$ .

**Example 205.1.** Let  $X = \mathbb{R}$  and  $f(x) = \cos x$ . Then the orbits of all  $x_0 \in \mathbb{R}$  converge to the unique fixed point of  $\cos x$ ,  $x \approx 0.739$ , as one can check by pressing repeatedly the  $\cos$  button on a calculator.

A closed and bounded subset  $A \subset X$  is an *attractor* if  $A$  is *invariant*—that is,  $f(A) = A$ —and  $A$  is *attracting*—that is, the distance between  $f^k(x_0)$  and  $A$  converges to zero as  $k \rightarrow \infty$ —for all points  $x_0 \in V$  in some neighborhood  $V$  of  $A$ . The set  $V$  is called the *basin of attraction* of  $A$ .

A *repeller* for  $f$  is a closed bounded set from which nearby initial conditions are running away under iteration. Speaking loosely, one can identify attractors and repellers by reversing the arrow of time.

A fractal attractor is called a *strange attractor* [Ruelle and Takens, 1971] to explain the onset of turbulence.

Dynamics on a strange attractor is usually chaotic. The word *chaos* was introduced by Li and Yorke. They proved that if a map has period 3 then it has all other periods; they summarized this result in the title of the paper, "Period Three Implies Chaos" [Li and Yorke, 1975]. The current technical meaning of *chaos* is, however, different.

We say that  $f$  is *chaotic* if the following three conditions hold.

1. There is a point  $x$  whose orbit is dense in  $A$ —that is, it comes arbitrarily close to any point in  $A$ .
2. The periodic points of  $f$  are dense in  $A$ .
3. The function  $f$  has sensitive dependence on initial conditions; that is, there is a number  $\delta > 0$ , such that for any  $x \in X$  there are points  $y$  arbitrarily close to  $x$ , such that  $|f^k(x) - f^k(y)| \geq \delta$  for some  $k$ .

It is convenient sometimes to study a whole family of dynamical systems,  $(X, f_\lambda)$ , indexed by an experimental control parameter  $\lambda$ . Varying  $\lambda$  changes the quality of the dynamics (the geometric nature of the attractor) through a sequence of transitions that may lead to chaos (turbulence). Several scenarios for the onset of chaos have been proposed [Ruelle, 1989]:

1. Ruelle-Takens scenario through quasiperiodicity
2. Feigenbaum scenario through periodic doubling
3. Pomeau-Manneville scenario through intermittency

The next example illustrates scenario 2.

**Example 205.2.** Let  $X = [0, 1]$  and let  $f_\lambda : X \rightarrow X$  be the logistic family  $f_\lambda(x) = \lambda x(1 - x)$  (205.2)

originally introduced to model the development of certain populations. If the population is  $x_k$  at the end of the  $k$ th year, it is assumed to be  $x_{k+1} = f_\lambda(x_k)$  at the end of the  $(k + 1)$  th year.

As  $\lambda$  increases from 0 to 4,  $f_\lambda$  presents a wide spectrum of behavior, ranging from the simplest to the most complex. [Falconer, 1990, p. 176] .

For  $0 < \lambda < 3$  ,  $f$  has a stable fixed point,  $x_\lambda = 1 - 1/\lambda$  , which attracts all orbits starting at  $x_0 \in (0, 1)$  .

When  $\lambda$  increases through the value  $\lambda = \lambda_1 = 3$  , the stable fixed point  $x_\lambda$  splits into a stable period 2 orbit,  $\{y_\lambda^1, y_\lambda^2\}$  . For example, if  $\lambda = 3.38$  , almost all trajectories settle asymptotically into this periodic cycle.

At  $\lambda = \lambda_2 = 1 + \sqrt{6} \approx 3.45$  , the period 2 orbit becomes unstable and splits into a stable period 4 orbit. Let  $\lambda_k$  denote the value of  $\lambda$  at which the  $k$ th period doubling occurs. Then  $\lambda_k$  increases to the critical value  $\lambda_\infty \approx 3.57$  , where chaos sets in.

For  $\lambda \in (\lambda_\infty, 4)$  chaos and periodicity alternate, depending on specific window values of  $\lambda$ .

Feigenbaum discovered that  $\lambda_\infty$  is a universal constant, in the sense that

$$\lambda_\infty - \lambda_k \approx C\delta^k \quad (205.3)$$

for very general families of functions.

If the time step is allowed to converge to zero, Eq. (205.1) becomes an autonomous (time-independent) differential equation,

$$\frac{dx}{dt} = f(x)$$

which determines a *continuous dynamical system* . The family of solution curves corresponding to various initial conditions  $x_0 \in X$  represents the trajectories on which the system evolves as time  $t \rightarrow \pm\infty$  .

**Example 205.3.** The most famous continuous system with chaotic behavior is described by Lorentz [1963], who studied Rayleigh-Benard convection. A fluid contained between two rigid plates is subjected to gravity. The top plate is maintained at the temperature  $T_0$  and the bottom plate is maintained at a higher temperature,  $T_0 + \Delta T$  . Experiment shows that, for a certain range of small values of  $\Delta T$  , the fluid will execute a convective cellular flow. For bigger values of  $\Delta T$  the flow becomes chaotic. This is similar to what happens in the earth's atmosphere. The specific equations of Lorentz are

$$\frac{dx}{dt} = -\sigma x + \sigma y$$

$$\frac{dy}{dt} = -xy + rx - y$$

$$\frac{dz}{dt} = xy - bz$$

where  $\sigma = 10$ ,  $r = 28$ , and  $b = 8/3$ .

The system has sensitive dependence on initial conditions: Orbits that start close soon become uncorrelated, which explains the well-observed fact that the weather is essentially unpredictable for periods longer than five days.

The Lorentz strange attractor lives in  $R^3$  and its two-dimensional projections have a characteristic butterfly shape [Falconer, 1990].

## 205.2 Fractals

---

The term *fractal* was coined by Mandelbrot [1982] to denote sets with fractured structure. He proposed the idea that fractals are the rule rather than the exception in nature: trees, coastlines, flames, riverbeds, and so forth all exhibit fractal behavior.

A fractal is a set whose topological dimension is *strictly* smaller than its Hausdorff dimension.

The topological dimension,  $\text{TDim}$ , of a set takes only integer values and is, roughly speaking, what we intuitively mean by dimension:  $\text{TDim}(\text{point}) = 0$ ;  $\text{TDim}(\text{continuous curve}) = 1$ ;  $\text{TDim}(\text{surface}) = 2$ , and so forth.

The Hausdorff dimension provides a way to distinguish between topologically equivalent attractors with various levels of geometric complexity.

The diameter of  $U \subset R^n$  is  $|U| = \sup\{|x - y| : x, y \in U\}$ . Let  $F \subset R^n$ . For  $\varepsilon > 0$ , an  $\varepsilon$ -cover of  $F$  is any countable collection of sets  $U_i$  with  $|U_i| < \varepsilon$ , and such that  $F \subset \cup_i U_i$ . Let  $s \geq 0$ . For any  $\varepsilon > 0$ , we define

$$\mathcal{H}_\varepsilon^s(F) = \inf \left\{ \sum_{i=1}^{\infty} |U_i|^s : \{U_i\} \text{ is an } \varepsilon\text{-cover of } F \right\}$$

As  $\varepsilon$  decreases to 0,  $\mathcal{H}_\varepsilon^s(F)$  increases and thus approaches a limit,

$$\mathcal{H}^s(F) = \lim_{\varepsilon \rightarrow 0} \mathcal{H}_\varepsilon^s(F)$$

which we define to be the *s-dimensional Hausdorff measure* of  $F$ .

For any subset  $F \subset R^n$ , there is a unique number  $D = \text{HDim}(F) \geq 0$  such that for  $0 \leq s < D$ ,  $\mathcal{H}^s(F) = \infty$  and for  $D < s < \infty$ ,  $\mathcal{H}^s(F) = 0$ . This number  $D$  is called the *Hausdorff dimension* of  $F$ . If  $F$  has Hausdorff dimension  $D$ , the  $D$ -dimensional Hausdorff measure of  $F$  may take any value in  $[0, \infty]$ . It always holds that  $\text{TDim} \leq \text{HDim}$ .

The Hausdorff dimension is difficult to estimate, both theoretically and numerically. As a substitute, one often uses a simpler concept, the *fractal dimension*, which conveys similar information about the geometry of a set yet is easier to compute.

Let  $F \subset R^n$  be a nonempty bounded subset of some euclidian space, and let  $\mathcal{N}(F, \varepsilon)$  be the smallest number of cubes of side length equal to  $\varepsilon$  required to cover  $F$ . The fractal dimension (also called *box dimension*, *Kolmogorov entropy*, or *capacity*) is defined as the limit



$$\text{FDim}(F) = \lim_{\varepsilon \rightarrow 0} \frac{\log \mathcal{N}(F, \varepsilon)}{-\log \varepsilon}$$

For any set,  $\text{TDim} \leq \text{HDim} \leq \text{FDim}$  .

The dimension of the Lorentz attractor was estimated numerically to be around 2.04 and the dimension of the attractor corresponding to  $\lambda_\infty$  , from Example 205.2, was estimated numerically to be  $\approx 0.538$  [ [Falconer, 1990, p. 175](#)].

The only class of fractals whose dimension is given by a simple formula is that of *self-similar sets*.

A map  $W : R^n \rightarrow R^n$  is called a *contractive similitude* if  $|W(x) - W(y)| = s|x - y|$  , for any  $x, y \in R^n$  , for some  $0 < s < 1$  . The number  $s$  is the *contraction ratio* of  $W$ .

A set  $A \subset R^n$  is *self-similar* if it can be tiled with smaller copies of itself,

$$A = \cup_{i=1}^N W_i(A) \quad (205.4)$$

where  $\{W_i\}$  are contractive similitudes.

If the nonoverlapping condition  $W_i(A) \cap W_j(A) = \emptyset$  holds, then

$$\text{HDim}(A) = \text{FDim}(A) = D \quad (205.5)$$

where  $D$  is the unique real solution to the equation  $\sum_{i=1}^N s_i^D = 1$  and  $s_i$  is the contraction ratio of  $W_i$ .

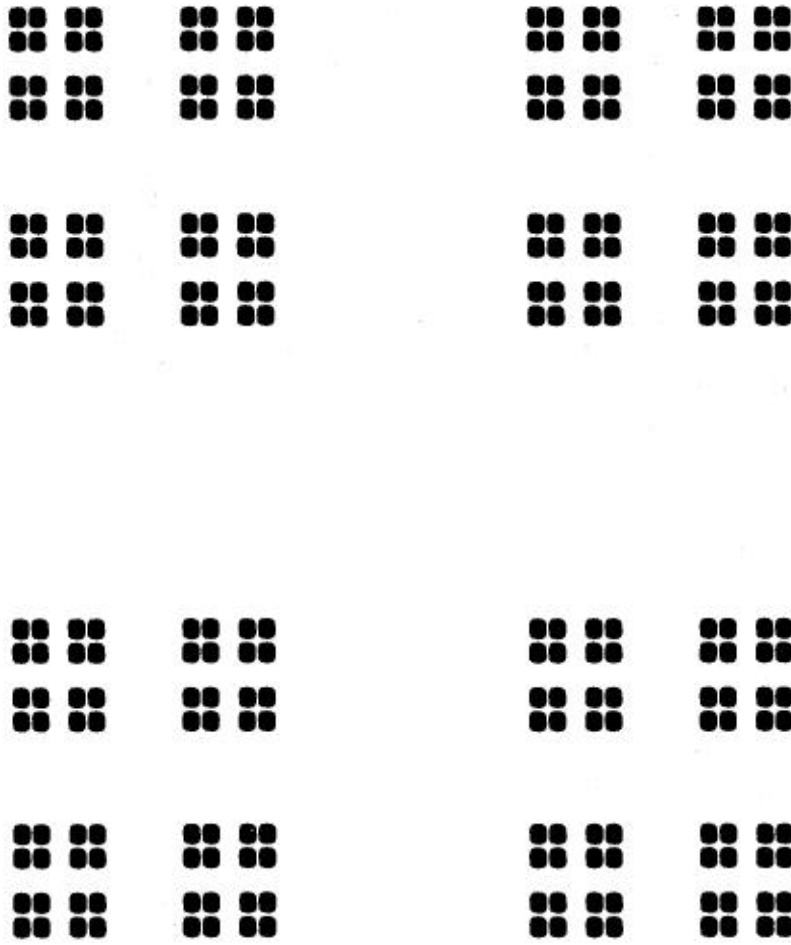
The simplest example of a self-similar set is the ternary Cantor set  $C$  ([Fig. 205.1](#)). From the closed interval  $[0, 1]$  , remove the middle third interval  $(\frac{1}{3}, \frac{2}{3})$  . Two closed intervals remain— $[0, \frac{1}{3}]$  and  $[\frac{2}{3}, 1]$  . Repeat the procedure for all remaining intervals iteratively. The Cantor set  $C$  is the fractal "dust" that remains. All points of the form  $\{k3^{-n} : n \geq 0, 0 \leq k \leq 3^n - 1\}$  belong to  $C$ , but there are many more.

**Figure 205.1** Ternary Cantor set  $C$ .



[Figure 205.2](#) illustrates the analogous two-dimensional construction, which is easier to picture.

**Figure 205.2** Two-dimensional construction of Cantor set  $C$ .



For the Cantor set in Fig. 205.1,  $W_1 = x/3$  and  $W_2 = x/3 + 2/3$ ,  $s_1 = s_2 = 1/3$ , hence Eq. (205.5) yields  $\text{HDim}(C) = \text{FDim}(C) = \log 2 / \log 3$ . For the Cantor set in Fig. 205.2,  $N = 4$ ,  $\{s_i = 1/3 : i = 1, \dots, 4\}$  and Eq. (205.5) yields  $\text{HDim} = \text{FDim} = \log 4 / \log 3$ . The topological dimension of both Cantor sets is 0, a feature common to all "dust" (*totally disconnected*) sets.

## 205.3 Julia Sets

Julia sets are fractals with most singular shapes that are generated by the iteration of simple complex mappings. For instance, consider the quadratic polynomials  $f(z) = z^2 + c$ ,  $z \in C$ ,  $c \in C$  a complex parameter.

The *Julia set*  $J(f)$  of  $f$  is the closure of the set of repelling periodic points of  $f$ . The complement of the Julia set is called the *Fatou* (or *stable*) set. One can show that  $J(f)$  is nonempty, is closed and bounded, has empty interior, and is generally a fractal. The Julia set  $J = J(f)$  is both forward- and backward-invariant,  $J = f(J) = f^{-1}(J)$ , and  $f$  is chaotic on  $J$ .

**Example 205.4.** For  $c = 0$ ,  $f(z) = z^2$  and the corresponding Julia set  $J(f)$  is the unit circle  $\{|z| = 1\}$ .

**Example 205.5.** Let  $c \neq 0$  be a complex number so that  $|c|$  is small. Then the Julia set  $J(f)$  is a closed fractal curve, close to the unit circle.

In these two examples  $J(f)$  is the boundary between two regions of  $\mathbb{C}$  with distinct behaviors: if  $z_0$  is inside the Julia set curve, the iterates  $f^k(z_0)$  converge to the finite fixed point of  $f$ , and if  $z_0$  is outside the curve its iterates diverge to  $\infty$ .

One can show that if  $|c| < \frac{1}{4}$  then  $J(f)$  is a closed curve, whereas if  $|c| > \frac{1}{4}(5 + 2\sqrt{6})$ ,  $J(f)$  is totally disconnected. For several pictures of Julia sets, see [Falconer 1990, p. 214].

## References

- Barnsley, M. F. 1988. *Fractals Everywhere*. Academic Press, New York.
- Devaney, R. L. 1989. *An Introduction to Chaotic Dynamical Systems*. Addison-Wesley, Reading, MA.
- Falconer, K. J. 1990. *Fractal Geometry, Mathematical Foundations and Applications*. John Wiley & Sons, New York.
- Li, T. Y. and Yorke, J. A. 1975. Period three implies chaos. *Amer. Math. Monthly*. 82:985.
- Lorentz, E. N. 1963. Deterministic nonperiodic flow. *J. Atmos. Sci.* 20:130.
- Mandelbrot, B. B. 1982. *The Fractal Geometry of Nature*. Freeman, New York.
- Ott, E. 1993. *Chaos in Dynamical Systems*. Cambridge University Press, New York.
- Ruelle, D. 1989. *Chaotic Evolution and Strange Attractors*. Cambridge University Press,
- Ruelle, D. and Takens, F. 1971. On the nature of turbulence. *Comm. Math. Phys.* 20:167.

Ames, W. F. "Calculus of Variations"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

[206.1 The Euler Equation](#)[206.2 The Variation](#)[206.3 Constraints](#)**William F. Ames***Georgia Institute of Technology*

The basic problem in the *calculus of variations* is to determine a function such that a certain *functional*, often an integral involving that function and certain of its derivatives, takes on *maximum or minimum values*. As an example, find the function  $y(x)$  such that  $y(x_1) = y_1$ ,  $y(x_2) = y_2$  and the integral (functional)  $I = 2\pi \int_{x_1}^{x_2} y[1 + (y')^2]^{1/2} dx$  is a minimum. A second example concerns the transverse deformation  $u(x, t)$  of a beam. The energy functional  $I = \int_{t_1}^{t_2} \int_0^L [\frac{1}{2} \rho (\partial u / \partial t)^2 - \frac{1}{2} EI (\partial^2 u / \partial x^2)^2 + fu] dx dt$  is to be minimized.

---

**206.1 The Euler Equation**

---

The elementary part of the theory is concerned with a *necessary* condition (generally in the form of a differential equation with boundary conditions) that the required function must satisfy. To show mathematically that the function obtained actually maximizes (or minimizes) the integral is much more difficult than the corresponding problems of the differential calculus.

The *simplest case* is to determine a function  $y(x)$  that makes the integral  $I = \int_{x_1}^{x_2} F(x, y, y') dx$  stationary and that satisfies the prescribed end conditions  $y(x_1) = y_1$  and  $y(x_2) = y_2$ . Here we suppose  $F$  has continuous second partial derivatives with respect to  $x$ ,  $y$ , and  $y' = dy/dx$ . If  $y(x)$  is such a function, then it must satisfy the *Euler equation*  $(d/dx)(\partial F / \partial y') - (\partial F / \partial y) = 0$ , which is the required necessary condition. The indicated partial derivatives have been formed by treating  $x$ ,  $y$ , and  $y'$  as independent variables. Expanding the equation, the equivalent form

$F_{y'y'} y'' + F_{y'y} y' + (F_{y'x} - F_y) = 0$  is found. This is second order in  $y$  unless

$F_{y'y'} = (\partial^2 F) / [(\partial y')^2] = 0$ . An alternative form

$1/y' [d/dx(F - (\partial F / \partial y')(dy/dx)) - (\partial F / \partial x)] = 0$  is useful. Clearly, if  $F$  does not involve  $x$  explicitly  $[(\partial F / \partial x) = 0]$  a first integral of Euler's equation is  $F - y' (\partial F / \partial y') = c$ . If  $F$  does not involve  $y$  explicitly  $[(\partial F / \partial y) = 0]$  a first integral is  $(\partial F / \partial y') = c$ .

The Euler equation for  $I = 2\pi \int_{x_1}^{x_2} y[1 + (y')^2]^{1/2} dx$ ,  $y(x_1) = y_1$ ,  $y(x_2) = y_2$  is  $(d/dx) [yy' / [1 + (y')^2]^{1/2}] - [1 + (y')^2]^{1/2} = 0$  or after reduction  $yy'' - (y')^2 - 1 = 0$ . The

solution is  $y = c_1 \cosh(x/c_1 + c_2)$ , where  $c_1$  and  $c_2$  are integration constants. Thus the required minimal surface, if it exists, must be obtained by revolving a catenary. Can  $c_1$  and  $c_2$  be chosen so that the solution passes through the assigned points? The answer is found in the solution of a transcendental equation that has two, one, or no solutions, depending on the prescribed values of  $y_1$  and  $y_2$ .

## 206.2 The Variation

If  $F = F(x, y, y')$ , with  $x$  independent and  $y = y(x)$ , then the *first variation*  $\delta F$  of  $F$  is defined to be  $\delta F = (\partial F / \partial x) \delta y + (\partial F / \partial y) \delta y'$  and  $\delta y' = \delta(dy/dx) = (d/dx)(\delta y)$ ; —that is, they commute. Note that the first variation,  $\delta F$ , of a functional is a first-order change from curve to curve, whereas the differential of a function is a first-order approximation to the change in that function along a *particular curve*. The laws of  $\delta$  are as follows:  $\delta(c_1 F + c_2 G) = c_1 \delta F + c_2 \delta G$ ;  $\delta(FG) = F \delta G + G \delta F$ ;  $\delta(F/G) = (G \delta F - F \delta G)/G^2$ ; if  $x$  is an independent variable,  $\delta x = 0$ ; if  $u = u(x, y)$ ;  $(\partial/\partial x)(\delta u) = \delta(\partial u/\partial x)$ ,  $(\partial/\partial y)(\delta u) = \delta(\partial u/\partial y)$ .

A necessary condition that the integral  $I = \int_{x_1}^{x_2} F(x, y, y') dx$  be stationary is that its (first) variation vanish—that is,  $\delta I = \delta \int_{x_1}^{x_2} F(x, y, y') dx = 0$ . Carrying out the variation and integrating by parts yields  $\delta I = \int_{x_1}^{x_2} [(\partial F/\partial y) - (d/dx)(\partial F/\partial y')] \delta y dx + [(\partial F/\partial y') \delta y]_{x_1}^{x_2} = 0$ . The arbitrary nature of  $\delta y$  means the square bracket must vanish and the last term constitutes the *natural boundary conditions*.

**Example.** The Euler equation of  $\int_{x_1}^{x_2} F(x, y, y', y'') dx$  is  $(d^2/dx^2)(\partial F/\partial y'') - (d/dx)(\partial F/\partial y') + (\partial F/\partial y) = 0$ , with natural boundary conditions  $\{[(d/dx)(\partial F/\partial y'') - (\partial F/\partial y')] \delta y\}_{x_1}^{x_2} = 0$  and  $(\partial F/\partial y'') \delta y' g|_{x_1}^{x_2} = 0$ . The Euler equation of  $\int_{x_1}^{x_2} \int_{y_1}^{y_2} F(x, y, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy}) dx dy$  is  $(\partial^2/\partial x^2)(\partial F/\partial u_{xx}) + (\partial^2/\partial x \partial y)(\partial F/\partial u_{xy}) + (\partial^2/\partial y^2)(\partial F/\partial u_{yy}) - (\partial/\partial x)(\partial F/\partial u_x) - (\partial/\partial y)(\partial F/\partial u_y) + (\partial F/\partial u) = 0$ , and the natural boundary conditions are

$$\left[ \left( \frac{\partial}{\partial x} \left( \frac{\partial F}{\partial u_{xx}} \right) + \frac{\partial}{\partial y} \left( \frac{\partial F}{\partial u_{xy}} \right) - \frac{\partial F}{\partial u_x} \right) \delta u \right]_{x_1}^{x_2} = 0, \quad \left[ \left( \frac{\partial F}{\partial u_{xx}} \right) \delta u_x \right]_{x_1}^{x_2} = 0$$

$$\left[ \left( \frac{\partial}{\partial y} \left( \frac{\partial F}{\partial u_{yy}} \right) + \frac{\partial}{\partial x} \left( \frac{\partial F}{\partial u_{xy}} \right) - \frac{\partial F}{\partial u_y} \right) \delta u \right]_{y_1}^{y_2} = 0, \quad \left[ \left( \frac{\partial F}{\partial u_{yy}} \right) \delta u_y \right]_{y_1}^{y_2} = 0$$

In the more general case of  $I = \iiint_R F(x, y, u, v, u_x, u_y, v_x, v_y) dx dy dz$ , the condition  $\delta I = 0$  gives rise to the two Euler equations  $(\partial/\partial x)(\partial F/\partial u_x) + (\partial/\partial y)(\partial F/\partial u_y) - (\partial F/\partial u) = 0$  and  $(\partial/\partial x)(\partial F/\partial v_x) + (\partial/\partial y)(\partial F/\partial v_y) - (\partial F/\partial v) = 0$ . These are two PDEs in  $u$  and  $v$  that are linear or quasi-linear in  $u$  and  $v$ . The Euler equation for  $I = \iiint_R (u_x^2 + u_y^2 + u_z^2) dx dy dz$ , from  $\delta I = 0$ , is Laplace's equation  $u_{xx} + u_{yy} + u_{zz} = 0$ .

Variational problems are easily derived from the differential equation and associated boundary conditions by multiplying by the variation and integrating the appropriate number of times. To illustrate, let  $F(x)$ ,  $\rho(x)$ ,  $p(x)$ , and  $w$  be the tension, the linear mass density, the natural load, and (constant) angular velocity of a rotating string of length  $L$ . The equation of motion is  $(d/dx)[F(dy/dx)] + \rho w^2 y + p = 0$ . To formulate a corresponding variational problem, multiply all terms by a variation  $\delta y$  and integrate over  $(0, L)$  to obtain

$$\int_0^L \frac{d}{dx} \left( F \frac{dy}{dx} \right) \delta y dx + \int_0^L \rho w^2 y \delta y dx + \int_0^L p \delta y dx = 0$$

The second and third integrals are the variations of  $\frac{1}{2} \rho w^2 y^2$  and  $py$ , respectively. To treat the first integral, integrate by parts to obtain

$$\left[ F \frac{dy}{dx} \delta y \right]_0^L - \int_0^L F \frac{dy}{dx} \delta \frac{dy}{dx} dx = \left[ F \frac{dy}{dx} \delta y \right]_0^L - \int_0^L \frac{1}{2} F \delta \left( \frac{dy}{dx} \right)^2 dx = 0$$

So the variation formulation is

$$\delta \int_0^L \left[ \frac{1}{2} \rho w^2 y^2 + py - \frac{1}{2} F \left( \frac{dy}{dx} \right)^2 \right] dx + \left[ F \frac{dy}{dx} \delta y \right]_0^L = 0$$

The last term represents the *natural boundary conditions*. The term  $\frac{1}{2} \rho w^2 y^2$  is the kinetic energy per unit length, the term  $-py$  is the potential energy per unit length due to the radial force  $p(x)$ , and the term  $\frac{1}{2} F(dy/dx)^2$  is a first approximation to the potential energy per unit length due to the tension  $F(x)$  in the string. Thus the integral is often called the *energy integral*.

## 206.3 Constraints

The variations in some cases cannot be arbitrarily assigned because of one or more auxiliary conditions that are usually called *constraints*. A typical case is the functional  $\int_{x_1}^{x_2} F(x, u, v, u_x, v_x) dx$  with a constraint  $\phi(u, v) = 0$  relating  $u$  and  $v$ . If the variations of  $u$  and  $v$  ( $\delta u$  and  $\delta v$ ) vanish at the end points, then the variation of the integral becomes

$$\int_{x_1}^{x_2} \left\{ \left[ \frac{\partial F}{\partial u} - \frac{d}{dx} \left( \frac{\partial F}{\partial u_x} \right) \right] \delta u + \left[ \frac{\partial F}{\partial v} - \frac{d}{dx} \left( \frac{\partial F}{\partial v_x} \right) \right] \delta v \right\} dx = 0$$

The variation of the constraint  $\phi(u, v) = 0$ ,  $\phi_u \delta u + \phi_v \delta v = 0$  means that the variations cannot both be assigned arbitrarily inside  $(x_1, x_2)$ , so their coefficients need not vanish separately.

Multiply  $\phi_u \delta u + \phi_v \delta v = 0$  by a Lagrange multiplier  $\lambda$  (may be a function of  $x$ ) and integrate to find  $\int_{x_1}^{x_2} (\lambda \phi_u \delta u + \lambda \phi_v \delta v) dx = 0$ . Adding this to the previous result yields

$$\int_{x_1}^{x_2} \left\{ \left[ \frac{\partial F}{\partial u} - \frac{d}{dx} \left( \frac{\partial F}{\partial u_x} \right) + \lambda \phi_u \right] \delta u + \left[ \frac{\partial F}{\partial v} - \frac{d}{dx} \left( \frac{\partial F}{\partial v_x} \right) + \lambda \phi_v \right] \delta v \right\} dx = 0$$

which must hold for any  $\lambda$ . Assign  $\lambda$  so the first square bracket vanishes. Then  $\delta v$  can be assigned to vanish inside  $(x_1, x_2)$ , so the two systems

$$\frac{d}{dx} \left[ \frac{\partial F}{\partial u_x} \right] - \frac{\partial F}{\partial u} - \lambda \phi_u = 0, \quad \frac{d}{dx} \left[ \frac{\partial F}{\partial v_x} \right] - \frac{\partial F}{\partial v} - \lambda \phi_v = 0$$

plus the constraint  $\phi(u, v) = 0$  are three equations for  $u$ ,  $v$  and  $\lambda$ .

## References

- Gelfand, I. M. and Fomin, S. V. 1963. *Calculus of Variations*. Prentice Hall, Englewood Cliffs, NJ.
- Lanczos, C. 1949. *The Variational Principles of Mechanics*. Univ. of Toronto Press, Toronto.
- Schechter, R. S. 1967. *The Variational Method in Engineering*. McGraw-Hill, New York.
- Vujanovic, B. D. and Jones, S. E. 1989. *Variational Methods in Nonconservative Phenomena*. Academic Press, New York.
- Weinstock, R. 1952. *Calculus of Variations, with Applications to Physics and Engineering*. McGraw-Hill, New York.



Tong, Y. L. “Force-System Resultants and Equilibrium”  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

**207.1 Elementary Probability**

Random Variables and Probability Distributions • Expectations • Some Commonly Used Distributions • The Normal Distribution

**207.2 Random Sample and Sampling Distributions**

Random Sample and Related Statistics

**207.3 Normal Distribution-Related Sampling Distributions**

One-Sample Case • Two-Sample Case

**207.4 Confidence Intervals**

One-Sample Case • Two-Sample Case

**207.5 Testing Statistical Hypotheses**

One-Sample Case • Two-Sample Case

**207.6 A Numerical Example**

One-Sample Case • Two-Sample Case

**Y. L. Tong**

*Georgia Institute of Technology*

In most engineering experiments, the outcomes (and hence the observed data) appear in a random and nondeterministic fashion. For example, the operating time of a system before failure, the tensile strength of a certain type of material, and the number of defective items in a batch of produced items are all subject to random variations from one experiment to another. In engineering statistics, we apply the theory and methods of statistics to develop procedures for summarizing the data and making statistical inference and to obtain useful information with the presence of randomness and uncertainty.

---

**207.1 Elementary Probability**

---

**Random Variables and Probability Distributions**

Intuitively speaking, a random variable (denoted by  $X$ ,  $Y$ ,  $Z$ , etc.) takes a numerical value that depends on the outcome of the experiment. Because the outcome of an experiment is subject to random variation, the resulting numerical value is also random. In order to provide a stochastic model for describing the probability distribution of a random variable  $X$ , we generally classify random variables into two groups—the discrete type and the continuous type. The discrete random variables are those which, technically speaking, take a finite number or a countably infinite number

of possible numerical values (in most engineering applications, they take nonnegative integer values). Continuous random variables involve outcome variables such as time, length, distance, area, and volume. We specify a function  $f(x)$ , called the probability density function (p.d.f.) of a random variable  $X$ , such that the probability that the random variable  $X$  takes a value in a set  $A$  (of real numbers) is given by

$$P[X \in A] = \begin{cases} \sum_{x \in A} f(x) & \text{for all sets } A \text{ if } X \text{ is discrete} \\ \int_A f(x) dx & \text{for all intervals } A \text{ if } X \text{ is continuous} \end{cases} \quad (207.1)$$

By letting  $A$  be the set of all values that are less than or equal to a fixed number  $t$  (i.e.,  $A = (-\infty, t]$ ), the probability function  $P[X \leq t]$ , denoted by  $F(t)$ , is called the distribution function of  $X$ . We note that, by calculus, if  $X$  is a continuous random variable and if  $F(x)$  is differentiable, then  $f(x) = (d/dx)F(x)$ .

## Expectations

In many applications, the result of an experiment with a numerical outcome  $X$  is a specific function of  $X[u(X)$ , say]. Because  $X$  is a random variable,  $u(X)$  itself is also a random variable. We define the expected value of  $u(X)$  by

$$E[u(X)] = \begin{cases} \sum_x u(x)f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} u(x)f(x) dx & \text{if } X \text{ is continuous} \end{cases} \quad (207.2)$$

provided that, of course, the sum or the integral exists. In particular, if  $u(x) = x$ , then  $E(X) \equiv \mu$  is called the mean of  $X$  (of the distribution) and  $E(X - \mu)^2 \equiv \sigma^2$  is called the variance of  $X$  (of the distribution). The mean is a measurement of the central tendency, and the variance is a measurement of the dispersion of the distribution.

## Some Commonly Used Distributions

There are many well-known distributions which are useful in engineering statistics. Among the discrete distributions, the hypergeometric and binomial distributions have applications in acceptance sampling problems and quality control, and the Poisson distribution is useful for studying queueing theory and other related problems. Among the continuous distributions, the uniform distribution concerns random numbers and can be applied in simulation studies, the exponential and gamma distributions are closely related to the Poisson distribution and, together with the Weibull distribution, have important applications in life testing and reliability studies. All of these distributions involve at least one unknown parameter; hence their means and variances also depend on the parameters. The reader is referred to textbooks in this area for details. For example, Hahn and Shapiro [1967, pp. 163–169 and pp. 120–134] comprehensively list these and other distributions on their p.d.f.s and the graphs, parameters, means, variances, with discussions and examples of their applications.

## The Normal Distribution

Perhaps *the* most important distribution in statistics and probability is the normal distribution (also known as the Gaussian distribution). This distribution involves two parameters,  $\mu$  and  $\sigma^2$ , and its p.d.f. is given by

$$f(x) = f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad (207.3)$$

for  $-\infty < \mu < \infty$ ,  $\sigma^2 > 0$ , and  $-\infty < x < \infty$ . It can be shown that, for a p.d.f. of this form, the values of  $\mu$  and  $\sigma^2$  are, respectively, that of the mean and the variance of the distribution. Further, the quantity  $\sigma = \sqrt{\sigma^2}$  is called the standard deviation of the distribution. We shall use the symbol  $X \sim \mathcal{N}(\mu, \sigma^2)$  to denote that  $X$  has a normal distribution with mean  $\mu$  and variance  $\sigma^2$ .

When plotting the p.d.f.  $f(x; \mu, \sigma^2)$  given in Eq. (207.3), we see that the resulting graph represents a bell-shaped curve and is symmetric about  $\mu$ . If a random variable  $Z$  has an  $\mathcal{N}(0, 1)$  distribution, then the p.d.f. of  $Z$  is given by [from Eq. (207.3)]

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad -\infty < z < \infty \quad (207.4)$$

The distribution function of  $Z$ ,

$$\Phi(z) = \int_{-\infty}^z \phi(u) du, \quad -\infty < z < \infty \quad (207.5)$$

cannot be given in a closed form; hence it has been tabulated. The table of  $\Phi(z)$  can be found in most textbooks in statistics and probability, including those listed in the references at the end of this chapter (we note in passing that, by the symmetry property,  $\Phi(z) + \Phi(-z) = 1$  holds for all  $z$ ).

## 207.2 Random Sample and Sampling Distributions

---

### Random Sample and Related Statistics

As noted in Box, Hunter, and Hunter [1978], in the design and analysis of engineering experiments, a study usually involves the following steps:

- (i) The choice of a suitable stochastic model by assuming that the observations follow a certain distribution. The functional form of the distribution (or the p.d.f.) is assumed to be known except the value(s) of the parameter(s).
- (ii) Design of experiments and collection of data.
- (iii) Summarization of data and computation of certain statistics.

(iv) Statistical inference (including the estimation of the parameters of the underlying distribution and the hypothesis-testing problems).

In order to make statistical inference concerning the parameter(s) of a distribution, it is essential to first study the sampling distributions. We say that  $X_1, X_2, \dots, X_n$  represent a random sample of size  $n$  if they are independent random variables and each of them has the same p.d.f.  $f(x)$ . (Due to space limitations, the notion of independence will not be carefully discussed here. But, nevertheless, we say that  $X_1, X_2, \dots, X_n$  are independent if

$$P[X_1 \in A_1, X_2 \in A_2, \dots, X_n \in A_n] = \prod_{i=1}^n P[X_i \in A_i] \quad (207.6)$$

holds for all sets  $A_1, A_2, \dots, A_n$ .) Because the parameters of the population are unknown, the population mean  $\mu$  and the population variance  $\sigma^2$  are unknown. In most commonly used distributions,  $\mu$  and  $\sigma^2$  can be estimated by the sample mean  $\bar{X}$  and the sample variance  $S^2$ , respectively, which are given by

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left[ \sum_{i=1}^n X_i^2 - n\bar{X}^2 \right] \quad (207.7)$$

(The second equality in the formula for  $S^2$  can be verified algebraically.) Now, because  $X_1, X_2, \dots, X_n$  are random variables,  $\bar{X}$  and  $S^2$  are also random variables. Each of them is called a statistic and has a probability distribution which also involves the unknown parameter(s). In probability theory, there are two fundamental results concerning their distributional properties.

**Theorem 1.**

(Weak Law of Large Numbers). As the sample size  $n$  becomes large,  $\bar{X}$  converges to  $\mu$  in probability and  $S^2$  converges to  $\sigma^2$  in probability. More precisely, for every fixed positive number  $\varepsilon > 0$ , we have

$$P[|\bar{X} - \mu| \leq \varepsilon] \rightarrow 1, \quad P[|S^2 - \sigma^2| \leq \varepsilon] \rightarrow 1 \quad (207.8)$$

as  $n \rightarrow \infty$ .

**Theorem 2.**

(Central Limit Theorem). As  $n$  becomes large, the distribution of the random variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \quad (207.9)$$

has approximately an  $\mathcal{N}(0, 1)$  distribution. More precisely,

$$P[Z \leq z] \rightarrow \Phi(z) \quad \text{for every fixed } z \text{ as } n \rightarrow \infty \quad (207.10)$$

## 207.3 Normal Distribution-Related Sampling Distributions

---

### One-Sample Case

Additional results exist when the observations come from a normal population. If  $X_1, X_2, \dots, X_n$  represent a random sample of size  $n$  from an  $\mathcal{N}(\mu, \sigma^2)$  population, then the following sampling distributions are useful.

**Fact 1.**

For every fixed  $n$ , the distribution of  $Z$  given in Eq.(207.9) has exactly an  $\mathcal{N}(0, 1)$  distribution.

**Fact 2.**

The distribution of the statistic  $T = \sqrt{n}(\bar{X} - \mu)/S$ , where  $S = \sqrt{S^2}$  is the sample standard deviation, is called a Student's  $t$  distribution with  $\nu = n - 1$  degrees of freedom; in symbols,  $t(n - 1)$ .

This distribution is useful for making inference on  $\mu$  when  $\sigma^2$  is unknown. A table of the percentiles can be found in most statistics textbooks.

**Fact 3.**

The distribution of the statistic  $W = (n - 1)S^2/\sigma^2$  is called a chi-squared distribution with  $\nu = n - 1$  degrees of freedom; in symbols,  $\chi^2(\nu)$ .

Such a distribution is useful in making inference on  $\sigma^2$ . A table of the percentiles can also be found in most statistics books.

### Two-Sample Case

In certain applications, we may be interested in the comparisons of two different treatments. Suppose that independent samples from treatments  $T_1$  and  $T_2$  are to be observed as shown in [Table 207.1](#). The difference of the population means  $(\mu_1 - \mu_2)$  and the ratio of the population variances can be estimated, respectively, by  $(\bar{X}_1 - \bar{X}_2)$  and  $S_1^2/S_2^2$ . The following facts summarize the distributions of these statistics.

**Table 207.1** Summary of Data for a Two-Sample Problem

Treatment	Observations	Distribution	Sample Size	Sample Mean	Sample Variance
$T_1$	$X_{11}, X_{12}, \dots, X_{1n_1}$	$\mathcal{N}(\mu_1, \sigma_1^2)$	$n_1$	$\bar{X}_1$	$S_1^2$
$T_2$	$X_{21}, X_{22}, \dots, X_{2n_2}$	$\mathcal{N}(\mu_2, \sigma_2^2)$	$n_2$	$\bar{X}_2$	$S_2^2$

**Fact 4.**

Under the assumption of normality,  $(\bar{X}_1 - \bar{X}_2)$  has an  $\mathcal{N}(\mu_1 - \mu_2, (\sigma_1^2/n_1) + (\sigma_2^2/n_2))$  distribution; or equivalently, for all  $n_1$  and  $n_2$ , the statistic

$$Z = [(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)]/(\sigma_1^2/n_1 + \sigma_2^2/n_2)^{1/2} \quad (207.11)$$

has an  $\mathcal{N}(0, 1)$  distribution.

**Fact 5.**

When  $\sigma_1^2 = \sigma_2^2 \equiv \sigma^2$ , the common population variance is estimated by

$$S_p^2 = (n_1 + n_2 - 2)^{-1} [(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2] \quad (207.12)$$

and  $(n_1 + n_2 - 2)S_p^2/\sigma^2$  has a  $\chi^2(n_1 + n_2 - 2)$  distribution.

**Fact 6.**

When  $\sigma_1^2 = \sigma_2^2$ , the statistic

$$T = [(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)]/S_p(1/n_1 + 1/n_2)^{1/2} \quad (207.13)$$

has a  $t(n_1 + n_2 - 2)$  distribution, where  $S_p = \sqrt{S_p^2}$ .

**Fact 7.**

The distribution of  $F = (S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2)$  is called an  $F$  distribution with degrees of freedom  $(n_1 - 1, n_2 - 1)$ ; in symbols,  $F(n_1 - 1, n_2 - 1)$ .

The percentiles of this distribution have also been tabulated and can be found in statistics books.

The distributions listed above (normal, Student's  $t$ , chi-squared, and  $F$ ) form an important part of classical statistical inference theory, and they are developed under the assumption that the observations follow a normal distribution. When the distribution of the population is not normal and inference on the population means is to be made, we conclude that (i) if the sample sizes  $n_1$  and  $n_2$  are large, then the statistic  $Z$  in Eq. (207.11) has an approximate  $\mathcal{N}(0, 1)$  distribution, (ii) in the small-sample case, the exact distribution of  $\bar{X}$  [of  $(\bar{X}_1 - \bar{X}_2)$ ] depends on the population p.d.f. There are several analytical methods for obtaining it, and those methods can be found in statistics textbooks.

## 207.4 Confidence Intervals

---

A method for estimating the population parameters based on the sample mean(s) and sample variance(s) involves the confidence intervals for the parameters.

## One-Sample Case

### Confidence Interval for $\mu$ When $\sigma^2$ Is Known

Consider the situation in which a random sample of size  $n$  is taken from an  $\mathcal{N}(\mu, \sigma^2)$  population and  $\sigma^2$  is known. An interval,  $I_1$ , of the form  $I_1 = (\bar{X} - d, \bar{X} + d)$  (with width  $2d$ ) is to be constructed as a *confidence interval* for  $\mu$ . If we make the assertion that  $\mu$  is in this interval (i.e.,  $\mu$  is bounded below by  $\bar{X} - d$ , and bounded above by  $\bar{X} + d$ ), then sometimes this assertion is correct and sometimes it is wrong, depending on the value of  $\bar{X}$  in a given experiment. If for a fixed  $\alpha$  value we would like to have a confidence probability (called confidence coefficient) such that

$$P[\mu \in I_1] = P[\bar{X} - d < \mu < \bar{X} + d] = 1 - \alpha \quad (207.14)$$

then we need to choose the value of  $d$  to satisfy  $d = z_{\alpha/2} \sigma / \sqrt{n}$ ; that is,

$$I_1 = \left( \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) \quad (207.15)$$

where  $z_{\alpha/2}$  is the  $(1 - \alpha/2)$  th percentile of the  $\mathcal{N}(0, 1)$  distribution such that  $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ . To see this, we note that, from the sampling distribution of  $\bar{X}$  (Fact 1), we have

$$\begin{aligned} P \left[ \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] &= P \left[ \frac{|\bar{X} - \mu|}{\sigma / \sqrt{n}} \leq z_{\alpha/2} \right] \\ &= \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) = 1 - \alpha. \end{aligned} \quad (207.16)$$

We further note that, even when the original population is not normal, by Theorem 2 the confidence probability is approximately  $(1 - \alpha)$  when the sample size is reasonably large.

### Confidence Interval for $\mu$ When $\sigma^2$ Is Unknown

Assume that the observations are from an  $\mathcal{N}(\mu, \sigma^2)$  population. When  $\sigma^2$  is unknown, by Fact 2 and a similar argument we see that

$$I_2 = \left( \bar{X} - t_{\alpha/2}(n-1) \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1) \frac{S}{\sqrt{n}} \right) \quad (207.17)$$

is a confidence interval for  $\mu$  with confidence probability  $1 - \alpha$ , where  $t_{\alpha/2}(n-1)$  is the  $(1 - \alpha/2)$  th percentile of the  $t(n-1)$  distribution.

### Confidence Interval for $\sigma^2$



If, under the same assumption of normality, a confidence interval for  $\sigma^2$  is needed when  $\mu$  is unknown, then

$$I_3 = ((n-1)S^2/\chi_{1-\alpha/2}^2(n-1), (n-1)S^2/\chi_{\alpha/2}^2(n-1)) \quad (207.18)$$

has a confidence probability  $1 - \alpha$ , when  $\chi_{1-\alpha/2}^2(n-1)$  and  $\chi_{\alpha/2}^2(n-1)$  are the  $(\alpha/2)$  th and  $(1 - \alpha/2)$  th percentiles, respectively, of the  $\chi^2(n-1)$  distribution.

## Two-Sample Case

### Confidence Intervals for $\mu_1 - \mu_2$ When $\sigma_1^2$ and $\sigma_2^2$ Are Known

Consider an experiment that involves the comparison of two treatments,  $T_1$  and  $T_2$ , as indicated in [Table 207.1](#). If a confidence interval for  $\delta = \mu_1 - \mu_2$  is needed when  $\sigma_1^2$  and  $\sigma_2^2$  are known then, by Fact 4 and a similar argument, the confidence interval

$$I_4 = \left( (\bar{X}_1 - \bar{X}_2) - z_{\alpha/2} \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}, (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2} \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} \right) \quad (207.19)$$

has a confidence probability  $1 - \alpha$ .

### Confidence Interval for $\mu_1 - \mu_2$ When $\sigma_1^2$ and $\sigma_2^2$ Are Unknown but Equal

Under the additional assumption that  $\sigma_1^2 = \sigma_2^2$  but the common variance is unknown, then by Fact 6 the confidence interval

$$I_5 = [(\bar{X}_1 - \bar{X}_2) - d, (\bar{X}_1 - \bar{X}_2) + d] \quad (207.20)$$

has a confidence probability  $1 - \alpha$ , where

$$d = t_{\alpha/2}(n_1 + n_2 - 2)S_p(1/n_1 + 1/n_2)^{1/2} \quad (207.21)$$

### Confidence Interval for $\sigma_1^2/\sigma_2^2$

A confidence interval for the ratio of the variances  $\sigma_2^2/\sigma_1^2$  can be obtained from the  $F$  distribution (see Fact 7), and the confidence interval

$$I_6 = \left( F_{1-\alpha/2}(n_1 - 1, n_2 - 1) \frac{S_2^2}{S_1^2}, F_{\alpha/2}(n_1 - 1, n_2 - 1) \frac{S_2^2}{S_1^2} \right) \quad (207.22)$$

has a confidence probability  $1 - \alpha$ , where  $F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$  and  $F_{\alpha/2}(n_1 - 1, n_2 - 1)$  are, respectively, the  $(\alpha/2)$  th and  $(1 - \alpha/2)$  th percentiles of the  $F(n_1 - 1, n_2 - 1)$  distribution.

## 207.5 Testing Statistical Hypotheses

---

A statistical hypothesis concerns a statement or assertion about the true value of the parameter in a given distribution. In the two-hypothesis problems, we deal with a null hypothesis and an alternative hypothesis, denoted by  $H_0$  and  $H_1$ , respectively. A decision is to be made, based on the data of the experiment, to either accept  $H_0$  (hence reject  $H_1$ ) or reject  $H_0$  (hence accept  $H_1$ ). In such a two-action problem, there are two types of errors we may commit: the type I error is to reject  $H_0$  when it is true, and the type II error is to accept  $H_0$  when it is false. As a standard practice, we do not reject  $H_0$  unless there is significant evidence indicating that it may be false (in doing so, the burden of proof that  $H_0$  is false is on the experimenter). Thus, we usually choose a small fixed number,  $\alpha$  (such as 0.05 or 0.01), such that the probability of committing a type I error is at most  $\alpha$ . With such a given  $\alpha$ , we can then determine the region in the data space for the rejection of  $H_0$  (called the critical region).

### One-Sample Case

Suppose that  $X_1, X_2, \dots, X_n$  represent a random sample of size  $n$  from an  $\mathcal{N}(\mu, \sigma^2)$  population, and  $\bar{X}$  and  $S^2$  are, respectively, the sample mean and sample variance.

#### Test for Mean

In testing

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu = \mu_1 (\mu_1 > \mu_0) \text{ or } H_1 : \mu > \mu_0$$

when  $\sigma^2$  is known, we reject  $H_0$  when  $\bar{X}$  is large. To determine the cut-off point, we note that (by Fact 1) the statistic  $Z_0 = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$  has an  $\mathcal{N}(0, 1)$  distribution under  $H_0$ . Thus if we decide to reject  $H_0$  when  $Z_0 > z_\alpha$ , then the probability of committing a type I error is  $\alpha$ . As a consequence, we apply the decision rule

$$d_1: \text{reject } H_0 \text{ if and only if } \bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Similarly, from the distribution of  $Z_0$  under  $H_0$ , we can obtain the critical region for the other types of hypotheses. When  $\sigma^2$  is unknown, then by Fact 2,  $T_0 = \sqrt{n}(\bar{X} - \mu_0)/S$  has a  $t(n-1)$  distribution under  $H_0$ . Thus, the corresponding tests can be obtained by substituting  $t_\alpha(n-1)$  for  $z_\alpha$  and  $S$  for  $\sigma$ . The tests for the various one-sided and two-sided hypotheses are summarized in [Table 207.2](#). For each set of hypotheses, the critical region given on the first line is for the case when  $\sigma^2$  is known, and that given on the second line is for the case when  $\sigma^2$  is unknown. Furthermore,  $t_\alpha$  and  $t_{\alpha/2}$  stand for  $t_\alpha(n-1)$  and  $t_{\alpha/2}(n-1)$ , respectively.

**Table 207.2** One-Sample Tests for Mean

Null Hypothesis $H_0$	Alternative Hypothesis $H_1$	Critical Region
$\mu = \mu_0$ or $\mu \leq \mu_0$	$\mu = \mu_1 > \mu_0$ or $\mu > \mu_0$	$\bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$ $\bar{X} > \mu_0 + t_\alpha \frac{S}{\sqrt{n}}$
$\mu = \mu_0$ or $\mu \geq \mu_0$	$\mu = \mu_1 < \mu_0$ or $\mu < \mu_0$	$\bar{X} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$ $\bar{X} < \mu_0 - t_\alpha \frac{S}{\sqrt{n}}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$ \bar{X} - \mu_0  > z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ $ \bar{X} - \mu_0  > t_{\alpha/2} \frac{S}{\sqrt{n}}$

**Test for Variance**

In testing hypotheses concerning the variance  $\sigma^2$  of a normal distribution, we use Fact 3 to assert that, under  $H_0 : \sigma^2 = \sigma_0^2$ , the distribution of  $w_0 = (n-1)S^2/\sigma_0^2$  is  $\chi^2(n-1)$ . The corresponding tests and critical regions are summarized in Table 207.3 ( $\chi_\alpha^2$  and  $\chi_{\alpha/2}^2$  stand for  $\chi_\alpha^2(n-1)$  and  $\chi_{\alpha/2}^2(n-1)$ , respectively).

**Table 207.3** One-Sample Tests for Variance

Null Hypothesis $H_0$	Alternative Hypothesis $H_1$	Critical Region
$\sigma^2 = \sigma_0^2$ or $\sigma^2 \leq \sigma_0^2$	$\sigma^2 = \sigma_1^2 > \sigma_0^2$ or $\sigma^2 > \sigma_0^2$	$(S^2/\sigma_0^2) > \frac{1}{n-1} \chi_\alpha^2$
$\sigma^2 = \sigma_0^2$ or $\sigma^2 \geq \sigma_0^2$	$\sigma^2 = \sigma_1^2 < \sigma_0^2$ or $\sigma^2 < \sigma_0^2$	$(S^2/\sigma_0^2) < \frac{1}{n-1} \chi_{1-\alpha}^2$
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$(S^2/\sigma_0^2) > \frac{1}{n-1} \chi_{\alpha/2}^2$ or $(S^2/\sigma_0^2) < \frac{1}{n-1} \chi_{1-\alpha/2}^2$

**Two-Sample Case**

In comparing the means and variances of two normal populations, we once again refer to Table 207.1 for notation and assumptions.

**Test for Difference of Two Means**

Let  $\delta = \mu_1 - \mu_2$  be the difference of the two population means. In testing  $H_0 : \delta = \delta_0$  versus a one-sided or two-sided alternative hypothesis, we note that, for

$$\tau = (\sigma_1^2/n_1 + \sigma_2^2/n_2)^{1/2} \quad (207.23)$$

and

$$v = S_p(1/n_1 + 1/n_2)^{1/2} \quad (207.24)$$

$Z_0 = [(\bar{X}_1 - \bar{X}_2) - \delta_0]/\tau$  has an  $\mathcal{N}(0, 1)$  distribution under  $H_0$ , and  $T_0 = [(\bar{X}_1 - \bar{X}_2) - \delta_0]/v$

has a  $t(n_1 + n_2 - 2)$  distribution under  $H_0$  when  $\sigma_1^2 = \sigma_2^2$ . Using these results, the corresponding critical regions for one-sided and two-sided tests can be obtained, and they are listed in [Table 207.4](#). Note that, as in the one-sample case, the critical region given on the first line for each set of hypotheses is for the case of known variances, and that given on the second line is for the case in which the variances are equal but unknown. Further,  $t_\alpha$  and  $t_{\alpha/2}$  stand for  $t_\alpha(n_1 + n_2 - 2)$  and  $t_{\alpha/2}(n_1 + n_2 - 2)$ , respectively.

**Table 207.4** Two-Sample Tests for Difference of Two Means

Null Hypothesis $H_0$	Alternative Hypothesis $H_1$	Critical Region
$\delta = \delta_0$ or $\delta \leq \delta_0$	$\delta = \delta_1 > \delta_0$ or $\delta > \delta_0$	$(\bar{X}_1 - \bar{X}_2) > \delta_0 + z_\alpha \tau$ $(\bar{X}_1 - \bar{X}_2) > \delta_0 + t_\alpha v$
$\delta = \delta_0$ or $\delta \geq \delta_0$	$\delta = \delta_1 < \delta_0$ or $\delta < \delta_0$	$(\bar{X}_1 - \bar{X}_2) < \delta_0 - z_\alpha \tau$ $(\bar{X}_1 - \bar{X}_2) < \delta_0 - t_\alpha v$
$\delta = \delta_0$	$\delta \neq \delta_0$	$ (\bar{X}_1 - \bar{X}_2) - \delta_0  > z_{\alpha/2} \tau$ $ (\bar{X}_1 - \bar{X}_2) - \delta_0  > t_{\alpha/2} v$

## 207.6 A Numerical Example

In the following, we provide a numerical example for illustrating the construction of confidence intervals and hypothesis-testing procedures. The example is given along the line of applications in Wadsworth [1990, p. 4.21] with artificial data.

Suppose that two processes,  $T_1$  and  $T_2$ , for manufacturing steel pins are in operation, and that a random sample of four pins (of five pins) was taken from the process  $T_1$  (the process  $T_2$ ) with the following results (in units of inches):

$$T_1: 0.7608, 0.7596, 0.7622, 0.7638$$

$$T_2: 0.7546, 0.7561, 0.7526, 0.7572, 0.7565$$

Simple calculation shows that the observed values of sample means, sample variances, and sample standard deviations are

$$\bar{X}_1 = 0.7616, \quad S_1^2 = 3.178914 \cdot 10^{-6}, \quad S_1 = 1.7830 \cdot 10^{-3}$$

$$\bar{X}_2 = 0.7554, \quad S_2^2 = 3.516674 \cdot 10^{-6}, \quad S_2 = 1.8753 \cdot 10^{-3}$$

## One-Sample Case

Let us first consider confidence intervals for the parameters of the first process,  $T_1$ , only.

1. Assume that, based on previous knowledge on processes of this type, the variance is known

to be  $\sigma_1^2 = 1.80^2 \cdot 10^{-6}$  ( $\sigma_1 = 0.0018$ ). Then, from the normal table [Ross, 1987, p. 482], we have  $z_{0.025} = 1.96$ . Thus, a 95% confidence interval for  $\mu_1$  is

$$(0.7616 - 1.96 \times 0.0018/\sqrt{4}, 0.7616 + 1.96 \times 0.0018/\sqrt{4})$$

or (0.7598, 0.7634) (after rounding off to the fourth decimal place).

2. If  $\sigma_1^2$  is unknown and a 95% confidence interval for  $\mu_1$  is needed, then, for  $t_{0.025}(3) = 3.182$  [Ross, 1987, p. 484], the confidence interval is

$$(0.7616 - 3.182 \times 0.001783/\sqrt{4}, 0.7616 + 3.182 \times 0.001783/\sqrt{4})$$

or (0.7588, 0.7644).

3. From the chi-squared table with  $4-1 = 3$  degrees of freedom, we have [Ross, 1987, p. 483]  $\chi_{0.975}^2 = 0.216$ ,  $\chi_{0.025}^2 = 9.348$ . Thus, a 95% confidence interval for  $\sigma_1^2$  is  $(3 \times 3.178914 \cdot 10^{-6}/9.348, 3.178914 \cdot 10^{-6}/0.216)$ , or  $(1.0202 \cdot 10^{-6}, 44.15158 \cdot 10^{-6})$ .
4. In testing the hypotheses

$$H_0 : \mu_1 = 0.76 \quad \text{vs.} \quad H_1 : \mu_1 > 0.76$$

with  $\alpha = 0.01$  when  $\sigma_1^2$  is unknown, the critical region is

$\bar{x}_1 > 0.76 + 4.541 \times 0.001783/\sqrt{4} = 0.7640$ . Because the observed value of  $\bar{x}_1$  is 0.7616,  $H_0$  is accepted. That is, we assert that there is no significant evidence to call for the rejection of  $H_0$ .

## Two-Sample Case

If we assume that the two populations have a common unknown variance, we can use the Student's  $t$  distribution (with degree of freedom  $\nu = 4 + 5 - 2 = 7$ ) to obtain confidence intervals and to test hypotheses for  $\mu_1 - \mu_2$ . We first note that the data given above yield

$$\begin{aligned} S_p^2 &= \frac{1}{7}(3 \times 3.178414 + 4 \times 3.516674) \cdot 10^{-6} \\ &= 3.371920 \cdot 10^{-6}, \\ S_p &= 1.836279 \cdot 10^{-3}, \quad v = S_p \sqrt{1/4 + 1/5} = 1.231813 \cdot 10^{-3} \end{aligned}$$

and  $\bar{X}_1 - \bar{X}_2 = 0.0062$ .

1. A 98% confidence interval for  $\mu_1 - \mu_2$  is  $(0.0062 - 2.998 \cdot v, 0.0062 + 2.998v)$  or (0.0025, 0.0099).
2. In testing the hypotheses  $H_0 : \mu_1 = \mu_2$  (i.e.,  $\mu_1 - \mu_2 = 0$ ) vs.  $H_1 : \mu_1 > \mu_2$  with  $\alpha = 0.05$ , the critical region is  $(X_1 - X_2) > 1.895v = 2.3344 \cdot 10^{-3}$ . Thus,  $H_0$  is rejected (i.e., we

- conclude that there is significant evidence to indicate that  $\mu_1 > \mu_2$  may be true).
3. In testing the hypotheses  $H_0 : \mu_1 = \mu_2$  vs  $H_1 : \mu_1 \neq \mu_2$  with  $\alpha = 0.02$ , the critical region is  $|X_1 - X_2| > 2.998v = 3.6930 \cdot 10^{-3}$ . Thus,  $H_0$  is rejected. We note this conclusion is consistent with the result that, with confidence probability  $1 - \alpha = 0.98$ , the confidence interval for  $(\mu_1 - \mu_2)$  does not contain the origin.

## References

- Bowker, A. H. and Lieberman, G. J. 1972. *Engineering Statistics*, 2d ed. Prentice Hall, Englewood Cliffs, NJ.
- Box, G. E. P., Hunter, W. G., and Hunter, J. S. 1978. *Statistics for Experimenters*. John Wiley & Sons, New York.
- Hahn, G. J. and Shapiro, S. S. 1967. *Statistical Models in Engineering*. John Wiley & Sons, New York.
- Hines, W. W. and Montgomery, D. G. 1980. *Probability and Statistics in Engineering and Management Science*. John Wiley & Sons, New York.
- Hogg, R. V. and Ledolter, J. 1992. *Engineering Statistics*. Macmillan, New York.
- Ross, S. M. 1987. *Introduction to Probability and Statistics for Engineers and Scientists*. John Wiley & Sons, New York.
- Wadsworth, H. M. (Ed.) 1990. *Handbook of Statistical Methods for Engineers and Scientists*. McGraw-Hill, New York.

## Further Information

Other important topics in engineering probability and statistics include sampling inspection and quality (process) control, reliability, regression analysis and prediction, design of engineering experiments, and analysis of variance. Due to space limitations, these topics are not treated in this chapter. The reader is referred to textbooks in this area for further information. There are many well-written books that cover most of these topics. The short list of references above consists of a small sample of them.

Cain, G. "Optimization"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# 208

## Optimization

---

208.1 Linear Programming

208.2 Unconstrained Nonlinear Programming

208.3 Constrained Nonlinear Programming

**George Cain**

*Georgia Institute of Technology*

---

### 208.1 Linear Programming

---

Let  $\mathbf{A}$  be an  $m \times n$  matrix,  $\mathbf{b}$  a column vector with  $m$  components, and  $\mathbf{c}$  a column vector with  $n$  components. Suppose  $m < n$ , and assume the rank of  $\mathbf{A}$  is  $m$ . The standard linear programming problem is to find, among all nonnegative solutions of  $\mathbf{Ax} = \mathbf{b}$ , one that minimizes

$$\mathbf{c}^T \mathbf{x} = c_1 x_1 + c_2 x_2 + \cdots + c_n x_n$$

This problem is called a *linear* program. Each solution of the system  $\mathbf{Ax} = \mathbf{b}$  is called a *feasible* solution, and the *feasible set* is the collection of all *feasible solutions*. The function  $\mathbf{c}^T \mathbf{x} = c_1 x_1 + c_2 x_2 + \cdots + c_n x_n$  is the *cost function*, or the *objective function*. A solution to the linear program is called an *optimal feasible solution*.

Let  $\mathbf{B}$  be an  $m \times m$  submatrix of  $\mathbf{A}$  made up of  $m$  linearly independent columns of  $\mathbf{A}$ , and let  $\mathbf{C}$  be the  $m \times (n - m)$  matrix made up of the remaining columns of  $\mathbf{A}$ . Let  $\mathbf{x}_B$  be the vector consisting of the components of  $\mathbf{x}$  corresponding to the columns of  $\mathbf{A}$  that make up  $\mathbf{B}$ , and let  $\mathbf{x}_C$  be the vector of the remaining components of  $\mathbf{x}$ , that is, the components of  $\mathbf{x}$  that correspond to the columns of  $\mathbf{C}$ . Then the equation  $\mathbf{Ax} = \mathbf{b}$  may be written  $\mathbf{Bx}_B + \mathbf{Cx}_C = \mathbf{b}$ . A solution of  $\mathbf{Bx}_B = \mathbf{b}$  together with  $\mathbf{x}_C = \mathbf{0}$  gives a solution  $\mathbf{x}$  of the system  $\mathbf{Ax} = \mathbf{b}$ . Such a solution is called a *basic solution*, and if it is, in addition, nonnegative, it is a *basic feasible solution*. If it is also optimal, it is an *optimal basic feasible solution*. The components of a basic solution are called *basic variables*.

The Fundamental Theorem of Linear Programming says that if there is a feasible solution, there is a basic feasible solution, and if there is an optimal feasible solution, there is an optimal basic feasible solution. The linear programming problem is thus reduced to searching among the set of basic solutions for an optimal solution. This set is, of course, finite, containing as many as  $n!/[m!(n - m)!]$  points. In practice, this will be a very large number, making it imperative that one use some efficient search procedure in seeking an optimal solution. The most important of



such procedures is the *simplex method*, details of which may be found in the references.

The problem of finding a solution of  $Ax \leq b$  that minimizes  $c^T x$  can be reduced to the standard problem by appending to the vector  $x$  an additional  $m$  nonnegative components, called *slack variables*. The vector  $x$  is replaced by  $z$ , where  $z^T = [x_1 x_2 \dots x_n \ s_1 s_2 \dots s_m]$ , and the matrix  $A$  is replaced by  $B = [A \ I]$ , where  $I$  is the  $m \times m$  identity matrix. The equation  $Ax = b$  is thus replaced by  $Bz = Ax + s = b$ , where  $s^T = [s_1 s_2 \dots s_m]$ . Similarly, if inequalities are reversed so that we have  $Ax \geq b$ , we simply append  $-s$  to the vector  $x$ . In this case, the additional variables are called *surplus variables*.

Associated with every linear programming problem is a corresponding dual problem. If the *primal* problem is to minimize  $c^T x$  subject to  $Ax \geq b$ , and  $x \geq 0$ , the corresponding *dual* problem is to maximize  $y^T b$  subject to  $y^T A \leq c^T$ . If either the primal problem or the dual problem has an optimal solution, so also does the other. Moreover, if  $x_p$  is an optimal solution for the primal problem and  $y_d$  is an optimal solution for the corresponding dual problem,  $c^T x_p = y_d^T b$ .

## 208.2 Unconstrained Nonlinear Programming

---

The problem of minimizing or maximizing a sufficiently smooth nonlinear function  $f(x)$  of  $n$  variables,  $x^T = [x_1 x_2 \dots x_n]$ , with no restrictions on  $x$  is essentially an ordinary problem in calculus. At a minimizer or maximizer  $x^*$ , it must be true that the gradient of  $f$  vanishes:

$$\nabla f(x^*) = 0$$

Thus  $x^*$  will be in the set of all solutions of this system of  $n$  generally nonlinear equations. The solution of the system can be, of course, a nontrivial undertaking. There are many recipes for solving systems of nonlinear equations. A method specifically designed for minimizing  $f$  is the *method of steepest descent*. It is an old and honorable algorithm, and the one on which most other more complicated algorithms for unconstrained optimization are based. The method is based on the fact that at any point  $x$ , the direction of maximum decrease of  $f$  is in the direction of  $-\nabla f(x)$ . The algorithm searches in this direction for a minimum, recomputes  $-\nabla f(x)$  at this point, and continues iteratively. Explicitly:

1. Choose an initial point  $x_0$ .
2. Assume  $x_k$  has been computed; then compute  $y_k = \nabla f(x_k)$ , and let  $t_k \geq 0$  be a local minimum of  $g(t) = f(x_k - ty_k)$ . Then  $x_{k+1} = x_k - t_k y_k$ .
3. Replace  $k$  by  $k + 1$ , and repeat step 2 until  $t_k$  is small enough.

Under reasonably general conditions, the sequence  $(x_k)$  converges to a minimum of  $f$ .

## 208.3 Constrained Nonlinear Programming

---

The problem of finding the maximum or minimum of a function  $f(x)$  of  $n$  variables subject to the constraints

$$\mathbf{a}(\mathbf{x}) = \begin{bmatrix} a_1(x_1, x_2, \dots, x_n) \\ a_2(x_1, x_2, \dots, x_n) \\ \vdots \\ a_m(x_1, x_2, \dots, x_n) \end{bmatrix} = \begin{bmatrix} b_1 \\ b \\ \vdots \\ b_m \end{bmatrix} = \mathbf{b}$$

is made into an unconstrained problem by introducing the new function  $L(\mathbf{x})$  :

$$L(\mathbf{x}) = f(\mathbf{x}) + \mathbf{z}^T \mathbf{a}(\mathbf{x})$$

where  $\mathbf{z}^T = [\lambda_1 \lambda_2 \dots \lambda_m]$  is the vector of *Lagrange multipliers*. Now the requirement that  $\nabla L(\mathbf{x}) = 0$  , together with the constraints  $\mathbf{a}(\mathbf{x}) = \mathbf{b}$  , give a system of  $n + m$  equations

$$\begin{aligned} \nabla f(\mathbf{x}) + \mathbf{z}^T \nabla \mathbf{a}(\mathbf{x}) &= 0 \\ \mathbf{a}(\mathbf{x}) &= \mathbf{b} \end{aligned}$$

for the  $n + m$  unknowns  $x_1, x_2, \dots, x_n, \lambda_1, \lambda_2, \dots, \lambda_m$  that must be satisfied by the minimizer (or maximizer)  $\mathbf{x}$  .

The problem of inequality constraints is significantly more complicated in the nonlinear case than in the linear case. Consider the problem of minimizing  $f(\mathbf{x})$  subject to  $m$  equality constraints  $\mathbf{a}(\mathbf{x}) = \mathbf{b}$  , and  $p$  inequality constraints  $\mathbf{c}(\mathbf{x}) \leq \mathbf{d}$  [thus  $\mathbf{a}(\mathbf{x})$  and  $\mathbf{b}$  are vectors of  $m$  components, and  $\mathbf{c}(\mathbf{x})$  and  $\mathbf{d}$  are vectors of  $p$  components]. A point  $\mathbf{x}^*$  that satisfies the constraints is a *regular point* if the collection

$$\{\nabla a_1(\mathbf{x}^*), \nabla a_2(\mathbf{x}^*), \dots, \nabla a_m(\mathbf{x}^*)\} \cup \{\nabla c_j(\mathbf{x}^*) : j \in J\}$$

where

$$J = \{j : c_j(\mathbf{x}^*) = d_j\}$$

is linearly independent. If  $\mathbf{x}^*$  is a local minimum for the constrained problem and if it is a regular point, there is a vector  $\mathbf{z}$  with  $m$  components and a vector  $\mathbf{w} \geq 0$  with  $p$  components such that

$$\begin{aligned} \nabla f(\mathbf{x}^*) + \mathbf{z}^T \nabla \mathbf{a}(\mathbf{x}^*) + \mathbf{w}^T \nabla \mathbf{c}(\mathbf{x}^*) &= 0 \\ \mathbf{w}^T (\mathbf{c}(\mathbf{x}^*) - \mathbf{d}) &= 0 \end{aligned}$$

These are the *Kuhn-Tucker conditions*. Note that in order to solve these equations, one needs to know for which  $j$  it is true that  $c_j(\mathbf{x}^*) = 0$  . (Such a constraint is said to be *active*.)

## References

- Luenberger, D. C. 1984. *Linear and Nonlinear Programming*, 2nd ed. Addison-Wesley, Reading, MA.
- Peressini, A. L., Sullivan, F. E., and Uhl, J. J., Jr. 1988. *The Mathematics of Nonlinear Programming*. Springer-Verlag, New York.

Ames, W. F. "Numerical Methods"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

**209.1 Linear Algebra Equations**

Direct Methods • Iterative Methods

**209.2 Nonlinear Equations in One Variable**

Special Methods for Polynomials • The Graeffe Root-Squaring Technique

**209.3 General Methods for Nonlinear Equations in One Variable**

Successive Substitutions

**209.4 Numerical Solution of Simultaneous Nonlinear Equations**

The Method of Successive Substitutions • The Newton-Raphson Procedure • Methods of Perturbation • The Method of Wegstein • The Method of Continuity

**209.5 Interpolation and Finite Differences**

Linear Interpolation • Divided Differences of Higher Order and Higher-Order Interpolation • Lagrange Interpolation Formulas • Other Difference Methods (Equally Spaced Ordinates) • Inverse Interpolation

**209.6 Numerical Differentiation**

The Use of Interpolation Formulas • Smoothing Techniques • Least Squares Methods

**209.7 Numerical Integration**

Newton Cotes Formulas (Equally Spaced Ordinates) • Gaussian Integration Formulas (Unequally Spaced Abscissas) • Two-Dimensional Formula

**209.8 Numerical Solution of Ordinary Differential Equations**

The Modified Euler Method • Modified Adam's Method • Runge-Kutta Methods • Equations of Higher Order and Simultaneous Differential Equations

**209.9 Numerical Solution of Integral Equations****209.10 Numerical Methods for Partial Differential Equations**

Finite Difference Methods • Weighted Residual Methods (WRMs) • Finite Elements • Method of Lines

**209.11 Discrete and Fast Fourier Transforms**

DFT Properties

**209.12 Software**

General Packages • Special Packages for Linear Systems • Ordinary Differential Equations Packages • Partial Differential Equations Packages

**William F. Ames***Georgia Institute of Technology***Introduction**

Since many mathematical models of physical phenomena are not solvable by available mathematical methods one must often resort to approximate or numerical methods. These

procedures do not yield exact results in the mathematical sense. This inexact nature of numerical results means we must pay attention to the errors. The two errors that concern us here are *round-off errors* and *truncation errors*.

Round-off errors arise as a consequence of using a number specified by  $m$  correct digits to approximate a number which requires more than  $m$  digits for its exact specification. For example, using 3.14159 to approximate the irrational number  $\pi$ . Such errors may be especially serious in matrix inversion or in any area where a very large number of numerical operations are required. Some attempts at handling these errors are called *enclosure methods* [Adams and Kulisch, 1993].

Truncation errors arise from the substitution of a finite number of steps for an infinite sequence of steps (usually an iteration) which would yield the exact result. For example, the iteration  $y_n(x) = 1 + \int_0^x xty_{n-1}(t)dt$ ,  $y(0) = 1$  is only carried out for a *few steps*, but it converges in *infinitely* many steps.

The study of some errors in a computation is related to the theory of probability. In what follows, a relation for the error will be given in certain instances.

## 209.1 Linear Algebra Equations

---

A problem often met is the determination of the solution vector  $u = (u_1, u_2, \dots, u_n)^T$  for the set of linear equations  $Au = v$  where  $A$  is the  $n \times n$  square matrix with coefficients  $a_{ij}$  ( $i, j = 1, \dots, n$ ) and  $v = (v_1, \dots, v_n)^T$ , and  $i$  denotes the row index and  $j$  the column index.

There are many numerical methods for finding the solution,  $u$ , of  $Au = v$ . The direct inversion of  $A$  is usually too expensive and is not often carried out unless it is needed elsewhere. We shall only list a few methods. One can check the literature for the many methods and computer software available. Some of the software is listed in the References section at the end of this chapter. The methods are usually subdivided into *direct* (once through) or *iterative* (repeated) procedures.

In what follows, it will often be convenient to partition the matrix  $A$  into the form  $A = U + D + L$ , where  $U$ ,  $D$ , and  $L$  are matrices having the same elements as  $A$ , respectively, above the main diagonal, on the main diagonal, and below the main diagonal, and zeros elsewhere. Thus,

$$U = \begin{bmatrix} 0 & a_{12} & & \cdots & a_{1n} \\ 0 & 0 & a_{23} & \cdots & a_{2n} \\ \vdots & \cdots & & \cdots & \\ 0 & 0 & \cdots & \cdots & 0 \end{bmatrix}$$

We also assume the  $u_j$ s are not all zero and  $\det A \neq 0$  so the solution is unique.

## Direct Methods

### Gauss Reduction

This classical method has spawned many variations. It consists of dividing the first equation by  $a_{11}$  (if  $a_{11} = 0$ , reorder the equations to find an  $a_{11} \neq 0$ ) and using the result to eliminate the terms in  $u_1$

from each of the succeeding equations. Next, the modified second equation is divided by  $a'_{22}$  (if  $a'_{22} = 0$ , a reordering of the modified equations may be necessary) and the resulting equation is used to eliminate all terms in  $u_2$  in the succeeding modified equations. This elimination is done  $n$  times resulting in a triangular system:

$$\begin{array}{ccccccc} u_1 + a'_{12}u_2 + \cdots & + a'_{1n}u_n & = & v'_1 \\ 0 & + u_2 & + \cdots & + a'_{2n}u_n & = & v'_2 \\ & & \cdots & & & \\ 0 & + \cdots & + u_{n-1} & + a'_{n-1,n}u_n & = & v'_{n-1} \\ & & & & u_n & = & v'_n \end{array}$$

where  $a'_{ij}$  and  $v'_j$  represent the specific numerical values obtained by this process. The solution is obtained by working backward from the last equation. Various modifications, such as the Gauss-Jordan reduction, the Gauss-Doolittle reduction, and the Crout reduction, are described in the classical reference authored by Bodewig [1956]. Direct methods prove very useful for sparse matrices and banded matrices that often arise in numerical calculation for differential equations. Many of these are available in computer packages such as IMSL, Maple, Matlab, and Mathematica.

### The Tridiagonal Algorithm

When the linear equations are tridiagonal, the system

$$\begin{array}{l} b_1 u_1 + c_1 u_2 = d_1 \\ a_i u_{i-1} + b_i u_i + c_i u_{i+1} = d_i \\ a_n u_{n-1} + b_n u_n = d_n, \quad i = 2, 3, \dots, n-1 \end{array}$$

can be solved explicitly for the unknowns, thereby eliminating any matrix operations.

The Gaussian elimination process transforms the system into a simpler one of *upper bidiagonal* form. We designate the coefficients of this new system by  $a'_i$ ,  $b'_i$ ,  $c'_i$ , and  $d'_i$ , and we note that

$$\begin{array}{l} a'_i = 0, \quad i = 2, 3, \dots, n \\ b'_i = 1, \quad i = 1, 2, \dots, n \end{array}$$

The coefficients  $c'_i$  and  $d'_i$  are calculated successively from the relations

$$c'_1 = \frac{c_1}{b_1} \quad d'_1 = \frac{d_1}{b_1}$$

$$c'_{i+1} = \frac{c_{i+1}}{b_{i+1} - a_{i+1} c'_i}$$

$$d'_{i+1} = \frac{d_{i+1} - a_{i+1} d'_i}{b_{i+1} - a_{i+1} c'_i}, \quad i = 1, 2, \dots, n-1$$

and, of course,  $c_n = 0$ .

Having completed the elimination, we examine the new system and see that the  $n$ th equation is now

$$u_n = d'_n$$

Substituting this value into the  $(n-1)$ st equation,

$$u_{n-1} + c'_{n-1} u_n = d'_{n-1}$$

we have

$$u_{n-1} = d'_{n-1} - c'_{n-1} u_n$$

Thus, starting with  $u_n$ , we have successively the solution for  $u_i$  as

$$u_i = d'_i - c'_i u_{i+1}, \quad i = n-1, n-2, \dots, 1$$

### Algorithm for Pentadiagonal Matrix

The equations to be solved are

$$a_i u_{i-2} + b_i u_{i-1} + c_i u_i + d_i u_{i+1} + e_i u_{i+2} = f_i$$

for  $1 \leq i \leq R$  with  $a_1 = b_1 = a_2 = e_{R-1} = d_R = e_R = 0$ .

The algorithm is as follows. First, compute

$$\delta_1 = d_1/c_1$$

$$\lambda_1 = e_1/c_1$$

$$\gamma_1 = f_1/c_1$$

and



$$\begin{aligned}
\mu_2 &= c_2 - b_2 \delta_1 \\
\delta_2 &= (d_2 - b_2 \lambda_1) / \mu_2 \\
\lambda_2 &= e_2 / \mu_2 \\
\gamma_2 &= (f - b_2 \gamma_1) / \mu_2
\end{aligned}$$

Then, for  $3 \leq i \leq R - 2$  , compute

$$\begin{aligned}
\beta_i &= b_i - a_i \delta_{i-2} \\
\mu_i &= c_i - \beta_i \delta_{i-1} - a_i \lambda_{i-2} \\
\delta_i &= (d_i - \beta_i \lambda_{i-1}) / \mu_i \\
\lambda_i &= e_i / \mu_i \\
\gamma_i &= (f_i - \beta_i \gamma_{i-1} - a_i \gamma_{i-2}) / \mu_i
\end{aligned}$$

Next, compute

$$\begin{aligned}
\beta_{R-1} &= b_{R-1} - a_{R-1} \delta_{R-3} \\
\mu_{R-1} &= c_{R-1} - \beta_{R-1} \delta_{R-2} - a_{R-1} \lambda_{R-3} \\
\delta_{R-1} &= (d_{R-1} - \beta_{R-1} \lambda_{R-2}) / \mu_{R-1} \\
\gamma_{R-1} &= (f_{R-1} - \beta_{R-1} \gamma_{R-2} - a_{R-1} \gamma_{R-3}) / \mu_{R-1}
\end{aligned}$$

and

$$\begin{aligned}
\beta_R &= b_R - a_R \delta_{R-2} \\
\mu_R &= c_R - \beta_R \delta_{R-1} - a_R \lambda_{R-2} \\
\gamma_R &= (f_R - \beta_R \gamma_{R-1} - a_R \gamma_{R-2}) / \mu_R
\end{aligned}$$

The  $\beta_i$  and  $\mu_i$  are used only to compute  $\delta_i$ ,  $\lambda_i$ , and  $\gamma_i$ , and need not be stored after they are computed. The  $\delta_i$ ,  $\lambda_i$ , and  $\gamma_i$  must be stored, as they are used in the back solution. This is

$$\begin{aligned}
u_R &= \gamma_R \\
u_{R-1} &= \gamma_{R-1} - \delta_{R-1} u_R
\end{aligned}$$

and

$$u_i = \gamma_i - \delta_i u_{i+1} - \lambda_i u_{i+2}$$

for  $R - 2 \geq i \geq 1$  .

### General Band Algorithm

The equations are of the form

$$A_j^{(M)} X_{j-M} + A_j^{(M-1)} X_{j-M+1} + \cdots + A_j^{(2)} X_{j-2} + A_j^{(1)} X_{j-1} + B_j X_j \\ + C_j^{(1)} X_{j+1} + C_j^{(2)} X_{j+2} + \cdots + C_j^{(M-1)} X_{j+M-1} + C_j^{(M)} X_{j+M} = D_j$$

for  $1 \leq j \leq N$  ,  $N \geq M$  . The algorithm used is as follows:

$$\alpha_j^{(k)} = A_j^{(k)} = 0, \quad \text{for } k \geq j \\ C_j^{(k)} = 0, \quad \text{for } k \geq N + 1 - j$$

The forward solution ( $j = 1, \dots, N$ ) is

$$\alpha_j^{(k)} = A_j^{(k)} - \sum_{p=k+1}^{p=M} \alpha_j^{(p)} W_{j-p}^{(p-k)}, \quad k = M, \dots, 1 \\ \beta_j = B_j - \sum_{p=1}^M \alpha_j^{(p)} W_{j-p}^{(p)} \\ W_j^{(k)} = \left( C_j^{(k)} - \sum_{p=k+1}^{p=M} \alpha_j^{(p-k)} W_{j-(p-k)}^{(p)} \right) / \beta_j, \quad k = 1, \dots, M \\ \gamma_j = \left( D_j - \sum_{p=1}^M \alpha_j^{(p)} \gamma_{j-p} \right) / \beta_j$$

The back solution ( $j = N, \dots, 1$ ) is

$$X_j = \gamma_j - \sum_{p=1}^M W_j^{(p)} X_{j+p}$$

### Cholesky Decomposition

When the matrix  $A$  is a symmetric and positive definite, as it is for many discretizations of self-adjoint positive definite boundary value problems, one can improve considerably on the band procedures by using the Cholesky decomposition. For the system  $Au = v$ , the matrix  $A$  can be

written in the form

$$A = (I + L)D(I + U)$$

where  $L$  is lower triangular,  $U$  is upper triangular, and  $D$  is diagonal. If  $A = A'$  ( $A'$  represents the transpose of  $A$ ), then

$$A = A' = (I + U)'D(I + L)'$$

Hence, because of the uniqueness of the decomposition,

$$I + L = (I + U)' = I + U'$$

and therefore,

$$A = (I + U)'D(I + U)$$

that is,

$$A = B'B, \text{ where } B = \sqrt{D}(I + U)$$

The system  $Au = v$  is then solved by solving the two triangular system

$$B'w = v$$

followed by

$$Bu = w$$

To carry out the decomposition  $A = B'B$ , all elements of the first row of  $A$ , and of the derived system, are divided by the square root of the (positive) leading coefficient. This yields smaller rounding errors than the banded methods because the relative error of  $\sqrt{a}$  is only half as large as that of  $a$  itself. Also, taking the square root brings numbers nearer to each other (i.e., the new coefficients do not differ as widely as the original ones do). The actual computation of  $B = (b_{ij}), j > i$ , is given in the following:

$$\begin{aligned}
b_{11} &= (a_{11})^{1/2}, & b_{1j} &= a_{1j}/b_{11}, & j &\geq 2 \\
b_{22} &= (a_{22} - b_{12}^2)^{1/2}, & b_{2j} &= (a_{2j} - b_{12}b_{1j})/b_{22} \\
b_{33} &= (a_{33} - b_{13}^2 - b_{23}^2)^{1/2}, & b_{3j} &= (a_{3j} - b_{13}b_{1j} - b_{23}b_{2j})/b_{33} \\
&\vdots \\
b_{ii} &= \left( a_{ii} - \sum_{k=1}^{i-1} b_{ki}^2 \right)^{1/2}, & b_{ij} &= \left( a_{ij} - \sum_{k=1}^{i-1} b_{ki}b_{kj} \right) / b_{ii}, & i &\geq 2, j \geq 2
\end{aligned}$$

## Iterative Methods

Iterative methods consist of repeated application of an often simple algorithm. They yield the exact answer only as the limit of a sequence. They can be programmed to take care of zeros in  $A$  and are self-correcting. Their structure permits the use of convergence accelerators, such as overrelaxation, Aitkins acceleration, or Chebyshev acceleration.

Let  $a_{ii} > 0$  for all  $i$  and  $\det A \neq 0$ . With  $A = U + D + L$  as previously described, several iteration methods are described for  $(U + D + L)u = v$ .

### Jacobi Method (Iteration by total steps)

Since  $u = -D^{-1}[U + L]u + D^{-1}v$ , the iteration  $u^{(k)}$  is  $u^{(k)} = -D^{-1}[U + L]u^{(k-1)} + D^{-1}v$ . This procedure has a slow convergent rate designated by  $R$ ,  $0 < R \ll 1$ .

### Gauss-Seidel Method (Iteration by single steps)

$u^{(k)} = -(L + D)^{-1} U u^{(k-1)} + (L + D)^{-1} v$ . Convergence rate is  $2R$ , twice as fast as that of the Jacobi method.

### Gauss-Seidel with Successive Overrelaxation (SOR)

Let  $\bar{u}_i^{(k)}$  be the  $i$ th components of the Gauss-Seidel iteration. The SOR technique is defined by

$$u_i^{(k)} = (1 - \omega)u_i^{(k-1)} + \omega\bar{u}_i^{(k)}$$

where  $1 < \omega < 2$  is the overrelaxation parameter. The full iteration is

$u^{(k)} = (D + \omega L)^{-1} \{[(1 - \omega)D - \omega U]u^{(k-1)} + \omega v\}$ . Optimal values of  $\omega$  can be computed and depend upon the properties of  $A$  [Ames, 1993]. With optimal values of  $\omega$ , the convergence rate of this method is  $2R\sqrt{2}$  which is much larger than that for Gauss-Seidel ( $R$  is usually much less than one).

For other acceleration techniques, see the literature [Ames, 1993].

## 209.2 Nonlinear Equations in One Variable

---

### Special Methods for Polynomials

The polynomial  $P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n = 0$ , with real coefficients  $a_j, j = 0, \dots, n$ , has exactly  $n$  roots which may be real or complex.

If all the coefficients of  $P(x)$  are integers, then any rational roots, say  $r/s$  ( $r$  and  $s$  are integers with no common factors), of  $P(x) = 0$  must be such that  $r$  is an integral divisor of  $a_n$  and  $s$  is an integral division of  $a_0$ . Any polynomial with rational coefficients may be converted into one with integral coefficients by multiplying the polynomial by the lowest common multiple of the denominators of the coefficients.

**Example.**  $x^4 - 5x^2/3 + x/5 + 3 = 0$ . The lowest common multiple of the denominators is 15. Multiplying by 15, which does not change the roots, gives  $15x^4 - 25x^2 + 3x + 45 = 0$ . The only possible rational roots  $r/s$  are such that  $r$  may have the value  $\pm 45, \pm 15, \pm 5, \pm 3$ , and  $\pm 1$ , while  $s$  may have the values  $\pm 15, \pm 5, \pm 3$ , and  $\pm 1$ . All possible rational roots, with no common factors, are formed using all possible quotients.

If  $a_0 > 0$ , the first negative coefficient is preceded by  $k$  coefficients which are positive or zero, and  $G$  is the largest of the absolute values of the negative coefficients, then each real root is less than  $1 + \sqrt[k]{G/a_0}$  (upper bound on the real roots). For a lower bound to the real roots, apply the criterion to  $P(-x) = 0$ .

**Example.**  $P(x) = x^5 + 3x^4 - 2x^3 - 12x + 2 = 0$ . Here  $a_0 = 1$ ,  $G = 12$ , and  $k = 2$ . Thus, the upper bound for the real roots is  $1 + \sqrt[2]{12} \approx 4.464$ . For the lower bound,  $P(-x) = -x^5 + 3x^4 + 2x^3 + 12x + 2 = 0$ , which is equivalent to  $x^5 - 3x^4 - 2x^3 - 12x - 2 = 0$ . Here  $k = 1$ ,  $G = 12$ , and  $a_0 = 1$ . A lower bound is  $-(1+12) = -13$ . Hence all real roots lie in  $-13 < x < 1 + \sqrt[2]{12}$ .

A useful *Descartes rule of signs* for the number of positive or negative real roots is available by observation for polynomials with real coefficients. The number of positive real roots is either equal to the number of sign changes,  $n$ , or is less than  $n$  by a positive *even* integer. The number of negative real roots is either equal to the number of sign changes,  $n$ , of  $P(-x)$ , or is less than  $n$  by a positive even integer.

**Example.**  $P(x) = x^5 - 3x^3 - 2x^2 + x - 1 = 0$ . There are three sign changes, so  $P(x)$  has either three or one positive roots. Since  $P(-x) = -x^5 + 3x^3 - 2x^2 - x - 1 = 0$ , there are either two or zero negative roots.

### The Graeffe Root-Squaring Technique

This is an iterative method for finding the roots of the algebraic equation

$$f(x) = a_0x^p + a_1x^{p-1} + \cdots + a_{p-1}x + a_p = 0$$

If the roots are  $r_1, r_2, r_3, \dots$ , then one can write

$$S_p = r_1^p \left( 1 + \frac{r_2^p}{r_1^p} + \frac{r_3^p}{r_1^p} + \cdots \right)$$

and if one root is larger than all the others, say  $r_1$ , then for large enough  $p$  all terms (other than 1) would become negligible. Thus,

$$S_p \approx r_1^p$$

or

$$\lim_{p \rightarrow \infty} S_p^{1/p} = r_1$$

The Graeffe procedure provides an efficient way for computing  $S_p$  via a sequence of equations such that the roots of each equation are the squares of the roots of the preceding equations in the sequence. This serves the purpose of ultimately obtaining an equation whose roots are so widely separated in magnitude that they may be read approximately from the equation by inspection. The basic procedure is illustrated for a polynomial of degree 4:

$$f(x) = a_0x^4 + a_1x^3 + a_2x^2 + a_3x + a_4 = 0$$

Rewrite this as

$$a_0x^4 + a_2x^2 + a_4 = -a_1x^3 - a_3x$$

and square both sides so that upon grouping

$$a_0^2x^8 + (2a_0a_2 - a_1^2)x^6 + (2a_0a_4 - 2a_1a_3 + a_2^2)x^4 + (2a_2a_4 - a_3^2)x^2 + a_4^2 = 0$$

Because this involves only even powers of  $x$ , we may set  $y = x^2$  and rewrite it as

$$a_0^2y^4 + (2a_0a_2 - a_1^2)y^3 + (2a_0a_4 - 2a_1a_3 + a_2^2)y^2 + (2a_2a_4 - a_3^2)y + a_4^2 = 0$$

whose roots are the squares of the original equation. If we repeat this process again, the new equation has roots which are the fourth power, and so on. After  $p$  such operations, the roots are  $2^p$  (original roots). If at any stage we write the coefficients of the unknown in sequence

$$a_0^{(p)} \quad a_1^{(p)} \quad a_2^{(p)} \quad a_3^{(p)} \quad a_4^{(p)}$$

then, to get the new sequence  $a^{(p+1)}$ , write  $a^{(p+1)}_i = 2a^{(p)}_0$  (times the symmetric coefficient) with respect to  $a^{(p)}_i - 2a^{(p)}_1$  (times the symmetric coefficient)  $-\cdots (-1)^i a^{(p)}_{i/2}$ . Now if the roots are  $r_1, r_2, r_3$ , and  $r_4$ , then  $a_1/a_0 = -\sum_{i=1}^4 r_i$ ,  $a_2/a_0 = -\sum_{i=1}^4 r_i^2, \dots, a_1^{(p)}/a_0^{(p)} = -\sum_{i=1}^4 r_i^{2^p}$ . If the roots

are all distinct and  $r_1$  is the largest in magnitude, then eventually

$$r_1^{2^p} \approx -\frac{a_1^{(p)}}{a_0^{(p)}}$$

And, if  $r_2$  is the next largest in magnitude, then

$$r_2^{2^p} \approx -\frac{a_2^{(p)}}{a_1^{(p)}}$$

And, in general,  $a_n^{(p)} / a_{n-1}^{(p)} \approx -r_n^{2^p}$ . This procedure is easily generalized to polynomials of arbitrary degree and specialized to the case of multiple and complex roots.

Other methods include Bernoulli iteration, Bairstow iteration, and Lin iteration. These may be found in the cited literature. In addition, the methods given below may be used for the numerical solution of polynomials.

## 209.3 General Methods for Nonlinear Equations in One Variable

---

### Successive Substitutions

Let  $f(x) = 0$  be the nonlinear equation to be solved. If this is rewritten as  $x = F(x)$ , then an iterative scheme can be set up in the form  $x_{k+1} = F(x_k)$ . To start the iteration, an initial guess must be obtained graphically or otherwise. The convergence or divergence of the procedure depends upon the method of writing  $x = F(x)$ , of which there will usually be several forms. A general rule to ensure convergence cannot be given. However, if  $a$  is a root of  $f(x) = 0$ , a necessary condition for convergence is that  $|F'(x)| < 1$  in that interval about  $a$  in which the iteration proceeds (this means the iteration cannot converge unless  $|F'(x)| < 1$ , but it does not ensure convergence). This process is called *first order* because the error in  $x_{k+1}$  is proportional to the first power of the error in  $x_k$ .

**Example.**  $f(x) = x^3 - x - 1 = 0$ . A rough plot shows a real root of approximately 1.3. The equation can be written in the form  $x = F(x)$  in several ways, such as  $x = x^3 - 1$ ,  $x = 1/(x^2 - 1)$ , and  $x = (1 + x)^{1/3}$ . In the first case,  $F'(x) = 3x^2 = 5.07$  at  $x = 1.3$ ; in the second,  $F(1.3) = 5.46$ ; only in the third case is  $F'(1.3) < 1$ . Hence, only the third iterative process has a chance to converge. This is illustrated in the iteration table below.

Step $k$	$x = \frac{1}{x^2 - 1}$	$x = x^3 - 1$	$x = (1 + x)^{1/3}$
0	1.3	1.3	1.3
1	1.4493	1.197	1.32

2	0.9087	0.7150	1.3238
3	-5.737	-0.6345	1.3247
4	...	...	1.3247

---

## 209.4 Numerical Solution of Simultaneous Nonlinear Equations

---

The techniques illustrated here will be demonstrated for two simultaneous equations— $f(x,y) = 0$  and  $g(x,y) = 0$ . They immediately generalize to more than two simultaneous equations.

### The Method of Successive Substitutions

The two simultaneous equations can be written in various ways in equivalent forms

$$x = F(x, y)$$

$$y = G(x, y)$$

and the method of successive substitutions can be based on

$$x_{k+1} = F(x_k, y_k)$$

$$y_{k+1} = G(x_k, y_k)$$

Again, the procedure is of the first order and a necessary condition for convergence is

$$\left| \frac{\partial F}{\partial x} \right| + \left| \frac{\partial F}{\partial y} \right| < 1 \quad \left| \frac{\partial G}{\partial x} \right| + \left| \frac{\partial G}{\partial y} \right| < 1$$

in the iteration neighborhood of the true solution.

### The Newton-Raphson Procedure

Using the two simultaneous equations, start from an approximation, say  $(x_0, y_0)$ , obtained graphically or from a two-way table. Then, solve successively the linear equations

$$\Delta x_k \frac{\partial f}{\partial x}(x_k, y_k) + \Delta y_k \frac{\partial f}{\partial y}(x_k, y_k) = -f(x_k, y_k)$$

$$\Delta x_k \frac{\partial g}{\partial x}(x_k, y_k) + \Delta y_k \frac{\partial g}{\partial y}(x_k, y_k) = -g(x_k, y_k)$$



for  $\Delta x_k$  and  $\Delta y_k$ . Then, the  $k + 1$  approximation is given from  $x_{k+1} = x_k + \Delta x_k$ ,  $y_{k+1} = y_k + \Delta y_k$ . A modification consists in solving the equations with  $(x_k, y_k)$  replaced by  $(x_0, y_0)$  (or another suitable pair later on in the iteration) in the derivatives. This means the derivatives (and therefore the coefficients of  $\Delta x_k, \Delta y_k$ ) are independent of  $k$ . Hence, the results become

$$\Delta x_k = \frac{-f(x_k, y_k)(\partial g/\partial y)(x_0, y_0) + g(x_k, y_k)(\partial f/\partial y)(x_0, y_0)}{(\partial f/\partial x)(x_0, y_0)(\partial g/\partial y)(x_0, y_0) - (\partial f/\partial y)(x_0, y_0)(\partial g/\partial x)(x_0, y_0)}$$

$$\Delta y_k = \frac{-g(x_k, y_k)(\partial f/\partial x)(x_0, y_0) + f(x_k, y_k)(\partial g/\partial x)(x_0, y_0)}{(\partial f/\partial x)(x_0, y_0)(\partial g/\partial y)(x_0, y_0) - (\partial f/\partial y)(x_0, y_0)(\partial g/\partial x)(x_0, y_0)}$$

and  $x_{k+1} = \Delta x_k + x_k$ ,  $y_{k+1} = \Delta y_k + y_k$ . Such an alteration of the basic technique reduces the rapidity of convergence.

### Example

$$f(x, y) = 4x^2 + 6x - 4xy + 2y^2 - 3$$

$$g(x, y) = 2x^2 - 4xy + y^2$$

By plotting, one of the approximate roots is found to be  $x_0 = 0.4$ ,  $y_0 = 0.3$ . At this point, there results  $\partial f/\partial x = 8$ ,  $\partial f/\partial y = -0.4$ ,  $\partial g/\partial x = 0.4$ , and  $\partial g/\partial y = -1$ . Hence,

$$x_{k+1} = x_k + \Delta x_k = x_k + \frac{-f(x_k, y_k) - 0.4g(x_k, y_k)}{8(-1) - (-0.4)(0.4)}$$

$$= x_k - 0.12755f(x_k, y_k) - 0.05102g(x_k, y_k)$$

and

$$y_{k+1} = y_k - 0.05102f(x_k, y_k) + 1.02041g(x_k, y_k)$$

The first few iteration steps are shown in the following table.

Step $k$	$x_k$	$y_k$	$f(x_k, y_k)$	$g(x_k, y_k)$
0	0.4	0.3	-0.26	0.07
1	0.43673	0.24184	0.078	0.0175
2	0.42672	0.25573	-0.0170	-0.007
3	0.42925	0.24943	0.0077	0.0010

## Methods of Perturbation

Let  $f(x) = 0$  be the equation. In general, the iterative relation is

$$x_{k+1} = x_k - \frac{f(x_k)}{\alpha_k}$$

where the iteration begins with  $x_0$  as an initial approximation and  $\alpha_k$  is some functional.

### The Newton-Raphson Procedure

This variant chooses  $\alpha_k = f'(x_k)$  where  $f' = df/dx$  and geometrically consists of replacing the graph of  $f(x)$  by the tangent line at  $x = x_k$  in each successive step. If  $f'(x)$  and  $f''(x)$  have the same sign throughout an interval  $a \leq x \leq b$  containing the solution, with  $f(a)$  and  $f(b)$  of opposite signs, then the process converges starting from any  $x_0$  in the interval  $a \leq x \leq b$ . The process is second order.

### Example

$$f(x) = x - 1 + \frac{(0.5)^x - 0.5}{0.3}$$

$$f'(x) = 1 - 2.3105[0.5]^x$$

An approximate root (obtained graphically) is 2.

Step $k$	$x_k$	$f(x_k)$	$f'(x_k)$
0	2	0.1667	0.4224
1	1.605	-0.002	0.2655
1	1.6125	-0.0005	...

### The Method of False Position

This variant is commenced by finding  $x_0$  and  $x_1$  such that  $f(x_0)$  and  $f(x_1)$  are of opposite signs. Then,  $\alpha_1 =$  slope of secant line joining  $[x_0, f(x_0)]$  and  $[x_1, f(x_1)]$  so that

$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} f(x_1)$$

In each following step,  $\alpha_k$  is the slope of the line joining  $[x_k, f(x_k)]$  to the most recently determined point where  $f(x_j)$  has the opposite sign from that of  $f(x_k)$ . This method is of first order.

### The Method of Wegstein

This is a variant of the method of successive substitutions which forces or accelerates convergence.

The iterative procedure  $x_{k+1} = F(x_k)$  is revised by setting  $\hat{x}_{k+1} = F(x_k)$  and then taking  $x_{k+1} = qx_k + (1 - q)\hat{x}_{k+1}$ . Wegstein found that suitably chosen  $qs$  are related to the basic process as follows:

Behavior of Successive Substitution Process	Range of Optimum $q$
Oscillatory convergence	$0 < q < 1/2$
Oscillatory divergence	$1/2 < q < 1$
Monotonic convergence	$q < 0$
Monotonic divergence	$1 < q$

At each step,  $q$  may be calculated to give a locally optimum value by setting

$$q = \frac{x_{k+1} - x_k}{x_{k+1} - 2x_k + x_{k-1}}$$

## The Method of Continuity

In the case of  $n$  equations in  $n$  unknowns, when  $n$  is large, determining the approximate solution may involve considerable effort. In such a case, the method of continuity is admirably suited for use on either digital or analog computers. It consists basically of the introduction of an extra variable into the  $n$  equations

$$f_i(x_1, x_2, \dots, x_n) = 0, \quad i = 1, \dots, n$$

and replacing them by

$$f_i(x_1, x_2, \dots, x_n, \lambda) = 0, \quad i = 1, \dots, n$$

where  $\lambda$  is introduced in such a way that the functions depend in a simple way upon  $\lambda$  and reduce to an easily solvable system for  $\lambda = 0$  and to the original equations for  $\lambda = 1$ . A system of ordinary differential equations, with independent variable  $\lambda$ , is then constructed by differentiating with respect to  $\lambda$ . There results

$$\sum_{j=1}^n \frac{\partial f_i}{\partial x_j} \frac{dx_j}{d\lambda} + \frac{\partial f_i}{\partial \lambda} = 0$$

where  $x_1, \dots, x_n$  are considered as functions of  $\lambda$ . The equations are integrated, with initial conditions obtained with  $\lambda = 0$ , from  $\lambda = 0$  to  $\lambda = 1$ . If the solution can be continued to  $\lambda = 1$ , the values of  $x_1, \dots, x_n$  for  $\lambda = 1$  will be a solution of the original equations. If the integration becomes infinite, the parameter  $\lambda$  must be introduced in a different fashion. Integration of the differential equations (which are usually nonlinear in  $\lambda$ ) may be accomplished on an analog computer or by digital means using techniques described in a later section entitled "Numerical

## Solution of Ordinary Differential Equations."

### Example

$$f(x, y) = 2 + x + y - x^2 + 8xy + y^3 = 0$$

$$g(x, y) = 1 + 2x + 3y + x^2 + xy - ye^x = 0$$

Introduce  $\lambda$  as

$$f(x, y, \lambda) = (2 + x + y) + \lambda(-x^2 + 8xy + y^3) = 0$$

$$g(x, y, \lambda) = (1 + 2x - 3y) + \lambda(x^2 + xy - ye^x) = 0$$

For  $\lambda = 1$ , these reduce to the original equations, but, for  $\lambda = 0$ , they are the linear systems

$$x + y = -2$$

$$2x - 3y = -1$$

which has the unique solution  $x = -1.4$ ,  $y = -0.6$ . The differential equations in this case become

$$\frac{\partial f}{\partial x} \frac{dx}{d\lambda} + \frac{\partial f}{\partial y} \frac{dy}{d\lambda} = -\frac{\partial f}{\partial \lambda}$$

$$\frac{\partial g}{\partial x} \frac{dx}{d\lambda} + \frac{\partial g}{\partial y} \frac{dy}{d\lambda} = -\frac{\partial g}{\partial \lambda}$$

or

$$\frac{dx}{d\lambda} = \frac{\frac{\partial f}{\partial y} \frac{\partial g}{\partial \lambda} - \frac{\partial f}{\partial \lambda} \frac{\partial g}{\partial y}}{\frac{\partial f}{\partial x} \frac{\partial g}{\partial y} - \frac{\partial f}{\partial y} \frac{\partial g}{\partial x}}$$

$$\frac{dy}{d\lambda} = \frac{\frac{\partial f}{\partial \lambda} \frac{\partial g}{\partial x} - \frac{\partial f}{\partial x} \frac{\partial g}{\partial \lambda}}{\frac{\partial f}{\partial x} \frac{\partial g}{\partial y} - \frac{\partial f}{\partial y} \frac{\partial g}{\partial x}}$$

Integrating in  $\lambda$ , with initial values  $x = -1.4$  and  $y = -0.6$  at  $\lambda = 0$ , from  $\lambda = 0$  to  $\lambda = 1$  gives the solution.

## 209.5 Interpolation and Finite Differences

---

The practicing engineer constantly finds it necessary to refer to tables as sources of information. Consequently, interpolation, or that procedure of "reading between the lines of the table," is a necessary topic in numerical analysis.

### Linear Interpolation

If a function  $f(x)$  is approximately linear in a certain range, then the ratio  $[f(x_1) - f(x_0)]/(x_1 - x_0) = f[x_0, x_1]$  is approximately independent of  $x_0$  and  $x_1$  in the range. The linear approximation to the function  $f(x)$ ,  $x_0 < x < x_1$ , then leads to the interpolation formula

$$\begin{aligned} f(x) &\approx f(x_0) + (x - x_0)f[x_0, x_1] \approx f(x_0) + \frac{x - x_0}{x_1 - x_0}[f(x_1) - f(x_0)] \\ &\approx \frac{1}{x_1 - x_0}[(x_1 - x)f(x_0) - (x_0 - x)f(x_1)] \end{aligned}$$

### Divided Differences of Higher Order and Higher-Order Interpolation

The first-order divided difference  $f[x_0, x_1]$  was defined above. Divided differences of second and higher order are defined iteratively by

$$\begin{aligned} f[x_0, x_1, x_2] &= \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} \\ &\vdots \\ f[x_0, x_1, \dots, x_k] &= \frac{f[x_1, \dots, x_k] - f[x_0, x_1, \dots, x_{k-1}]}{x_k - x_0} \end{aligned}$$

and a convenient form for computational purposes is

$$f[x_0, x_1, \dots, x_k] = \sum_{j=0}^k ' \frac{f(x_j)}{(x_j - x_0)(x_j - x_1) \cdots (x_j - x_k)}$$

for any  $k \geq 0$ , where the ' means the term  $(x_j - x_j)$  is omitted in the denominator. For example,

$$f[x_0, x_1, x_2] = \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)}$$

If the accuracy afforded by a linear approximation is inadequate, a generally more accurate result may be based upon the assumption that  $f(x)$  may be approximated by a polynomial of degree 2 or higher over certain ranges. This assumption leads to *Newton's fundamental interpolation formula* with divided differences:

$$f(x) \approx f(x_0) + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] \\ + (x - x_0)(x - x_1) \cdots (x - x_{n-1})f[x_0, x_1, \dots, x_n] + E_n(x)$$

where  $E_n(x)$  = error =  $[1/(n+1)] f^{(n+1)}(\xi) \pi(x)$  where

$\min(x_0, \dots, x_n) < \xi < \max(x_0, x_1, \dots, x_n, x)$  and  $\pi(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$ . In order to use this most effectively, one may first form a divided-difference table. For example, for third-order interpolation, the difference table is

$x_0$	$f(x_0)$			
		$f[x_0, x_1]$		
$x_1$	$f(x_1)$		$f[x_0, x_1, x_2]$	
		$f[x_1, x_2]$		$f[x_0, x_1, x_2, x_3]$
$x_2$	$f(x_2)$		$f[x_1, x_2, x_3]$	
		$f[x_2, x_3]$		
$x_3$	$f(x_3)$			

where each entry is given by taking the difference between diagonally adjacent entries to the left, divided by the abscissas corresponding to the ordinates intercepted by the diagonals passing through the calculated entry.

**Example.** Calculate by third-order interpolation the value of  $\cosh 0.83$  given  $\cosh 0.60$ ,  $\cosh 0.80$ ,  $\cosh 0.90$ , and  $\cosh 1.10$ .

$x_0 = 0.60$	1.185 47			
		0.7598		
$x_1 = 0.80$	1.337 43		0.6560	
		0.9566		0.1586
$x_2 = 0.90$	1.433 09		0.7353	
		1.1772		
$x_3 = 1.10$	1.668 52			

With  $n = 3$ , we have

$$\cosh 0.83 \approx 1.185 47 + (0.23)(0.7598) + (0.23)(0.03)(0.6560) \\ + (0.23)(0.03)(-0.07)(0.1586) = 1.364 64$$

which varies from the true value by 0.000 04.

## Lagrange Interpolation Formulas

The Newton formulas are expressed in terms of divided differences. It is often useful to have interpolation formulas expressed explicitly in terms of the ordinates involved. This is accomplished by the Lagrange interpolation polynomial of degree  $n$ :

$$y(x) = \sum_{j=0}^n \frac{\pi(x)}{(x - x_j)\pi'(x_j)} f(x_j)$$

where

$$\begin{aligned}\pi(x) &= (x - x_0)(x - x_1) \cdots (x - x_n) \\ \pi'(x_j) &= (x_j - x_0)(x_j - x_1) \cdots (x_j - x_n)\end{aligned}$$

where  $(x_j - x_j)$  is the omitted factor. Thus,

$$f(x) = y(x) + E_n(x)$$

$$E_n(x) = \frac{1}{(n+1)!} \pi(x) f^{(n+1)}(\varepsilon)$$

**Example.** The interpolation polynomial of degree 3 is

$$\begin{aligned}y(x) &= \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} f(x_0) + \frac{(x - x_0)(x - x_2)(x - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)} f(x_1) \\ &+ \frac{(x - x_0)(x - x_1)(x - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)} f(x_2) + \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)} f(x_3)\end{aligned}$$

Thus, directly from the data

$x$	0	1	3	4
$f(x)$	1	1	-1	2

we have as an interpolation polynomial  $y(x)$  for  $f(x)$ :

$$\begin{aligned}y(x) &= 1 \cdot \frac{(x - 1)(x - 3)(x - 4)}{(0 - 1)(0 - 3)(0 - 4)} + 1 \cdot \frac{x(x - 3)(x - 4)}{(1 - 0)(1 - 3)(1 - 4)} \\ &- 1 \cdot \frac{x(x - 1)(x - 4)}{(3 - 0)(3 - 1)(3 - 4)} + 2 \cdot \frac{(x - 0)(x - 1)(x - 3)}{(4 - 0)(4 - 1)(4 - 3)}\end{aligned}$$

## Other Difference Methods (Equally Spaced Ordinates)

### Backward Differences

The backward differences denoted by

$$\begin{aligned}\nabla f(x) &= f(x) - f(x - h) \\ \nabla^2 f(x) &= \nabla f(x) - \nabla f(x - h) \\ &\dots \\ \nabla^n f(x) &= \nabla^{n-1} f(x) - \nabla^{n-1} f(x - h)\end{aligned}$$

are useful for calculation near the end of tabulated data.

### Central Differences

The central difference denoted by

$$\begin{aligned}\delta f(x) &= f\left(x + \frac{h}{2}\right) - f\left(x - \frac{h}{2}\right) \\ \delta^n f(x) &= \delta^{n-1} f\left(x + \frac{h}{2}\right) - \delta^{n-1} f\left(x - \frac{h}{2}\right)\end{aligned}$$

is useful for calculating at the interior points of tabulated data.

Also to be found in the literature are Gaussian, Stirling, Bessel, Everett, Comrie differences, and so forth.

## Inverse Interpolation

This is the process of finding the value of the independent variable or abscissa corresponding to a given value of the function when the latter is between two tabulated values of the abscissa. One method of accomplishing this is to use Lagrange's interpolation formula in the form

$$x = \psi(y) = \sum_{j=0}^n \frac{\pi(y)}{(y - y_j)\pi'(y_j)} x_j$$

where  $x$  is expressed as a function of  $y$ . Other methods revolve about methods of iteration.

## 209.6 Numerical Differentiation

---

Numerical differentiation should be avoided wherever possible, particularly when data are empirical and subject to appreciable observation errors. Errors in data can affect numerical derivatives quite strongly (i.e., differentiation is a roughening process). When such a calculation



must be made, it is usually desirable first to *smooth* the data to a certain extent.

## The Use of Interpolation Formulas

If the data are given over equidistant values of the independent variable  $x$ , an interpolation formula, such as the Newton formula, may be used, and the resulting formula differentiated analytically. If the independent variable is not at equidistant values, then Lagrange's formulas must be used. By differentiating three- and five-point Lagrange interpolation formulas, the following differentiation formulas result for equally spaced tabular points.

### Three-point Formulas

Let  $x_0$ ,  $x_1$ , and  $x_2$  be the three points.

$$f'(x_0) = \frac{1}{2h} [-3f(x_0) + 4f(x_1) - f(x_2)] + \frac{h^2}{3} f'''(\varepsilon)$$

$$f'(x_1) = \frac{1}{2h} [-f(x_0) + f(x_2)] + \frac{h^2}{6} f'''(\varepsilon)$$

$$f'(x_2) = \frac{1}{2h} [f(x_0) - 4f(x_1) + 3f(x_2)] + \frac{h^2}{3} f'''(\varepsilon)$$

where the last term is an error term and  $\min_j x_j < \varepsilon < \max_j x_j$ .

### Five-point Formulas

Let  $x_0$ ,  $x_1$ ,  $x_2$ ,  $x_3$ , and  $x_4$  be the five values of the equally spaced independent variable and  $f_j = f(x_j)$ .

$$f'(x_0) = \frac{1}{12h} [-25f_0 + 48f_1 - 36f_2 + 16f_3 - 3f_4] + \frac{h^4}{5} f^{(v)}(\varepsilon)$$

$$f'(x_1) = \frac{1}{12h} [-3f_0 - 10f_1 + 18f_2 - 6f_3 + f_4] - \frac{h^4}{20} f^{(v)}(\varepsilon)$$

$$f'(x_2) = \frac{1}{12h} [f_0 - 8f_1 + 8f_3 - f_4] + \frac{h^4}{30} f^{(v)}(\varepsilon)$$

$$f'(x_3) = \frac{1}{12h} [-f_0 + 6f_1 - 18f_2 + 10f_3 + 3f_4] - \frac{h^4}{20} f^{(v)}(\varepsilon)$$

$$f'(x_4) = \frac{1}{12h} [3f_0 - 16f_1 + 36f_2 - 48f_3 + 25f_4] + \frac{h^4}{5} f^{(v)}(\varepsilon)$$

and the last term is again an error term.

## Smoothing Techniques

These techniques involve the approximation of the tabular data by a least squares fit of the data

using some known functional form, usually a polynomial. In place of approximating  $f(x)$  by a single least squares polynomial of degree  $n$  over the entire range of the tabulation, it is often desirable to replace each tabulated value by the value taken on by a least squares polynomial of degree  $n$  relevant to a subrange of  $2M + 1$  points centered, where possible, at the point for which the entry is to be modified. Thus, each smoothed value replaces a tabulated value. Let  $f_j = f(x_j)$  be the tabular points and  $y_j$  = smoothed values. A first-degree least squares with three points would be

$$\begin{aligned}y_0 &= \frac{1}{6}[5f_0 + 2f_1 - f_2] \\y_1 &= \frac{1}{3}[f_0 + f_1 + f_2] \\y_2 &= \frac{1}{6}[-f_0 + 2f_1 + 5f_2]\end{aligned}$$

A first-degree least squares with five points would be

$$\begin{aligned}y_0 &= \frac{1}{5}[3f_0 + 2f_1 + f_2 - f_4] \\y_1 &= \frac{1}{10}[4f_0 + 3f_1 + 2f_2 + f_3] \\y_2 &= \frac{1}{5}[f_0 + f_1 + f_2 + f_3 + f_4] \\y_3 &= \frac{1}{10}[f_0 + 2f_1 + 3f_2 + 4f_3] \\y_4 &= \frac{1}{5}[-f_0 + f_2 + 2f_3 + 3f_4]\end{aligned}$$

Thus, for example, if first-degree, five-point least squares are used, the central formula is used for all values except the first two and the last two, where the off-center formulas are used. A third-degree least squares with seven points would be

$$\begin{aligned}y_0 &= \frac{1}{42}[39f_0 + 8f_1 - 4f_2 - 4f_3 + f_4 + 4f_5 - 2f_6] \\y_1 &= \frac{1}{42}[8f_0 + 19f_1 + 16f_2 + 6f_3 - 4f_4 - 7f_5 + 4f_6] \\y_2 &= \frac{1}{42}[-4f_0 + 16f_1 + 19f_2 + 12f_3 + 2f_4 - 4f_5 + f_6] \\y_3 &= \frac{1}{21}[-2f_0 + 3f_1 + 6f_2 + 7f_3 + 6f_4 + 3f_5 - 2f_6] \\y_4 &= \frac{1}{42}[f_0 - 4f_1 + 2f_2 + 12f_3 + 19f_4 + 16f_5 - 4f_6] \\y_5 &= \frac{1}{42}[4f_0 - 7f_1 - 4f_2 + 6f_3 + 16f_4 + 19f_5 + 8f_6] \\y_6 &= \frac{1}{42}[-2f_0 + 4f_1 + f_2 - 4f_3 - 4f_4 + 8f_5 + 39f_6]\end{aligned}$$

Additional smoothing formulas may be found in the references. After the data are smoothed, any of the interpolation polynomials, or an appropriate least squares polynomial, may be fitted and the results used to obtain the derivative.

## Least Squares Methods

### Parabolic

For five evenly spaced neighboring abscissas labeled  $x_{-2}$ ,  $x_{-1}$ ,  $x_0$ ,  $x_1$ , and  $x_2$ , and their ordinates  $f_{-2}$ ,  $f_{-1}$ ,  $f_0$ ,  $f_1$ , and  $f_2$ , assume a parabola is fit by least squares. There results for all interior points, except the first and last two points of the data, the formula for the numerical derivative:

$$f'_0 = \frac{1}{10h} [-2f_{-2} - f_{-1} + f_1 + 2f_2]$$

For the first two data points designated by 0 and  $h$ :

$$f'(0) = \frac{1}{20h} [-21f(0) + 13f(h) + 17f(2h) - 9f(3h)]$$

$$f'(h) = \frac{1}{20h} [-11f(0) + 3f(h) + 7f(2h) + f(3h)]$$

and for the last two given by  $\alpha - h$  and  $\alpha$ :

$$f'(\alpha - h) = \frac{1}{20h} [-11f(\alpha) + 3f(\alpha - h) + 7f(\alpha - 2h) + f(\alpha - 3h)]$$

$$f'(\alpha) = \frac{1}{20h} [-21f(\alpha) + 13f(\alpha - h) + 17f(\alpha - 2h) - 9f(\alpha - 3h)]$$

### Quartic (Douglas-Avakian)

A fourth-degree polynomial  $y = a + bx + cx^2 + dx^3 + ex^4$  is fitted to seven adjacent equidistant points (spacing  $h$ ) after a translation of coordinates has been made so that  $x = 0$  corresponds to the central point of the seven. Thus, these may be called  $-3h$ ,  $-2h$ ,  $-h$ ,  $0$ ,  $h$ ,  $2h$ , and  $3h$ . Let  $k =$  coefficient of  $h$  for the seven points. That is, in  $-3h$ ,  $k = -3$ . Then, the coefficients for the polynomial are

$$\begin{aligned}
a &= \frac{524 \sum f(kh) - 245 \sum k^2 f(kh) + 21 \sum k^4 f(kh)}{924} \\
b &= \frac{397 \sum k f(kh)}{1512h} - \frac{7 \sum k^3 f(kh)}{216h} \\
c &= \frac{-840 \sum f(kh) + 679 \sum k^2 f(kh) - 67 \sum k^4 f(kh)}{3168h^2} \\
d &= \frac{-7 \sum k f(kh) + \sum k^3 f(kh)}{216h^3} \\
e &= \frac{72 \sum f(kh) - 67 \sum k^2 f(kh) + 7 \sum k^4 f(kh)}{3168h^4}
\end{aligned}$$

where all summations run from  $k = -3$  to  $k = +3$  and  $f(kh)$  = tabular value at  $kh$ . The slope of the polynomial at  $x = 0$  is  $dy/dx = b$ .

## 209.7 Numerical Integration

---

Numerical evaluation of the finite integral  $\int_a^b f(x) dx$  is carried out by a variety of methods. A few are given here.

### Newton-Cotes Formulas (Equally Spaced Ordinates)

#### Trapezoidal Rule

This formula consists of subdividing the interval  $a \leq x \leq b$  into  $n$  subintervals  $a$  to  $a + h$ ,  $a + h$  to  $a + 2h$ , ..., and replacing the graph of  $f(x)$  by the result of joining the ends of adjacent ordinates by line segments. If  $f_j = f(x_j) = f(a + jh)$ ,  $f_0 = f(a)$ , and  $f_n = f(b)$ , the integration formula is

$$\int_a^b f(x) dx = \frac{h}{2} [f_0 + 2f_1 + 2f_2 + \cdots + 2f_{n-1} + f_n] + E_n$$

where  $|E_n| = (nh^3/12)|f''(\varepsilon)| = [(b-a)^3/12n^2]|f''(\varepsilon)|$ ,  $a < \varepsilon < b$ . This procedure is not of high accuracy. However, if  $f''(x)$  is continuous in  $a < x < b$ , the error goes to zero as  $1/n^2$ ,  $n \rightarrow \infty$ .

#### Parabolic Rule (Simpson's Rule)

This procedure consists of subdividing the interval  $a < x < b$  into  $n/2$  subintervals, each of length  $2h$ , where  $n$  is an even integer. Using the notation as above the integration formula is

$$\begin{aligned}
\int_a^b f(x) dx &= \frac{h}{3} [f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 4f_{n-3} \\
&\quad + 2f_{n-2} + 4f_{n-1} + f_n] + E_n
\end{aligned}$$

where

$$|E_n| = \frac{nh^5}{180} |f^{(iv)}(\varepsilon)| = \frac{(b-a)^3}{180n^4} |f^{(iv)}(\varepsilon)| \quad a < \varepsilon < b$$

This method approximates  $f(x)$  by a parabola on each subinterval. This rule is generally more accurate than the trapezoidal rule. It is the most widely used integration formula.

### Weddle's Rule

This procedure consists of subdividing the integral  $a < x < b$  into  $n/6$  subintervals, each of length  $6h$ , where  $n$  is a multiple of 6. Using the notation from the trapezoidal rule, there results

$$\begin{aligned} \int_a^b f(x)dx &= \frac{3h}{10} [f_0 + 5f_1 + f_2 + 6f_3 + f_4 + 5f_5 + 2f_6 + 5f_7 + f_8 + \cdots \\ &\quad + 6f_{n-3} + f_{n-2} + 5f_{n-1} + f_n] + E_n \end{aligned}$$

Note that the coefficients of  $f_j$  follow the rule 1, 5, 1, 6, 1, 5, 2, 5, 1, 6, 1, 5, 2, 5, etc. . . This procedure consists of approximating  $f(x)$  by a polynomial of degree 6 on each subinterval. Here,

$$E_n = \frac{nh^7}{1400} [10f^{(6)}(\varepsilon_1) + 9h^2f^{(8)}(\varepsilon_2)]$$

## Gaussian Integration Formulas (Unequally Spaced Abscissas)

These formulas are capable of yielding comparable accuracy with fewer ordinates than the equally spaced formulas. The ordinates are obtained by optimizing the distribution of the abscissas rather than by arbitrary choice. For the details of these formulas, Hildebrand [1956] is an excellent reference.

## Two-Dimensional Formula

Formulas for two-way integration over a rectangle, circle, ellipse, and so forth, may be developed by a double application of one-dimensional integration formulas. The two-dimensional generalization of the parabolic rule is given here. Consider the iterated integral  $\int_a^b \int_c^d f(x, y) dx dy$ . Subdivide  $c < x < d$  into  $m$  (even) subintervals of length  $h = (d - c)/m$ , and  $a < y < b$  into  $n$  (even) subintervals of length  $k = (b - a)/n$ . This gives a subdivision of the rectangle  $a \leq y \leq b$  and  $c \leq x \leq d$  into subrectangles. Let  $x_j = c + jh$ ,  $y_j = a + jk$ , and  $f_{i,j} = f(x_i, y_j)$ . Then,

$$\begin{aligned} \int_a^b \int_c^d f(x, y) dx dy &= \frac{hk}{9} [(f_{0,0} + 4f_{1,0} + 2f_{2,0} + \cdots + f_{m,0}) \\ &\quad + 4(f_{0,1} + 4f_{1,1} + 2f_{2,1} + \cdots + f_{m,1}) \\ &\quad + 2(f_{0,2} + 4f_{1,2} + 2f_{2,2} + \cdots + f_{m,2}) + \cdots \\ &\quad + (f_{0,n} + 4f_{1,n} + 2f_{2,n} + \cdots + f_{m,n})] + E_{m,n} \end{aligned}$$

where

$$E_{m,n} = -\frac{hk}{90} \left[ mh^4 \frac{\partial^4 f(\varepsilon_1, \eta_1)}{\partial x^4} + nk^4 \frac{\partial^4 f(\varepsilon_2, \eta_2)}{\partial y^4} \right]$$

where  $\varepsilon_1$  and  $\varepsilon_2$  lie in  $c < x < d$ , and  $\eta_1$  and  $\eta_2$  lie in  $a < y < b$ .

## 209.8 Numerical Solution of Ordinary Differential Equations

---

A number of methods have been devised to solve ordinary differential equations numerically. The general references contain some information. A numerical solution of a differential equation means a table of values of the function  $y$  and its derivatives over only a limited part of the range of the independent variable. Every differential equation of order  $n$  can be rewritten as  $n$  first-order differential equations. Therefore, the methods given below will be for first-order equations, and the generalization to simultaneous systems will be developed later.

### The Modified Euler Method

This method is simple and yields modest accuracy. If extreme accuracy is desired, a more sophisticated method should be selected. Let the first-order differential equation be  $dy/dx = f(x, y)$  with the initial condition  $(x_0, y_0)$  (i.e.,  $y = y_0$  when  $x = x_0$ ). The procedure is as follows.

- Step 1. From the given initial conditions  $(x_0, y_0)$ , compute  $y'_0 = f(x_0, y_0)$  and  $y''_0 = [\partial f(x_0, y_0)/\partial x] + [\partial f(x_0, y_0)/\partial y]y'_0$ . Then, determine  $y_1 = y_0 + hy'_0 + (h^2/2)y''_0$ , where  $h$  = subdivision of the independent variable.
- Step 2. Determine  $y'_1 = f(x_1, y_1)$ , where  $x_1 = x_0 + h$ . These prepare us for the following.

#### Predictor Steps

- Step 3. For  $n \geq 1$ , calculate  $(y_{n+1})_1 = y_{n-1} + 2hy'_n$ .
- Step 4. Calculate  $(y'_{n+1})_1 = f[x_{n+1}, (y_{n+1})_1]$ .

#### Corrector Steps

- Step 5. Calculate  $(y_{n+1})_2 = y_n + (h/2)[(y'_{n+1})_1 + y'_n]$ , where  $y_n$  and  $y'_n$  without the

subscripts are the previous values obtained by this process (or by steps 1 and 2).

Step 6.  $(y'_{n+1})_2 = f[x_{n+1}, (y_{n+1})_2]$  .

Step 7. Repeat the corrector steps 5 and 6 if necessary until the desired accuracy is produced in  $y_{n+1}, y'_{n+1}$ .

**Example.** Consider the equation  $y' = 2y^2 + x$  with the initial conditions  $y_0 = 1$  when  $x_0 = 0$ . Let  $h = 0.1$ . A few steps of the computation are illustrated.

Step	
1	$y'_0 = 2y_0^2 + x_0 = 2$ $y''_0 = 1 + 4y_0y'_0 = 1 + 8 = 9$ $y_1 = 1 + (0.1)(2) + [(0.1)^2/2]9 = 1.245$
2	$y'_1 = 2y_1^2 + x_1 = 3.100 + 0.1 = 3.200$
3	$(y_2)_1 = y_0 + 2hy'_1 = 1 + 2(0.1)3.200 = 1.640$
4	$(y'_2)_1 = 2(y_2)_1^2 + x_2 + 5.592$
5	$(y_2)_2 = y_1 + (0.1/2)[(y'_2)_1 + y'_1] = 1.685$
6	$(y'_2)_2 = 2(y_2)_2^2 + x_2 = 5.878$
5 (repeat)	$(y_2)_3 = y_1 + (0.05)[(y'_2)_2 + y'_1] = 1.699$
6 (repeat)	$(y'_2)_3 = 2(y_2)_3^2 + x_2 = 5.974$

and so forth. This procedure may be programmed for a computer. A discussion of the truncation error of this process may be found in Milne [1953].

## Modified Adam's Method

The procedure given here was developed retaining third differences. It can then be considered as a more exact predictor-corrector method than the Euler method. The procedure is as follows for  $dy/dx = f(x,y)$  and  $h$  = interval size.

Steps 1 and 2 are the same as in the Euler method.

### Predictor Steps

Step 3.  $(y_{n+1})_1 = y_n + (h/24)[55y'_n - 59y'_{n-1} + 37y'_{n-2} - 9y'_{n-3}]$  , where  $y'_n, y'_{n-1}$  , etc, are calculated in step 1.

Step 4.  $(y'_{n+1})_1 = f[x_{n+1}, (y_{n+1})_1]$  .

### Corrector Steps

Step 5.  $(y_{n+1})_2 = y_n + (h/24)[9(y'_{n+1})_1 + 19y'_n - 5y'_{n-1} + y'_{n-2}]$  .

Step 6.  $(y'_{n+1})_2 = f[x_{n+1}, (y_{n+1})_2]$  .

Step 7. Iterate steps 5 and 6 if necessary.

## Runge-Kutta Methods

These methods are self-starting and are inherently stable. Kopal [1955] is a good reference for their derivation and discussion. Third- and fourth-order procedures are given below for  $dy/dx = f(x,y)$ , where  $h$  = interval size.

For third-order (error  $\approx h^4$ ),

$$\begin{aligned}k_0 &= hf(x_n, y_n) \\k_1 &= hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_0) \\k_2 &= hf(x_n + h, y_n + 2k_1 - k_0)\end{aligned}$$

and

$$y_{n+1} = y_n + \frac{1}{6}(k_0 + 4k_1 + k_2)$$

for all  $n \geq 0$ , with initial condition  $(x_0, y_0)$ .

For fourth-order (error  $\approx h^5$ ),

$$\begin{aligned}k_0 &= hf(x_n, y_n) \\k_1 &= hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_0) \\k_2 &= hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1) \\k_3 &= hf(x_n + h, y_n + k_2)\end{aligned}$$

and

$$y_{n+1} = y_n + \frac{1}{6}(k_0 + 2k_1 + 2k_2 + k_3)$$

**Example.** (Third-order) Let  $dy/dx = x - 2y$ , with initial condition  $y_0 = 1$  when  $x_0 = 0$ , and let  $h = 0.1$ . Clearly,  $x_n = nh$ . To calculate  $y_1$ , proceed as follows:

$$\begin{aligned}k_0 &= 0.1[x_0 - 2y_0] = -0.2 \\k_1 &= 0.1[0.05 - 2(1 - 0.1)] = -0.175 \\k_2 &= 0.1[0.1 - 2(1 - 0.35 + 0.2)] = -0.16 \\y_1 &= 1 + \frac{1}{6}(-0.2 - 0.7 - 0.16) = 0.8234\end{aligned}$$

## Equations of Higher Order and Simultaneous Differential Equations

Any differential equation of second- or higher order can be reduced to a simultaneous system of first-order equations by the introduction of auxiliary variables. Consider the following equations:



$$\frac{d^2 x}{dt^2} + xy \frac{dx}{dt} + z = e^x$$

$$\frac{d^2 y}{dt^2} + xy \frac{dy}{dt} = 7 + t^2$$

$$\frac{d^2 z}{dt^2} + xz \frac{dz}{dt} + x = e^x$$

In the new variables  $x_1 = x$ ,  $x_2 = y$ ,  $x_3 = z$ ,  $x_4 = dx_1/dt$ ,  $x_5 = dx_2/dt$ , and  $x_6 = dx_3/dt$ , the equations become

$$\frac{dx_1}{dt} = x_4$$

$$\frac{dx_2}{dt} = x_5$$

$$\frac{dx_3}{dt} = x_6$$

$$\frac{dx_4}{dt} = -x_1 x_2 x_4 - x_3 + e^{x_1}$$

$$\frac{dx_5}{dt} = -x_3 x_2 x_5 + 7 + t^2$$

$$\frac{dx_6}{dt} = -x_1 x_3 x_6 - x_1 + e^{x_1}$$

which is a system of the general form

$$\frac{dx_i}{dt} = f_i(t, x_1, x_2, x_3, \dots, x_n)$$

where  $i = 1, 2, \dots, n$ . Such systems may be solved by simultaneous application of any of the above numerical techniques. A Runge-Kutta method for

$$\frac{dx}{dt} = f(t, x, y)$$

$$\frac{dy}{dt} = g(t, x, y)$$

is given below. The fourth-order procedure is shown.

Starting at the initial conditions  $x_0$ ,  $y_0$ , and  $t_0$ , the next values  $x_1$  and  $y_1$  are computed via the equations below (where  $\Delta t = h$ ,  $t_j = h + t_{j-1}$ ):

$$\begin{aligned} k_0 &= hf(t_0, x_0, y_0) & l_0 &= hg(t_0, x_0, y_0) \\ k_1 &= hf\left(t_0 + \frac{h}{2}, x_0 + \frac{k_0}{2}, y_0 + \frac{l_0}{2}\right) & l_1 &= hg\left(t_0 + \frac{h}{2}, x_0 + \frac{k_0}{2}, y_0 + \frac{l_0}{2}\right) \\ k_2 &= hf\left(t_0 + \frac{h}{2}, x_0 + \frac{k_1}{2}, y_0 + \frac{l_1}{2}\right) & l_2 &= hg\left(t_0 + \frac{h}{2}, x_0 + \frac{k_1}{2}, y_0 + \frac{l_1}{2}\right) \\ k_3 &= hf(t_0 + h, x_0 + k_2, y_0 + l_2) & l_3 &= hg(t_0 + h, x_0 + k_2, y_0 + l_2) \end{aligned}$$

and

$$\begin{aligned} x_1 &= x_0 + \frac{1}{6}(k_0 + 2k_1 + 2k_2 + k_3) \\ y_1 &= y_0 + \frac{1}{6}(l_0 + 2l_1 + 2l_2 + l_3) \end{aligned}$$

To continue the computation, replace  $t_0$ ,  $x_0$ , and  $y_0$  in the above formulas by  $t_1 = t_0 + h$ ,  $x_1$ , and  $y_1$  just calculated. Extension of this method to more than two equations follows precisely this same pattern.

## 209.9 Numerical Solution of Integral Equations

---

This section considers a method of numerically solving the Fredholm integral equation of the second kind:

$$u(x) = f(x) + \lambda \int_a^b k(x, t)u(t)dt \quad \text{for } u(x)$$

The method discussed arises because a definite integral can be closely approximated by any of several numerical integration formulas (each of which arises by approximating the function by some polynomial over an interval). Thus, the definite integral can be replaced by an integration formula which becomes

$$u(x) = f(x) + \lambda(b-a) \left[ \sum_{i=1}^n c_i k(x, t_i)u(t_i) \right]$$

where  $t_1, \dots, t_n$  are points of subdivision of the  $t$  axis,  $a \leq t \leq b$ , and the  $c$ s are coefficients whose values depend upon the type of numerical integration formula used. Now, this must hold for all values of  $x$ , where  $a \leq x \leq b$ ; so it must hold for  $x = t_1, x = t_2, \dots, x = t_n$ . Substituting for  $x$

successively  $t_1, t_2, \dots, t_n$ , and setting  $u(t_i) = u_i$  and  $f(t_i) = f_i$ , we get  $n$  linear algebraic equations for the  $n$  unknowns  $u_1, \dots, u_n$ . That is,

$$u_i = f_i + (b - a)[c_1 k(t_i, t_1)u_1 + c_2 k(t_i, t_2)u_2 + \dots + c_n k(t_i, t_n)u_n], \quad i = 1, 2, \dots, n$$

These  $u_j$  may be solved for by the methods under the section entitled "Numerical Solution of Linear Equations."

## 209.10 Numerical Methods for Partial Differential Equations

---

The ultimate goal of numerical (discrete) methods for partial differential equations (PDEs) is the reduction of continuous systems (projections) to discrete systems that are suitable for high-speed computer solutions. The user must be cautioned that the seeming elementary nature of the techniques holds pitfalls that can be seriously misleading. These approximations often lead to difficult mathematical questions of adequacy, accuracy, convergence, stability, and consistency. Convergence is concerned with the approach of the approximate numerical solution to the exact solution as the number of mesh units increase indefinitely in some sense. Unless the numerical method can be shown to converge to the exact solution, the chosen method is unsatisfactory.

Stability deals in general with error growth in the calculation. As stated before, any numerical method involves truncation and round-off errors. These errors are not serious unless they grow as the computation proceeds (i.e., the method is unstable).

### Finite Difference Methods

In these methods, the derivatives are replaced by various finite differences. The methods will be illustrated for problems in two space dimensions  $(x, y)$  or  $(x, t)$ , where  $t$  is timelike. Using subdivisions  $\Delta x = h$  and  $\Delta y = k$  with  $u(ih, jk) = u_{i,j}$ , approximate

$$\begin{aligned} u_x|_{i,j} &= [(u_{i+1,j} - u_{i,j})/h] + O(h) \quad (\text{forward difference}), \text{ a first-order } [O(h)] \text{ method, or} \\ u_x|_{i,j} &= [(u_{i+1,j} - u_{i-1,j})/2h] + O(h^2) \quad (\text{central difference}), \text{ a second-order method. The second} \\ &\text{derivative is usually approximated with the second-order method} \\ u_{xx}|_{i,j} &= [(u_{i+1,j} - 2u_{i,j} + u_{i-1,j})/h^2] + O(h^2) \end{aligned}$$

**Example.** Using second-order differences for  $u_{xx}$  and  $u_{yy}$ , the five-point difference equation (with  $h = k$ ) for Laplace's equation  $u_{xx} + u_{yy} = 0$  is  $u_{i,j} = \frac{1}{4}[u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}]$ . The accuracy is  $O(h^2)$ . This model is called *implicit* because one must solve for the total number of unknowns at the unknown grid points  $(i, j)$  in terms of the given boundary data. In this case, the system of equations is a linear system.

**Example.** Using a forward-difference approximation for  $u_t$  and a second-order approximation for  $u_{xx}$ , the diffusion equation  $u_t = u_{xx}$  is approximated by the *explicit* formula  $u_{i,j+1} = ru_{i-1,j} + (1-2r)u_{i,j} + ru_{i+1,j}$ . This classic result permits step-by-step advancement in the  $t$  direction

beginning with the initial data at  $t = 0$  ( $j = 0$ ) and guided by the boundary data. Here, the term  $r = \Delta t / (\Delta x)^2 = k/h^2$  is restricted to be less than or equal to  $1/2$  for stability and the truncation error is  $O(k^2 + kh^2)$ .

The Crank-Nicolson implicit formula which approximates the diffusion equation  $u_t = u_{xx}$  is

$$\begin{aligned} -r\lambda u_{i-1,j+1} + (1 + 2r\lambda)u_{i,j+1} - r\lambda u_{i+1,j+1} &= r(1 - \lambda)u_{i-1,j} \\ &+ [1 - 2r(1 - \lambda)]u_{i,j} \\ &+ r(1 - \lambda)u_{i+1,j} \end{aligned}$$

The stability of this numerical method was analyzed by Crandall [Ames, 1993] where the  $\lambda$ ,  $r$  stability diagram is given.

Approximation of the time derivative in  $u_t = u_{xx}$  by a central difference leads to an always unstable approximation—the useless approximation

$$u_{i,j+1} = u_{i,j-1} + 2r(u_{i+1,j} - 2u_{i,j} + u_{i-1,j})$$

which is a warning to be careful.

The foregoing method is *symmetric* with respect to the point  $(i,j)$ , where the method is centered. Asymmetric methods have some computational advantages, so the Saul'yev method is described [Ames, 1993]. The algorithms ( $r = k/h^2$ )

$$(1 + r)u_{i,j+1} = u_{i,j} + r(u_{i-1,j+1} - u_{i,j} + u_{i+1,j}) \quad (\text{Saul'yev A})$$

$$(1 + r)u_{i,j+1} = u_{i,j} + r(u_{i+1,j+1} - u_{i,j} + u_{i-1,j}) \quad (\text{Saul'yev B})$$

are used as in any one of the following options:

1. Use Saul'yev A only and proceed line-by-line in the  $t(j)$  direction, but *always* from the left boundary on a line.
2. Use Saul'yev B only and proceed line-by-line in the  $t(j)$  direction, but *always* from the right boundary to the left on a line.
3. Alternate from line to line by first using Saul'yev A and then B, or the reverse. This is related to *alternating direction methods*.
4. Use Saul'yev A and Saul'yev B on the same line and average the results for the final answer (A first, and then B). This is equivalent to introducing the dummy variables  $P_{i,j}$  and  $Q_{i,j}$

such that

$$\begin{aligned} (1 + r)P_{i,j+1} &= U_{i,j} + r(P_{i-1,j+1} - U_{i,j} + U_{i+1,j}) \\ (1 + r)Q_{i,j+1} &= U_{i,j} + r(Q_{i+1,j+1} - U_{i,j} + U_{i-1,j}) \end{aligned}$$

and

$$U_{i,j+1} = \frac{1}{2}(P_{i,j+1} + Q_{i,j+1})$$

This averaging method has some computational advantage because of the possibility of truncation error cancellation. As an alternative, one can retain the  $P_{i,j}$  and  $Q_{i,j}$  from the previous step and replace  $U_{i,j}$  and  $U_{i+1,j}$  by  $P_{i,j}$  and  $P_{i+1,j}$ , respectively, and  $U_{i,j}$  and  $U_{i-1,j}$  by  $Q_{i,j}$  and  $Q_{i-1,j}$ , respectively.

## Weighted Residual Methods (WRMs)

To set the stage for the method of finite elements, we briefly describe the WRMs, which have several variations—the interior, boundary, and mixed methods. Suppose the equation is  $Lu = f$ , where  $L$  is the partial differential operator and  $f$  is a known function, of say  $x$  and  $y$ . The first step in WRM is to select a class of known basis functions  $b_i$  (e.g., trigonometric, Bessel, Legendre) to approximate  $u(x,y)$  as  $\sim \sum a_i b_i(x,y) = U(x,y,a)$ . Often, the  $b_i$  are selected so that  $U(x,y,a)$  satisfy the boundary conditions. This is essentially the *interior method*. If the  $b_i$  in  $U(x,y,a)$  are selected to satisfy the differential equations, but not the boundary conditions, the variant is called the *boundary method*. When neither the equation nor the boundary conditions are satisfied, the method is said to be *mixed*. The least ingenuity is required here. The usual method of choice is the interior method.

The second step is to select an optimal set of constants  $a_i$ ,  $i = 1, 2, \dots, n$ , by using the residual  $R_I(U) = LU - f$ . This is done here for the interior method. In the boundary method, there are a set of boundary residual  $R_B$ , and, in the mixed method, both  $R_I$  and  $R_B$ . Using the spatial average  $(w, v) = \int_V w v dV$ , the criterion for selecting the values of  $a_i$  is the requirement that the  $n$  spatial averages

$$(b_i, R_E(U)) = 0, \quad i = 1, 2, \dots, n$$

These represent  $n$  equations (linear if the operator  $L$  is linear and nonlinear otherwise) for the  $a_i$ .

Particular WRMs differ because of the choice of the  $b_j$ s. The most common follow.

1. *Subdomain* The domain  $V$  is divided into  $n$  smaller, not necessarily disjoint, subdomains  $V_j$  with  $w_j(x,y) = 1$  if  $(x,y)$  is in  $V_j$ , and 0 if  $(x,y)$  is not in  $V_j$ .
2. *Collocation* Select  $n$  points  $P_j = (x_j, y_j)$  in  $V$  with  $w_j(P_j) = \delta(P - P_j)$ , where  $\int_V \phi(P) \delta(P - P_j) dP = \phi(P_j)$  for all test functions  $\phi(P)$  which vanish outside the compact set  $V$ . Thus,  $(w_j, R_E) = \int_V \delta(P - P_j) R_E dV = R_E[U(P_j)] \equiv 0$  (i.e., the residual is set equal to zero at the  $n$  points  $P_j$ ).
3. *Least squares* Here, the functional  $I(a) = \int_V R_E^2 dV$ , where  $a = (a_1, \dots, a_n)$ , is to be made stationary with respect to the  $a_j$ . Thus,  $0 = \partial I / \partial a_j = 2 \int_V R_E (\partial R_E / \partial a_j) dV$ , with  $j = 1, 2, \dots, n$ . The  $w_j$  in this case are  $\partial R_E / \partial a_j$ .
4. *Bubnov-Galerkin* Choose  $w_j(P) = b_j(P)$ . This is perhaps the best-known method.

5. *Stationary Functional (Variational) Method* With  $\phi$  a variational integral (or other functional), set  $\partial\phi[U]/\partial a_j = 0$ , where  $j = 1, \dots, n$ , to generate the  $n$  algebraic equations.

**Example.**  $u_{xx} + u_{yy} = -2$ , with  $u = 0$  on the boundaries of the square  $x = \pm 1, y = \pm 1$ . Select an interior method with  $U = a_1(1 - x^2)(1 - y^2) + a_2x^2y^2(1 - x^2)(1 - y^2)$ , whereupon the residual  $R_E(U) = -2a_1(2 - x^2 - y^2) + 2a_2[(1 - 6x^2)y^2(1 - y^2) + (1 - 6y^2)x^2(1 - x^2)] + 2$ . Collocating at  $(\frac{1}{3}, \frac{1}{3})$  and  $(\frac{2}{3}, \frac{2}{3})$  gives the two linear equations  $-32a_1/9 + 32a_2/243 + 2 = 0$  and  $-20a_1/9 - 400a_2/243 + 2 = 0$  for  $a_1$  and  $a_2$ .

WRM methods can obviously be used as approximate methods. We have now set the stage for *finite elements*.

## Finite Elements

The WRM methods are more general than the *finite element* (FE) methods. FE methods require, in addition, that the basis functions be finite elements (i.e., functions that are zero except on a small part of the domain under consideration). A typical example of an often used basis is that of triangular elements. For a triangular element with Cartesian coordinates  $(x_1, y_1)$ ,  $(x_2, y_2)$ , and  $(x_3, y_3)$ , define natural coordinates  $L_1, L_2$ , and  $L_3$  ( $L_i \leftrightarrow (x_i, y_i)$ ) so that  $L_i = A_i/A$  where

$$A = \frac{1}{2} \det \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix}$$

is the area of the triangle and

$$A_1 = \frac{1}{2} \det \begin{bmatrix} 1 & x & y \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix}$$

$$A_2 = \frac{1}{2} \det \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x & y \\ 1 & x_3 & y_3 \end{bmatrix}$$

$$A_3 = \frac{1}{2} \det \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x & y \end{bmatrix}$$

Clearly  $L_1 + L_2 + L_3 = 1$ , and the  $L_i$  are one at node  $i$  and zero at the other nodes. In terms of the Cartesian coordinates,

$$\begin{bmatrix} L_1 \\ L_2 \\ L_3 \end{bmatrix} = \frac{1}{2A} \begin{bmatrix} x_2 y_3 - x_3 y_2, & y_2 - y_3, & x_3 - x_2 \\ x_3 y_1 - x_1 y_3, & y_3 - y_1, & x_1 - x_3 \\ x_1 y_2 - x_2 y_1, & y_1 - y_2, & x_2 - x_1 \end{bmatrix} \begin{bmatrix} 1 \\ x \\ y \end{bmatrix}$$

is the linear triangular element relation.

Tables of linear, quadratic, and cubic basis functions are given in the literature. Notice that while the linear basis needs three nodes, the quadratic requires six and the cubic basis ten. Various modifications, such as the Hermite basis, are described in the literature. Triangular elements are useful in approximating irregular domains.

For rectangular elements, the *chapeau* functions are often used. Let us illustrate with an example. Let  $u_{xx} + u_{yy} = Q$ ,  $0 < x < 2$ ,  $0 < y < 2$ ,  $u(x, 2) = 1$ ,  $u(0, y) = 1$ ,  $u_y(x, 0) = 0$ ,  $u_x(2, y) = 0$ , and  $Q(x, y) = Q_w \delta(x - 1) \delta(y - 1)$ ,

$$\delta(x) = \begin{cases} 0 & x \neq 0 \\ 1 & x = 0 \end{cases}$$

Using four equal rectangular elements, map the element  $I$  with vertices at (0,0), (0,1), (1,1), and (1,0) into the local (canonical) coordinates  $(\xi, \eta)$ ,  $-1 \leq \xi \leq 1$ ,  $-1 \leq \eta \leq 1$ , by means of  $x = \frac{1}{2}(\xi + 1)$ ,  $y = \frac{1}{2}(\eta + 1)$ . This mapping permits one to develop software that standardizes the treatment of all elements. Converting to  $(\xi, \eta)$  coordinates, our problem becomes  $u_{\xi\xi} + u_{\eta\eta} = \frac{1}{4}Q$ ,  $-1 \leq \xi \leq 1$ ,  $-1 \leq \eta \leq 1$ ,  $Q = Q_w \delta(\xi - 1) \delta(\eta - 1)$ .

First, a trial function  $\bar{u}(\xi, \eta)$  is defined as  $u(\xi, \eta) \approx \bar{u}(\xi, \eta) = \sum_{j=1}^4 A_j \phi_j(\xi, \eta)$  (in element  $I$ ) where the  $\phi_j$  are the two-dimensional chapeau functions

$$\begin{aligned} \phi_1 &= \left[ \frac{1}{2}(1 - \xi) \frac{1}{2}(1 - \eta) \right] & \phi_2 &= \left[ \frac{1}{2}(1 + \xi) \frac{1}{2}(1 - \eta) \right] \\ \phi_3 &= \left[ \frac{1}{2}(1 + \xi) \frac{1}{2}(1 + \eta) \right] & \phi_4 &= \left[ \frac{1}{2}(1 - \xi) \frac{1}{2}(1 + \eta) \right] \end{aligned}$$

Clearly  $\phi_i$  take the value one at node  $i$ , provide a bilinear approximation, and are nonzero only over elements adjacent to node  $i$ .

Second, the equation residual  $R_E = \nabla^2 \bar{u} - \frac{1}{4}Q$  is formed and a WRM procedure is selected to formulate the algebraic equations for the  $A_i$ . This is indicated using the Galerkin method. Thus, for element  $I$ , we have

$$\int \int_{D_I} (\bar{u}_{\xi\xi} + \bar{u}_{\eta\eta} - Q) \phi_i(\xi, \eta) d\xi d\eta = 0, \quad i = 1, \dots, 4$$

Applying Green's theorem, this result becomes

$$\int \int_{D_I} [\bar{u}_{\xi}(\phi_i)_{\xi} + \bar{u}_{\eta}(\phi_i)_{\eta} + \frac{1}{4}Q \phi_i] d\xi d\eta - \int_{\partial D_I} (\bar{u}_{\xi} c_{\xi} + \bar{u}_{\eta} c_{\eta}) \phi_i ds = 0, \quad i = 1, 2, \dots, 4$$

Using the same procedure in all four elements and recalling the property that the  $\phi_i$  in each element are nonzero only over elements adjacent to node  $i$  gives the following nine equations:

$$\sum_{e=1}^4 \left\{ \int_{D_e} \sum_{j=1}^9 A_j [(\phi_j)_\xi (\phi_i)_\xi + (\phi_j)_\eta (\phi_i)_\eta] + \frac{1}{4} Q \phi_i \right\} d\xi d\eta - \sum_{e=1}^4 \int_{\partial D_e} (\bar{u}_\xi c_\xi + \bar{u}_\eta c_\eta) \phi ds = 0, \quad i = 1, 2, \dots, 9$$

where the  $c_\xi$  and  $c_\eta$  are the direction cosines of the appropriate element ( $e$ ) boundary.

## Method of Lines

The *method of lines*, when used on PDEs in two dimensions, reduces the PDE to a system of ordinary differential equations (ODEs), usually by finite difference or finite element techniques. If the original problem is an initial value (boundary value) problem, then the resulting ODEs form an initial value (boundary value) problem. These ODEs are solved by ODE numerical methods.

**Example.**  $u_t = u_{xx} + u^2$ ,  $0 < x < 1$ ,  $0 < t$ , with the initial value  $u(x,0) = x$ , and boundary data  $u(0,t) = 0$ ,  $u(1,t) = \sin t$ . A discretization of the space variable ( $x$ ) is introduced and the time variable is left continuous. The approximation is  $\dot{u}_i = (u_{i+1} - 2u_i + u_{i-1})/h^2 + u_i^2$ . With  $h = 1/5$ , the equations become

$$\begin{aligned} u_0(t) &= 0 \\ \dot{u}_1 &= \frac{1}{25}[u_2 - 2u_1] + u_1^2 \\ \dot{u}_2 &= \frac{1}{25}[u_3 - 2u_2 + u_1] + u_2^2 \\ \dot{u}_3 &= \frac{1}{25}[u_4 - 2u_3 + u_2] + u_3^2 \\ \dot{u}_4 &= \frac{1}{25}[\sin t - 2u_4 + u_3] + u_4^2 \\ u_5 &= \sin t \end{aligned}$$

and  $u_1(0) = 0.2$ ,  $u_2(0) = 0.4$ ,  $u_3(0) = 0.6$ , and  $u_4(0) = 0.8$ .

## 209.11 Discrete and Fast Fourier Transforms

Let  $x(n)$  be a sequence that is nonzero only for a finite number of samples in the interval  $0 \leq n \leq N - 1$ . The quantity



$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-i(2\pi/N)nk}, \quad k = 0, 1, \dots, N-1$$

is called the *discrete Fourier transform* (DFT) of the sequence  $x(n)$ . Its inverse (IDFT) is given by

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k)e^{i(2\pi/N)nk}, \quad n = 0, 1, \dots, N-1 \quad (i^2 = -1)$$

Clearly, DFT and IDFT are finite sums and there are  $N$  frequency values. Also,  $X(k)$  is periodic in  $k$  with period  $N$ .

**Example.**  $x(0) = 1, x(1) = 2, x(2) = 3, x(3) = 4$

$$X(k) = \sum_{n=0}^3 x(n)e^{-i(2\pi/4)nk}, \quad k = 0, 1, 2, 3, 4$$

Thus,

$$X(0) = \sum_{n=0}^3 x(n) = 10$$

and  $X(1) = x(0) + x(1)e^{-i\pi/2} + x(2)e^{-i\pi} + x(3)e^{-i3\pi/2} = 1 - 2i - 3 + 4i = -2 + 2i$ ;  $X(2) = -2$ ;  $X(3) = -2 - 2i$ .

## DFT Properties

1. **Linearity:** If  $x_3(n) = ax_1(n) + bx_2(n)$ , then  $X_3(k) = aX_1(k) + bX_2(k)$ .
2. **Symmetry:** For  $x(n)$  real,  $\text{Re}[X(k)] = \text{Re}[X(N-k)]$ ,  $\text{Im}[X(k)] = -\text{Im}[X(N-k)]$ .
3. **Circular shift:** By a circular shift of a sequence defined in the interval  $0 \leq n \leq N-1$ , we mean that, as values *fall off* from one end of the sequence, they are appended to the other end. Denoting this by  $x(n \oplus m)$ , we see that positive  $m$  means shift left and negative  $m$  means shift right. Thus,  $x_2(n) = x_1(n \oplus m) \Leftrightarrow X_2(k) = X_1(k)e^{i(2\pi/N)km}$ .
4. **Duality:**  $x(n) \Leftrightarrow X(k)$  implies  $(1/N)X(n) \Leftrightarrow x(-k)$ .
5. **Z-transform relation:**  $X(k) = X(z)|_{z=e^{i(2\pi/N)k}}, k = 0, 1, \dots, N-1$ .
6. **Circular convolution:**  $x_3(n) = \sum_{m=0}^{N-1} x_1(m)x_2(n \ominus m) = \sum_{\ell=0}^{N-1} x_1(n \ominus \ell)x_2(\ell)$  where  $x_2(n \ominus m)$  corresponds to a circular shift to the right for positive  $m$ .

One fast algorithm for calculating DFTs is the radix-2 *fast Fourier transform* developed by J. W. Cooley and J. W. Tucker. Consider the two-point DFT  $X(k) = \sum_{n=0}^1 x(n)e^{-i(2\pi/2)nk}, k = 0, 1$ . Clearly,  $X(k) = x(0) + x(1)e^{-i\pi k}$ . So,  $X(0) = x(0) + x(1)$  and  $X(1) = x(0) - x(1)$ . This process can be extended to DFTs of length  $N = 2^r$ , where  $r$  is a positive integer. For  $N = 2^r$ , decompose the  $N$ -point DFT into *two*  $N/2$ -point DFTs. Then, decompose each  $N/2$ -point DFT into *two*  $N/4$ -point DFTs, and so on until eventually we have  $N/2$  *two*-point DFTs. Computing these as indicated above, we recombine them into  $N/4$  four-point DFTs and then  $N/8$  eight-point DFTs, and so on, until the DFT is computed. The total number of DFT operations (for large  $N$ ) is  $O(N^2)$ , and that of the FFT is  $O(N \log_2 N)$ , quite a saving for large  $N$ .

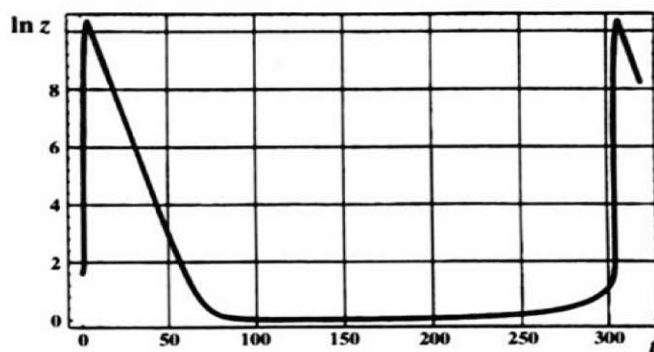


Figure 1

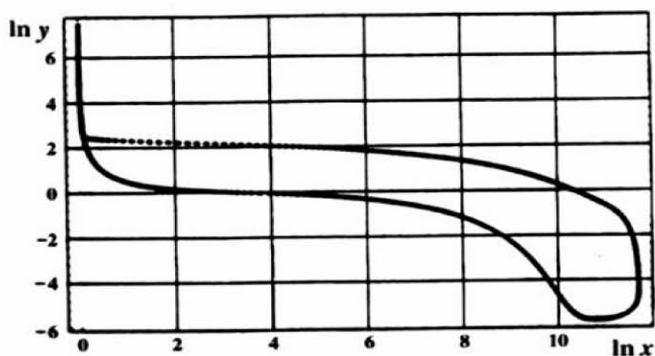


Figure 2

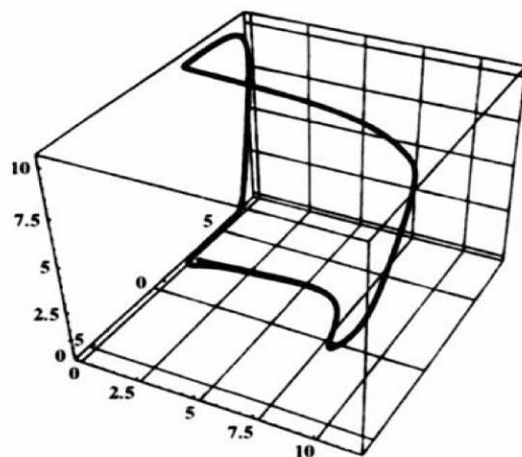


Figure 3

### THE OREGONATOR

The "Oregonator" is a periodic chemical reaction describable by three nonlinear first-order differential equations. The results (Fig. 1) illustrate the periodic nature of the major chemical versus time. Figure 2 shows the phase diagram of two of the reactants, and Fig. 3 is the three-dimensional phase diagram of all reactants. The numerical computation was done using a fourth-order Runge-Kutta method on Mathematica by Waltraud Rufeger at the Georgia Institute of Technology.

## 209.12 Software

---

Some available software is listed here.

### General Packages

General software packages include Maple, Mathematica, and Matlab. All contain algorithms for handling a large variety of both numerical and symbolic computations.

### Special Packages for Linear Systems

In the IMSL Library, there are three complementary linear system packages of note.

LINPACK is a collection of programs concerned with *direct* methods for general (or full) symmetric, symmetric positive definite, triangular, and tridiagonal matrices. There are also programs for least squares problems, along with the QR algorithm for eigensystems and the singular value decompositions of rectangular matrices. The programs are intended to be completely machine independent, fully portable, and run with good efficiency in most computing environments. The LINPACK User's Guide by Dongarra *et al.* is the basic reference.

ITPACK is a modular set of programs for iterative methods. The package is oriented toward the sparse matrices that arise in the solution of PDEs and other applications. While the programs apply to full matrices, that is rarely profitable. Four basic iteration methods and two convergence acceleration methods are in the package. There is a Jacobi, SOR (with optimum relaxation parameter estimated), symmetric SOR, and reduced system (red-black ordering) iteration, each with semi-iteration and conjugate gradient acceleration. All parameters for these iterations are automatically estimated. The practical and theoretical background for ITPACK is found in Hagemen and Young [1981].

YALEPACK is a substantial collection of programs for sparse matrix computations.

### Ordinary Differential Equations Packages

Also in IMSL, one finds such sophisticated software as DVERK, DGEAR, or DREBS for initial value problems. For two-point boundary value problems, one finds DTPTB (use of DVERK and multiple shooting) or DVCPR.

## Partial Differential Equations Packages

DISPL was developed and written at Argonne National Laboratory. DISPL is designed for nonlinear second-order PDEs (parabolic, elliptic, hyperbolic (some cases), and parabolic-elliptic). Boundary conditions of a general nature and material interfaces are allowed. The spatial dimension can be either one or two and in Cartesian, cylindrical, or spherical (one dimension only) geometry. The PDEs are reduced to ordinary DEs by Galerkin discretization of the spatial variables. The resulting ordinary DEs in the timelike variable are then solved by an ODE software package (such as GEAR). Software features include graphics capabilities, printed output, dump/restart facilities, and free format input. DISPL is intended to be an engineering and scientific tool and is not a finely tuned production code for a small set of problems. DISPL makes no effort to control the spatial discretization errors. It has been used to successfully solve a variety of problems in chemical transport, heat and mass transfer, pipe flow, etc.

PDELIB was developed and written at Los Alamos Scientific Laboratory. PDELIB is a library of subroutines to support the numerical solution of evolution equations with a timelike variable and one or two space variables. The routines are grouped into a dozen independent modules according to their function (i.e., accepting initial data, approximating spatial derivatives, advancing the solution in time). Each task is isolated in a distinct module. Within a module, the basic task is further refined into general-purpose flexible lower-level routines. PDELIB can be understood and used at different levels. Within a small period of time, a large class of problems can be solved by a novice. Moreover, it can provide a wide variety of outputs.

DSS/2 is a differential systems simulator developed at Lehigh University as a transportable numerical method of lines (NMOL) code. See also LEANS.

FORSIM is designed for the automated solution of sets of implicitly coupled PDEs of the form

$$\frac{\partial u_i}{\partial t} = \phi_i(x, t, u_i, u_j, \dots, (u_i)_x, \dots, (u_i)_{xx}, (u_j)_{xx}, \dots), \quad \text{for } i = 1, \dots, N$$

The user specifies the  $\phi_i$  in a simple FORTRAN subroutine. Finite difference formulas of any order may be selected for the spatial discretization and the spatial grid need not be equidistant. The resulting system of time-dependent ODEs is solved by the method of lines.

SLDGL is a program package for the self-adaptive solution of nonlinear systems of elliptic and parabolic PDEs in up to three space dimensions. Variable step size and variable order are permitted. The discretization error is estimated and used for the determination of the optimum grid and optimum orders. This is the most general of the codes described here (not for hyperbolic systems, of course). This package has seen extensive use in Europe.

FIDISOL (finite difference solver) is a program package for nonlinear systems of two- or three-dimensional elliptic and parabolic systems in rectangular domains or in domains that can be transformed analytically to rectangular domains. This package is actually a redesign of parts of SLDGL, primarily for the solution of large problems on vector computers. It has been tested on the CYBER 205, CRAY-1M, CRAY X-MP/22, and VP 200. The program vectorizes very well and uses the vector arithmetic efficiently. In addition to the numerical solution, a reliable error estimate is computed.

CAVE is a program package for conduction analysis via eigenvalues for three-dimensional geometries using the method of lines. In many problems, much time is saved because only a few terms suffice.

Many industrial and university computing services subscribe to the IMSL Software Library. Announcements of new software appear in *Directions*, a publication of IMSL. A brief description of some IMSL packages applicable to PDEs and associated problems is now given. In addition to those packages just described, two additional software packages bear mention. The first of these, the ELLPACK system, solves elliptic problems in two dimensions with general domains and in three dimensions with box-shaped domains. The system contains over 30 numerical methods modules, thereby providing a means of evaluating and comparing different methods for solving elliptic problems. ELLPACK has a special high-level language making it easy to use. New algorithms can be added or deleted from the system with ease.

Second, TWODEPEP is IMSL's general finite element system for two-dimensional elliptic, parabolic, and eigenvalue problems. The Galerkin finite elements available are triangles with quadratic, cubic, or quartic basic functions, with one edge curved when adjacent to a curved boundary, according to the isoparametric method. Nonlinear equations are solved by Newton's method, with the resulting linear system solved directly by Gauss elimination. PDE/PROTRAN is also available. It uses triangular elements with piecewise polynomials of degree 2, 3, or 4 to solve quite general steady state, time-dependent, and eigenvalue problems in general two-dimensional regions. There is a simple user input. Additional information may be obtained from IMSL. NASTRAN and STRUDL are two advanced finite element computer systems available from a variety of sources. Another, UNAFEM, has been extensively used.

## References

### General

- Adams, E. and Kulisch, U. (Eds.) 1993. *Scientific Computing with Automatic Result Verification*. Academic Press, Boston, MA.
- Gerald, C. F. and Wheatley, P. O. 1984. *Applied Numerical Analysis*. Addison-Wesley, Reading, MA.
- Hamming, R. W. 1962. *Numerical Methods for Scientists and Engineers*. McGraw-Hill, New York.
- Hildebrand, F. B. 1956. *Introduction to Numerical Analysis*. McGraw-Hill, New York.
- Isaacson, E. and Keller, H. B. 1966. *Analysis of Numerical Methods*. John Wiley & Sons, New York.
- Kopal, Z. 1955. *Numerical Analysis*. John Wiley & Sons, New York.
- Rice, J. R. 1993. *Numerical Methods, Software and Analysis*, 2d ed. Academic Press, Boston, MA.
- Stoer, J. and Bulirsch, R. 1976. *Introduction to Numerical Analysis*. Springer, New York.

### Linear Equations

- Bodewig, E. 1956. *Matrix Calculus*. Wiley (Interscience), New York.

- Hageman, L. A. and Young, D. M. 1981. *Applied Iterative Methods*. Academic Press, Boston, MA.
- Varga, R. S. 1962. *Matrix Iterative Numerical Analysis*. John Wiley & Sons, New York.
- Young, D. M. 1971. *Iterative Solution of Large Linear Systems*. Academic Press, Boston, MA.

## Ordinary Differential Equations

- Aiken, R. C. 1985. *Stiff Computation*. Oxford University Press, New York.
- Gear, C. W. 1971. *Numerical Initial Value Problems in Ordinary Differential Equations*. Prentice Hall, Englewood Cliffs, NJ.
- Keller, H. B. 1976. *Numerical Solutions of Two Point Boundary Value Problems*. SIAM, Philadelphia, PA.
- Lambert, J. D. 1973. *Computational Methods in Ordinary Differential Equations*. Cambridge University Press, New York.
- Milne, W. E. 1953. *Numerical Solution of Differential Equations*. John Wiley & Sons, New York.
- Rockey, K. C., Evans, H. R., Griffiths, D. W., and Nethercot, D. A. 1983. *The Finite Element Method 3/4 A Basic Introduction for Engineers*, 2d ed. Halstead Press, New York.
- Shampine, L. and Gear, C. W. 1979. A User's View of Solving Stiff Ordinary Differential Equations, *SIAM Rev.* 21:1-17.

## Partial Differential Equations

- Ames, W. F. 1993. *Numerical Methods for Partial Differential Equations*, 3d ed. Academic Press, Boston, MA.
- Brebbia, C. A. 1984. *Boundary Element Techniques in Computer Aided Engineering*. Martinus Nijhoff, Boston, MA.
- Burnett, D. S. 1987. *Finite Element Analysis*. Addison-Wesley, Reading, MA.
- Lapidus, L. and Pinder, G. F. 1982. *Numerical Solution of Partial Differential Equations in Science and Engineering*. John Wiley & Sons, New York.
- Roache, P. 1972. *Computational Fluid Dynamics*. Hermosa, Albuquerque, NM.

Ames, W. F. "Dimensional Analysis"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## 210.1 Units and Variables

## 210.2 Method of Dimensions

**William F. Ames***Georgia Institute of Technology***210.1 Units and Variables**

Dimensional analysis is a mathematical tool whose use will enable the scientist and engineer to save time in planning experiments and correlating results of experiments. The method arose in the science of mechanics. In this science three *units* are regarded as *fundamental*, namely, *length*, *mass*, and *time*. Other employed units are called derived units. For example, velocity and acceleration are derived units defined by reference to the two fundamental units—length and time. The units of force-momentum, mechanical energy, and power—are dependent on all three of the fundamental units. The study of electricity, heat transfer, and so forth requires the inclusion of other fundamental units such as charge.

Here the fundamental units adopted are *length* (in centimeters),  $L$ ; *mass* (in grams),  $M$ ; and *time* (in seconds),  $T$ . This is the CGS system of units. Others are the British, or foot, pound, second (FPS), system and the MKS system, which is closely allied to the CGS system. With the CGS system the dimensional formulas of some basic physical *variables* are given in [Tables 210.1](#) through [210.3](#).

**Table 210.1** Dimensional Physical Quantities

Physical Quantity	Exponents of Dimensions			Formula
	$L$	$M$	$T$	
Volume density	−3	1	0	$L^{-3} M$
Length per unit volume	−2	0	0	$L^{-2}$
Area density	−2	1	0	$L^{-2} M$
Curvature	−1	0	0	$L^{-1}$
Linear density	−1	1	0	$L^{-1} M$
Angle	0	0	0	
Mass	0	1	0	$M$
Length	1	0	0	$L$



Mass $\times$ length	1	1	0	$LM$
Area	2	0	0	$L^2$
Moment of inertia	2	1	0	$L^2 M$
Volume	3	0	0	$L^3$
Speed of density change	-3	1	-1	$L^{-3} MT^{-1}$
Velocity per unit volume	-2	0	-1	$L^{-2} T^{-1}$
Momentum per unit volume	-2	1	-1	$L^{-2} MT^{-1}$
Velocity per unit area	-1	0	-1	$L^{-1} T^{-1}$
Viscosity	-1	1	-1	$L^{-1} MT^{-1}$
Frequency	0	0	-1	$T^{-1}$
Mass per second	0	1	-1	$MT^{-1}$
Velocity	1	0	-1	$LT^{-1}$
Momentum	1	1	-1	$LMT^{-1}$
Kinematic viscosity	2	0	-1	$L^2 T^{-1}$
Action	2	1	-1	$L^2 MT^{-1}$
Volume per second	3	0	-1	$L^3 T^{-1}$
Acceleration of density change	-3	1	-2	$L^{-3} MT^{-2}$
Acceleration per unit volume	-2	0	-2	$L^{-2} T^{-2}$
Force per unit volume	-2	1	-2	$L^{-2} MT^{-2}$
Acceleration per unit area	-1	-0	-2	$L^{-1} T^{-2}$
Pressure	-1	1	-2	$L^{-1} MT^{-2}$
Angular acceleration	0	0	-2	$T^{-2}$
Surface tension	0	1	-2	$MT^{-2}$
Acceleration	1	0	-2	$LT^{-2}$
Force	1	1	-2	$LMT^{-2}$
Temperature	2	0	-2	$L^2 T^{-2}$
Energy, torque	2	1	-2	$L^2 MT^{-2}$
Rate of change of volume per second	3	0	-2	$L^3 T^{-2}$
Power	2	1	-3	$L^2 MT^{-3}$

**Table 210.2** Dimensional Thermal Quantities

Physical Quantity	Thermal Units	Dynamic Units
Temperature	$\theta$	$\theta$
Quantity of heat	$H$	$L^2 MT^{-2}$
Specific heat	Dimensionless	Dimensionless
Heat capacity per unit mass	$L^0 M^{-1} T^0 H \theta^{-1}$	$L^2 M^0 T^{-2} \theta^{-1}$
Heat capacity per unit volume	$L^{-3} M^0 T^0 H \theta^{-1}$	$L^{-1} MT^{-2} \theta^{-1}$
Temperature gradient	$L^{-1} M^0 T^0 H^0 \theta^{-1}$	$L^{-1} M^0 T^0 \theta^{-1}$
Conductivity	$L^{-1} M^0 T^{-1} H^0 \theta^{-1}$	$LMT^{-3} \theta^{-1}$
Entropy	$H \theta^{-1}$	$L^2 MT^{-2} \theta^{-1}$

**Table 210.3** Dimensional Magnetic and Electrical Quantities

Physical Quantity	Electromagnetic	Electrostatic
Magnetic pole, $P$	$L^{3/2} M^{1/2} T^{-1} \mu^{1/2}$	$L^{1/2} M^{1/2} T^0 \kappa^{-1/2}$
Strength of magnetic field, $H$	$L^{-1/2} M^{1/2} T^{-1} \mu^{-1/2}$	$L^{1/2} M^{1/2} T^{-2} \kappa^{1/2}$
Magnetic and electric induction, $\mu H, \kappa E$	$L^{-1/2} M^{1/2} T^{-1} \mu^{1/2}$	$L^{-1/2} M^{1/2} T^{-1} \kappa^{1/2}$
Magnetic and electric moments, $PL, QL$	$L^{5/2} M^{1/2} T^{-1} \mu^{1/2}$	$L^{5/2} M^{1/2} T^{-1} \kappa^{1/2}$
Electric current, $I$	$L^{1/2} M^{1/2} T^{-1} \mu^{-1/2}$	$L^{3/2} M^{1/2} T^{-2} \kappa^{1/2}$
Quantity of electricity, $Q$	$L^{1/2} M^{1/2} T^0 \mu^{-1/2}$	$L^{3/2} M^{1/2} T^{-1} \kappa^{1/2}$
Potential difference, $E$	$L^{3/2} M^{1/2} T^{-2} \mu^{1/2}$	$L^{1/2} M^{1/2} T^{-1} \kappa^{-1/2}$
Resistance, $R$	$LM^0 T^{-1} \mu$	$L^{-1} M^0 T \kappa^{-1}$
Capacitance, $C$	$L^{-1} M^0 T^2 \mu^{-1}$	$LM^0 T^0 \kappa$
Inductance, $ET/I$	$LM^0 T^0 \mu$	$L^{-1} M^0 T^2 \kappa^{-1}$
Permeability, $\mu$	$L^0 M^0 T^0 \mu$	$L^{-2} M^0 T^2 \kappa^{-1}$
Permittivity, $\kappa$	$L^{-2} M^0 T^2 \mu^{-1}$	$L^0 M^0 T^0 \kappa$

Some *variables* are dimensionless. For example, angle is defined as arc/radius; strain as volume/volume; and specific gravity as density/density (dimensions  $(L^{-3} M)/(L^{-3} M) = L^0 M^0$ ). In addition there are dimensional constants:  $c$ , the velocity of light in a vacuum, has dimensions  $LT^{-1}$ , and  $g$ , the gravitational constant, has dimensions  $L^3 M^{-1} T^{-2}$ . Some dimensionless constants also appear. An example is the Reynolds number  $N_{\text{Re}} = Dv\rho/\mu$  ( $D$  is diameter), which has dimensions  $(L \cdot LT^{-1} \cdot L^{-3} M)/(L^{-1} MT^{-1}) = L^0 M^0 T^0$ . Some dimensionless groups are listed in Table 210.4.

**Table 210.4** Dimensionless Groups in Engineering

Biot number	$N_{\text{Bi}}$	$hL/k$
Condensation number	$N_{\text{Co}}$	$(h/k)(\mu^2/\rho^2 g)^{1/3}$
Number used in condensation of vapors	$N_{\text{Cv}}$	$L^3 \rho^2 g \lambda / \kappa \mu \Delta t$
Euler number	$N_{\text{Eu}}$	$g_e(-dp)/\rho V^2$
Fourier number	$N_{\text{Fo}}$	$k\theta/\rho c L^2$
Froude number	$N_{\text{Fr}}$	$V^2/Lg$
Graetz number	$N_{\text{Gz}}$	$wc/kL$
Grashof number	$N_{\text{Gr}}$	$L^3 \rho^2 \beta g \Delta t / \mu^2$
Mach number	$N_{\text{Ma}}$	$V/V_a$
Nusselt number	$N_{\text{Nu}}$	$hD/k$
Peclet number	$N_{\text{Pe}}$	$DV\rho c/k$
Prandtl number	$N_{\text{Pr}}$	$c\mu/k$
Reynolds number	$N_{\text{Re}}$	$DV\rho/\mu$
Schmidt number	$N_{\text{Sc}}$	$\mu/\rho D_v$
Stanton number	$N_{\text{St}}$	$h/cV\rho$
Weber number	$N_{\text{We}}$	$LV^2\rho/\sigma g_e$

## 210.2 Method of Dimensions

The *method of dimensions* is based on the simple observation that the dimensions on both sides of an equation must be the same. To illustrate, suppose we wish to find the distance  $s$  traveled in a given time by a particle of mass  $m$  falling from rest under the uniform acceleration due to gravity  $g$ . Assuming  $s = f(g, t, m) = Cg^a t^b m^c$ , where  $C$  is an unknown dimensionless constant and the indices  $a$ ,  $b$ , and  $c$  are to be determined. Examining the dimensions in a dimensional formula we have

$$L = (LT^{-2})^a T^b M^c$$

Clearly  $a = 1$ ,  $b = 2$ , and  $c = 0$ , so  $s = Cgt^2$ . It is not within the realm of dimensional analysis to evaluate  $C$ .

Another example is the two-body problem of astronomy. Suppose a planet of mass  $m$  rotates about the sun (mass  $\overline{M}$ ) in an orbit that is an ellipse of major diameter  $D$ . The physical quantities that will affect the sought-for period,  $t$ , appear to be the mass of the sun, the mass of the planet, the diameter  $D$ , and the gravitational constant  $G$ . Thus we try  $t = C\overline{M}^a m^b D^c G^d + (\text{similar terms})$ , where the similar terms must have the same dimensions, that is, 0 in length, 0 in mass, and 1 in time. The dimensional equation becomes  $T = M^a M^b L^c (L^3 M^{-1} T^{-2})^d$ . Hence

$$0 = c + 3d \quad (\text{length})$$

$$0 = a + b - d \quad (\text{mass})$$

$$1 = -2d \quad (\text{time})$$

The solution of this system is  $d = -\frac{1}{2}$ ,  $c = \frac{3}{2}$ ,  $a$  arbitrary, and  $b = -\frac{1}{2} - a$ . Therefore  $t = C\overline{M}^a m^{(-1/2)-a} D^{3/2} G^{-1/2}$  or  $t^2 = C'(\overline{M}/m)^{2a} (D^3/Gm)$ , which is Kepler's law (the square of the period of a planet is proportional to the cube of the major axis of the orbit). The constants  $C'$  and  $a$  must be evaluated by experiment.

Many other examples are found in the literature. This section is closed with the determination of the height,  $h$ , that a liquid rises in a capillary tube. The relevant equation is  $h = C \cdot \rho^a r^b s^c g^d \theta^e$ , where  $\rho$  is the fluid density,  $r$  is the radius of the tube,  $s$  is the surface tension of the fluid,  $g$  is the acceleration of gravity, and  $\theta$  is the contact angle. The dimensional equation becomes  $L = (L^{-3} M)^a L^b (MT^{-2})^c (LT^{-2})^d$ , which gives rise to the following three linear equations:

$$1 = -3a + b + d \quad (\text{length})$$

$$0 = a + c \quad (\text{mass})$$

$$0 = -2c - 2d \quad (\text{time})$$

in the four unknowns  $a$ ,  $b$ ,  $c$ , and  $d$ . We may express any three of the unknowns in terms of the fourth. Choosing  $a$  we find  $c = -a$ ,  $d = a$ , and  $b = 1 + 2a$ , so  $h = cr(r^2\rho g/s)^a$ . Experiments show that  $h$  is inversely proportional to  $r$ , so that  $a = -1$ . Thus  $h = C(s/r\rho g)$ .

## References

Bridgeman, P. W. 1937. *Dimensional Analysis*. Yale University Press, New Haven, CT.  
Huntley, H. E. 1952. *Dimensional Analysis*. MacDonald, London.

Gallagher, R. S. "Computer Graphics Visualization"  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Computer Graphics Visualization

---

- 211.1 The Display of Objects in 3-D
- 211.2 Scalar Display Techniques
- 211.3 Vector and Tensor Field Display
- 211.4 Continuum Volume Visualization
- 211.5 Animation over Time
- 211.6 Summary

**Richard S. Gallagher**

*R. S. Gallagher and Associates*

Visualization is the process of seeing something. In the computer graphics field, the term *visualization* has a more specific meaning, and refers to computational techniques to display and understand behavior. The field of computer graphics visualization is still young; much of its fundamental work took place in the late 1980s. Today, it has grown to become one of the key commercial applications of computer graphics. Key technical aspects of computer graphics visualization include the display of objects in 3-D; display of scalar, vector, and tensor quantities; continuum volume display techniques; and animation over time.

## 211.1 The Display of Objects in 3-D

---

Most modern computer graphics equipment uses display technology similar to that of television, where a two-dimensional array of dots known as **pixels** have individual dot colors set to produce a composite image. Turning data into such a screen image involves representing the data as geometric components, such as lines or polygons, *transforming* these components to a particular view from the observer, coloring or *rendering* these components, and converting them to dots on the screen through a process known as **scan conversion**.

Transformations are generally performed through matrix multiplication of the coordinates of each component, effecting operations including *rotation*, *scaling*, and *translation*. The transformation of a point  $P$  to its transformed position  $P'$  takes the form:

$$[P'] = [T][P] = [r][s][t][P]$$

Scaling and translation matrices can be combined in the form:

$$\begin{bmatrix} s_x & 0 & 0 & \Delta x \\ 0 & s_y & 0 & \Delta y \\ 0 & 0 & s_z & \Delta z \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

which multiplies the vector  $[x, y, z, W]$ , with the fourth element being a nonzero *homogenous coordinate* factor of  $x, y$ , and  $z$ . The rotational transformation is added by multiplying this matrix by matrices corresponding to rotations about the  $x, y$ , and  $z$  axes:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta & 0 \\ 0 & \sin \theta & \cos \theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos \theta & 0 & \sin \theta & 0 \\ 0 & 1 & 0 & 0 \\ -\sin \theta & 0 & \cos \theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos \theta & -\sin \theta & 0 & 0 \\ \sin \theta & \cos \theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Other transformations include the simulation of *perspective*, which reduces the  $x$  and  $y$  coordinates in the current view as a function of the  $z$  coordinate, and *clipping*, which removes portions of components intersected by either screen boundaries or an arbitrary *clipping plane*.

Rendering techniques generally involve setting a color or color function across each component, often based on the effects of a light source from the observer to the screen. For example, a square polygon facing the observer may be red, while its color may change towards darker shades of red as the polygon is rotated, becoming completely dark when 90 degrees from the observer. Shading techniques may compute color as a continuous function across a component. Gouraud shading, for example, interpolates a color value across a polygon from its corner values, while another technique known as Phong shading computes color from an interpolation of the light source vector itself across the polygon. More advanced rendering techniques include **ray tracing**, which computes the behavior of light rays to simulate effects such as reflectance, shadows, and translucency, and *texture mapping*, which simulates a pattern or image across surfaces of the displayed model.

## 211.2 Scalar Display Techniques

---

One of the fundamental operations of visualization is the display of a single variable in space. Some of the visual techniques for displaying scalar field values include:

*Color coding* The outside visible surfaces of a model are color-coded with values corresponding to the scalar value (Fig. 211.1). For example, a structural model may have elements of high stress colored red, and low stress colored blue.

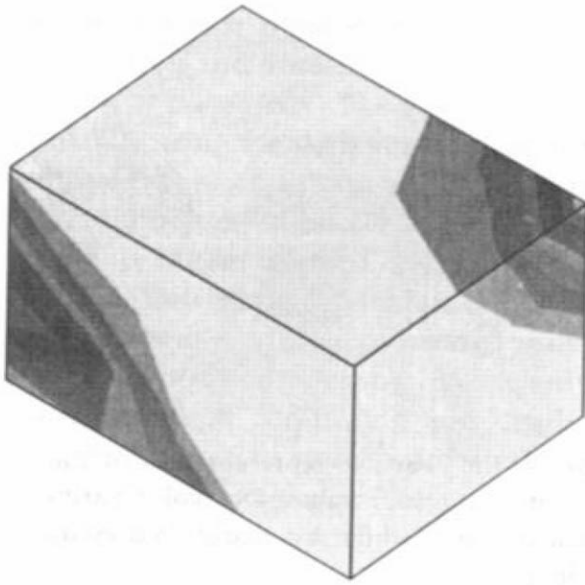
*Isovalue display* An **isovalue** is a region of constant scalar value. In the general case, an isovalue can be idealized as a point in one-dimensional space, a curve in two dimensions, and a surface in three dimensions (Fig. 211.2). One particular form of isovalue display, the *contour plot*, displays bands of color on model surfaces corresponding to ranges of isovalues.

Isosurfaces are surfaces of constant 3-D scalar value.

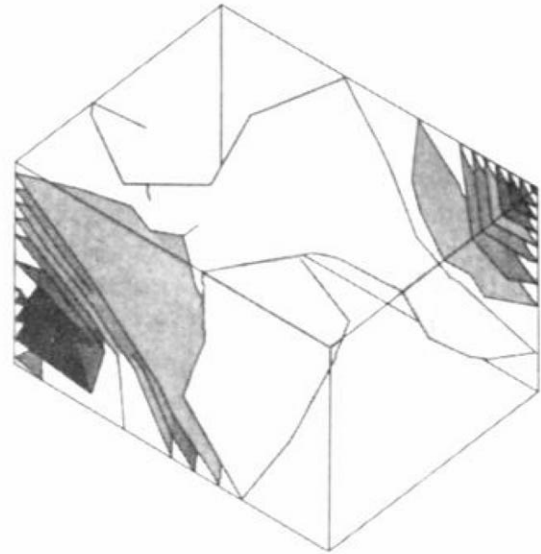
*Particle displays and implicit isovalues* A color-coded distribution of particles within a 3-D

scalar field gives an overview of values throughout the field and can be useful in situations where isosurfaces obscure portions of the interior model. Another application of particle display is to compute and display particles at points corresponding to a particular scalar value. As the density of such particles increases, the resulting image approaches the isosurface of the scalar value in the limit.

**Figure 211.1** Contour representation of a scalar value.



**Figure 211.2** Isosurfaces of a scalar value.



## 211.3 Vector and Tensor Field Display

---

The comprehensible display of multidimensional quantities remains a key area of visualization research. Particular applications include the display of flow fields, multivariate field analysis problems, and examination of derived quantities such as gradients. Some of the current techniques used in vector and tensor field display include:

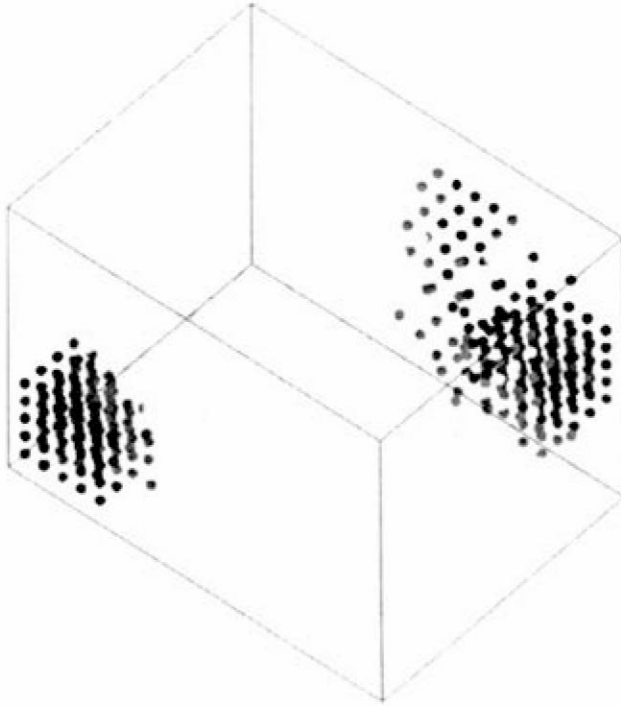
*Vector and tensor symbols* Symbols, often known as **glyphs**, use size, shape and color to show multiple values at a point in space. For example, a three-pointed triad may have each point colored and sized according to the value of different quantities at that point. More complex glyphs may vary the shape, direction, and color of multiple components to show complex states of behavior, such as the state of a tensor field.

*Vector and tensor field curves* Path curves can be created through a vector or tensor field, generally representing the path of a vector quantity from specified starting locations. Further dimensions of information can be displayed across these path curves by varying quantities such as color, thickness, and cross-sectional geometry.

*Particle field display* A distribution of particles within a field can represent a multivariate quantity by varying attributes such as particle density, color, size, and shape. As one example, [Fig. 211.3](#) shows a particle distribution whose density varies by the gradient, or rate of change, of a state of stress in the Cartesian  $x$ ,  $y$ , and  $z$  directions.



**Figure 211.3** Particle representation of the gradient of a scalar value. Dot color varies by scalar value, while dot density varies by gradient.



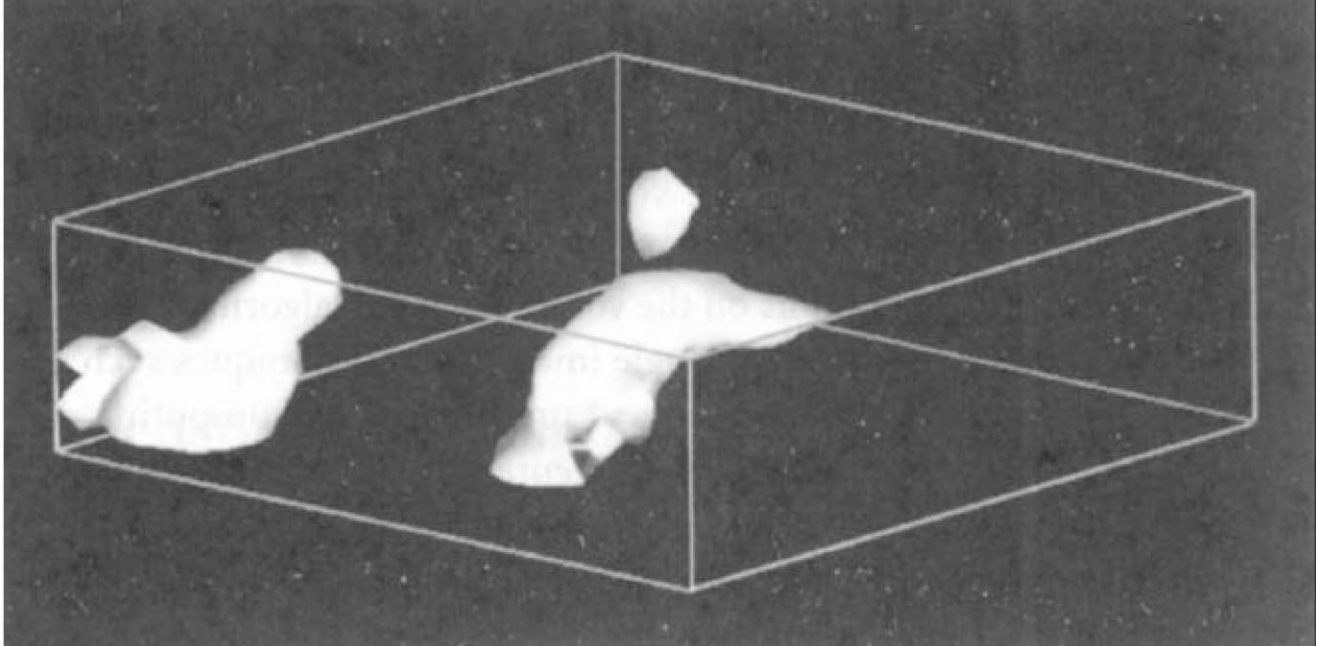
All of these techniques share a common approach of combining several visual techniques within a single image to display multivariate behavior. In addition, the use of animation (discussed below) adds the dimension of time and motion to evaluating complex behavior visually.

## 211.4 Continuum Volume Visualization

---

A three-dimensional field, particularly a sampled field such as a medical image, can be represented as a field of volume elements or **voxels**. These voxels can be viewed as tiny cubes connecting a regular array of sample points within the volume. Continuum volume visualization techniques generate imagery by performing operations on the voxel field. Such algorithms can allow an entire scalar-valued field to be represented within a single image, using techniques such as translucency and color variation, and can also provide a more accurate means of computing discrete images, such as isosurfaces (Fig. 211.4). Above all, a voxel representation of a volume keeps an entire volumes data in a form suited to direct visualization and analysis.

**Figure 211.4** Isovalues of density within a voxel dataset.



Many voxel-based techniques operate in either *image space* or *voxel space*. Image space techniques often involve casting rays through each dot of the screen image into the projected voxel array, computing contributions to the ray's color and opacity from each voxel the ray intersects. Such algorithms may work from front-to-back, stopping when a ray becomes opaque or exits the volume, or back-to-front. Preclassification of the voxel space can help optimize these techniques or allow computations to be performed in parallel.

Voxel space techniques evaluate voxels to create polygons which contribute to the image. One such example, the Marching Cubes algorithm patented in 1987 by General Electric, examines which vertices of a voxel are above or below a threshold value, and then orders these binary values into a bit string whose value is used to look up the topology of any isosurface polygons passing through the voxel.

## 211.5 Animation over Time

---

With increasing computing capabilities, a growing area of interest in visualization is the display of behavior which varies with time or motion.

At its most basic level, animation is simply the generation and playback of individual images to produce the sensation of continuous movement, which humans perceive at a rate between 10 and 30 frames per second. The steps involved in producing animated visualization sequences include:

*Interpolation of behavior* Within each individual frame, aspects, such as result values or model deformation, can be interpolated between the initial and final positions. While simple linear interpolation is most common, this interpolation function can incorporate issues, such as acceleration and deceleration, as well.

*Motion of the object* The position of the model can be interpolated linearly, or along a path,

generally using the same function of incrementation as would be used for behavior. Issues in the general case of object motion include continuity across multiple connecting path segments, and the combination of one or more viewing transformations, such as translation and rotation.

*Motion of the observer* The observer's position can be likened to the location of a camera within a scene. This camera position can also be interpolated along a path, moving into, out of, or around a scene, while accounting for display characteristics, such as perspective and lighting.

*Frame capture and playback* While animated sequences can be directly generated and displayed when adequate computing and display resources exist, animation above a certain level of complexity must be stored frame-by-frame for later playback. Frame capture can be accomplished in computer memory, as images saved in a given format on computer disk, or by using special-purpose hardware, such as videotape controllers or writable video discs.

## 211.6 Summary

---

Techniques such as these share a common purpose of making objects and behavior more comprehensible to the human observer. In particular, they provide engineers with a means to better understand complex three-dimensional behavior. Trends towards the future in this area include computing performance improvements ranging from better display throughput to massively parallel architectures, a greater degree of interactivity, improved user interfaces, and further development of display algorithms.

### Defining Terms

**Glyph:** A symbol displaying multivariate field values at a point in space.

**Isovalue:** A region of constant scalar value within a field of geometry. For example, an *isosurface* represents constant-value surfaces within a 3-D volume.

**Pixel:** A unit 2-D screen dot location in a computer graphics image.

**Ray Tracing:** A technique for realistic display of geometry and volumetric data, involving computing the path of a light ray from screen pixel locations to its projected or reflected locations within the 3-D model.

**Scan Conversion:** The process of converting geometric entities, such as lines and polygons, to projected screen dot (pixel) locations on a computer graphics display.

**Voxel:** A unit volume element within a regular 3-D array, describing a discrete volume space.

### References

- Foley, J. D., Van Dam, A., Feiner, S. K., and Hughes, J. 1990. *Computer Graphics, Principles and Practice*, 2d ed. Addison-Wesley, Reading, MA.
- Gallagher, R. S. (Ed.) 1994. *Computer Visualization*. CRC Press, Boca Raton, FL.
- Kaufman, A. (Ed.) 1990. *Volume Visualization*. IEEE Computer Society Press, Los Alamitos, CA.

Lorensen, W. and Cline, H. E. 1987. Marching Cubes: A High Resolution 3-D Surface Construction Algorithm. *Computer Graphics* 21(4), 163–169.

### **Further Information**

ACM SIGGRAPH, annual conference proceedings, ACM Press.

IEEE Visualization, annual conference proceedings (1990–present), IEEE Computer Society Press.

“Appendix: Mathematical Tables and Formulae”  
*The Engineering Handbook*.  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

# Appendix:

## Mathematical Tables and Formulae

---

### A.1 Greek Alphabet

### A.2 International System of Units (SI)

Definitions of SI Base Units • Names and Symbols for the SI Base Units • SI Derived Units with Special Names and Symbols • Units in Use Together with the SI

### A.3 Conversion Constants and Multipliers

Recommended Decimal Multiples and Submultiples • Conversion Factors — Metric to English • Conversion Factors — English to Metric • Conversion Factors — General • Temperature Factors • Conversion of Temperatures

### A.4 Physical Constants

General •  $\frac{1}{4}$  Constants • Constants Involving  $e$  • Numerical Constants

### A.5 Symbols and Terminology for Physical and Chemical Qualities

### A.6 Elementary Algebra and Geometry

Fundamental Properties (Real Numbers) • Exponents • Fractional Exponents • Irrational Exponents • Operations with Zero • Logarithms • Factorials • Binomial Theorem • Factors and Expansion • Progression • Complex Numbers • Permutations • Combinations • Algebraic Equations • Geometry

### A.7 Table of Derivatives

Additional Relations with Derivatives

### A.8 Integrals

Elementary Forms • Forms Containing  $(a + bx)$

### A.9 The Fourier Transforms

Fourier Transforms • Finite Sine Transforms • Finite Cosine Transforms • Fourier Sine Transforms • Fourier Cosine Transforms • Fourier Transforms

### A.10 Bessel Functions

Bessel Functions of the First Kind,  $J_n(x)$  (Also Called Simply *Bessel Functions*) • Bessel Functions of the Second Kind,  $Y_n(x)$  (Also Called *Neumann Functions* or *Weber Functions*)

### A.11 Legendre Functions

Associated Legendre Functions of the First Kind,  $P_n^m(x)$

### A.12 Table of Differential Equations

## A.1 Greek Alphabet

---

Greek Letter	Greek Name	English Equivalent	Greek Letter	Greek Name	English Equivalent
A $\alpha$	Alpha	a	N $\nu$	Nu	n
B $\beta$	Beta	b	$\Xi$ $\xi$	Xi	x
$\Gamma$ $\gamma$	Gamma	g	O $o$	Omicron	ō
$\Delta$ $\delta$	Delta	d	$\Pi$ $\pi$	Pi	p
E $\epsilon$	Epsilon	ē	P $\rho$	Rho	r
Z $\zeta$	Zeta	z	$\Sigma$ $\sigma$ $\varsigma$	Sigma	s
H $\eta$	Eta	ē	T $\tau$	Tau	t
$\Theta$ $\theta$ $\vartheta$	Theta	th	Y $\upsilon$	Upsilon	u
I $\iota$	Iota	i	$\Phi$ $\phi$ $\varphi$	Phi	ph
K $\kappa$	Kappa	k	X $\chi$	Chi	ch
$\Lambda$ $\lambda$	Lambda	l	$\Psi$ $\psi$	Psi	ps
M $\mu$	Mu	m	$\Omega$ $\omega$	Omega	ō

## A.2 International System of Units (SI)

---

The International System of units (SI) was adopted by the 11th General Conference on Weights and Measures (CGPM) in 1960. It is a coherent system of units built from seven *SI base units*, one for each of the seven dimensionally independent base quantities: the meter, kilogram, second, ampere, kelvin, mole, and candela, for the dimensions length, mass, time, electric current, thermodynamic temperature, amount of substance, and luminous intensity, respectively. The definitions of the SI base units are given below. The *SI derived units* are expressed as products of powers of the base units, analogous to the corresponding relations between physical quantities but with numerical factors equal to unity.

In the International System there is only one SI unit for each physical quantity. This is either the appropriate SI base unit itself or the appropriate SI derived unit. However, any of the approved decimal prefixes, called *SI prefixes*, may be used to construct decimal multiples or submultiples of SI units.

It is recommended that only SI units be used in science and technology (with SI prefixes where appropriate). Where there are special reasons for making an exception to this rule, it is recommended always to define the units used in terms of SI units. This section is based on information supplied by IUPAC.

### Definitions of SI Base Units

*Meter:* The meter is the length of path traveled by light in vacuum during a time interval

of  $1=299\,792\,458$  of a second (17th CGPM, 1983).

*Kilogram:* The kilogram is the unit of mass; it is equal to the mass of the international prototype of the kilogram (3rd CGPM, 1901).

*Second:* The second is the duration of  $9\,192\,631\,770$  periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom (13th CGPM, 1967).

*Ampere:* The ampere is that constant current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross section, and placed 1 meter apart in vacuum, would produce between these conductors a force equal to  $2 \times 10^{-7}$  newton per meter of length (9th CGPM, 1948).

*Kelvin:* The kelvin, unit of thermodynamic temperature, is the fraction  $1/273.16$  of the thermodynamic temperature of the triple point of water (13th CGPM, 1967).

*Mole:* The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kilogram of carbon-12. When the mole is used, the elementary entities must be specified and may be atoms, molecules, ions, electrons, or other particles, or specified groups of such particles (14th CGPM, 1971).

Examples of the use of the mole:

- 1 mol of  $\text{H}_2$  contains about  $6.022 \times 10^{23}$   $\text{H}_2$  molecules, or  $12.044 \times 10^{23}$  H atoms.
- 1 mol of  $\text{HgCl}$  has a mass of 236.04 g.
- 1 mol of  $\text{Hg}_2\text{Cl}_2$  has a mass of 472.08 g.
- 1 mol of  $\text{Hg}_2^{2+}$  has a mass of 401.18 g and a charge of 192.97 kC.
- 1 mol of  $\text{Fe}_{0.91}\text{S}$  has a mass of 82.88 g.
- 1 mol of  $\text{e}^-$  has a mass of  $548.60 \times 10^{-6}$  g and a charge of  $96.49$  kC.
- 1 mol of photons whose frequency is  $10^{14}$  Hz has energy of about 39.90 kJ.

*Candela:* The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency  $540 \times 10^{12}$  Hz and that has a radiant intensity in that direction of  $(1/683)$  watt per steradian (16th CGPM, 1979).

## Names and Symbols for the SI Base Units

Physical Quantity	Name of SI Unit	Symbol for SI Unit
Length	meter	m
Mass	kilogram	kg
Time	second	s
Electric current	ampere	A
Thermodynamic temperature	kelvin	K
Amount of substance	mole	mol
Luminous intensity	candela	cd



## SI Derived Units with Special Names and Symbols

Physical Quantity	Name of SI Unit	Symbol for SI Unit	Expression in Terms of SI Base Units
Frequency <sup>a</sup>	hertz	Hz	$s^{-1}$
Force	newton	N	$m \cdot kg \cdot s^{-2}$
Pressure, stress	pascal	Pa	$N \cdot m^{-2} = m^{-1} \cdot kg \cdot s^{-2}$
Energy, work, heat	joule	J	$N \cdot m = m^2 \cdot kg \cdot s^{-2}$
Power, radiant flux	watt	W	$J \cdot s^{-1} = m^2 \cdot kg \cdot s^{-3}$
Electric charge	coulomb	C	$A \cdot s$
Electric potential, electromotive force	volt	V	$J \cdot C^{-1} = m^2 \cdot kg \cdot s^{-3} \cdot A^{-1}$
Electric resistance	ohm	$\Omega$	$V \cdot A^{-1} = m^2 \cdot kg \cdot s^{-3} \cdot A^{-2}$
Electric conductance	siemens	S	$\Omega^{-1} = m^{-2} \cdot kg^{-1} \cdot s^3 \cdot A^2$
Electric capacitance	farad	F	$C \cdot V^{-1} = m^{-2} \cdot kg^{-1} \cdot s^4 \cdot A^2$
Magnetic flux density	tesla	T	$V \cdot s \cdot m^{-2} = kg \cdot s^{-2} \cdot A^{-1}$
Magnetic flux	weber	Wb	$V \cdot s = m^2 \cdot kg \cdot s^{-2} \cdot A^{-1}$
Inductance	henry	H	$V \cdot A^{-1} \cdot s = m^2 \cdot kg \cdot s^{-2} \cdot A^{-2}$
Celsius temperature <sup>b</sup>	degree Celsius	$^{\circ}C$	K
Luminous flux	lumen	lm	$cd \cdot sr$
Illuminance	lux	lx	$cd \cdot sr \cdot m^{-2}$
Activity (radioactive)	becquerel	Bq	$s^{-1}$
Absorbed dose (of radiation)	gray	Gy	$J \cdot kg^{-1} = m^2 \cdot s^{-2}$

Physical Quantity	Name of SI Unit	Symbol for SI Unit	Expression in Terms of SI Base Units
Dose equivalent (dose equivalent index)	sievert	Sv	$J \cdot kg^{-1} = m^2 \cdot s^{-2}$
Plane angle	radian	rad	1 = $m \cdot m^{-1}$
Solid angle	steradian	sr	1 = $m^2 \cdot m^{-2}$

<sup>a</sup>For radial (circular) frequency and for angular velocity the unit  $\text{rad s}^{-1}$ , or simply  $s^{-1}$ , should be used, and this may not be simplified to Hz. The unit Hz should be used only for frequency in the sense of cycles per second.

<sup>b</sup>The Celsius temperature  $\theta$  is defined by the equation

$$\theta/^{\circ}C = T/K - 237.15$$

The SI unit of Celsius temperature interval is the degree Celsius,  $^{\circ}C$ , which is equal to the kelvin, K.  $^{\circ}C$  should be treated as a single symbol, with no space between the  $^{\circ}$  and the letter C. (The symbol  $^{\circ}K$ , and the symbol  $^{\circ}$ , should no longer be used.)

## Units in Use Together with the SI

These units are not part of the SI, but it is recognized that they will continue to be used in appropriate contexts. SI prefixes may be attached to some of these units, such as milliliter, ml; millibar, mbar; mega-electronvolt, MeV; and kilotonne, kt.

Physical Quantity	Name of Unit	Symbol for Unit	Value in SI Units
Time	minute	min	60 s
Time	hour	h	3600 s
Time	day	d	86 400 s
Plane angle	degree	°	( $\pi/180$ ) rad
Plane angle	minute	'	( $\pi/10\,800$ ) rad
Plane angle	second	"	( $\pi/648\,000$ ) rad
Length	angstrom <sup>a</sup>	Å	10 <sup>-10</sup> m
Area	barn	b	10 <sup>-28</sup> m <sup>2</sup>
Volume	liter	l, L	dm <sup>3</sup> = 10 <sup>-3</sup> m <sup>3</sup>
Mass	tonne	t	Mg = 10 <sup>3</sup> kg
Pressure	bar <sup>a</sup>	bar	10 <sup>5</sup> Pa = 10 <sup>5</sup> N·m <sup>-2</sup>
Energy	electronvolt <sup>b</sup>	eV (= $e \times V$ )	$\approx 1.60218 \times 10^{-19}$ J
Mass	unified atomic mass unit <sup>b,c</sup>	u (= $m_a(^{12}\text{C})/12$ )	$\approx 1.66054 \times 10^{-27}$ kg

<sup>a</sup>The angstrom and the bar are approved by CIPM for “temporary use with SI units,” until CIPM makes a further recommendation. However, they should not be introduced where they are not used at present.

<sup>b</sup>The values of these units in terms of the corresponding SI units are not exact, since they depend on the values of the physical constants  $e$  (for the electronvolt) and  $N_A$  (for the unified atomic mass unit), which are determined by experiment.

<sup>c</sup>The unified atomic mass unit is also sometimes called the dalton, with symbol Da, although the name and symbol have not been approved by CGPM.

## A.3 Conversion Constants and Multipliers

### Recommended Decimal Multiples and Submultiples

Multiple or Submultiple	Prefix	Symbol	Multiple or Submultiple	Prefix	Symbol
10 <sup>18</sup>	exa	E	10 <sup>-1</sup>	deci	d
10 <sup>15</sup>	peta	P	10 <sup>-2</sup>	centi	c
10 <sup>12</sup>	tera	T	10 <sup>-3</sup>	milli	m
10 <sup>9</sup>	giga	G	10 <sup>-6</sup>	micro	μ (Greek mu)
10 <sup>6</sup>	mega	M	10 <sup>-9</sup>	nano	n
10 <sup>3</sup>	kilo	k	10 <sup>-12</sup>	pico	p
10 <sup>2</sup>	hecto	h	10 <sup>-15</sup>	femto	f
10	deca	da	10 <sup>-18</sup>	atto	a

## Conversion Factors—Metric to English

To Obtain	Multiply	By
Inches	Centimeters	0.393 700 787 4
Feet	Meters	3.280 839 895
Yards	Meters	1.093 613 298
Miles	Kilometers	0.621 371 192 2
Ounces	Grams	$3.527\,396\,195 \times 10^{-2}$
Pounds	Kilograms	2.204 622 622
Gallons (U.S. liquid)	Liters	0.264 172 052 4
Fluid ounces	Milliliters (cc)	$3.381\,402\,270 \times 10^{-2}$
Square inches	Square centimeters	0.155 000 310 0
Square feet	Square meters	10.763 910 42
Square yards	Square meters	1.195 990 046
Cubic inches	Milliliters (cc)	$6.102\,374\,409 \times 10^{-2}$
Cubic feet	Cubic meters	35.314 666 72
Cubic yards	Cubic meters	1.307 950 619

## Conversion Factors—English to Metric

To Obtain	Multiply	By <sup>a</sup>
Microns	Mils	<b>25.4</b>
Centimeters	Inches	<b>2.54</b>
Meters	Feet	<b>0.304 8</b>
Meters	Yards	<b>0.914 4</b>
Kilometers	Miles	<b>1.609 344</b>
Grams	Ounces	28.349 523 13
Kilograms	Pounds	<b>0.453 592 37</b>
Liters	Gallons (U.S. liquid)	<b>3.785 411 784</b>
Millimeters (cc)	Fluid ounces	29.573 529 56
Square centimeters	Square inches	<b>6.451 6</b>
Square meters	Square feet	<b>0.092 903 04</b>
Square meters	Square yards	<b>0.836 127 36</b>
Milliliters (cc)	Cubic inches	<b>16.387 064</b>
Cubic meters	Cubic feet	$2.831\,684\,659 \times 10^{-2}$
Cubic meters	Cubic yards	0.764 554 858

<sup>a</sup> Boldface numbers are exact; others are given to ten significant figures where so indicated by the multiplier factor.

## Conversion Factors—General

To Obtain	Multiply	By <sup>a</sup>
Atmospheres	Feet of water @ 4°C	$2.950 \times 10^{-2}$
Atmospheres	Inches of mercury @ 0°C	$3.342 \times 10^{-2}$
Atmospheres	Pounds per square inch	$6.804 \times 10^{-2}$
Btu	Foot-pounds	$1.285 \times 10^{-3}$
Btu	Joules	$9.480 \times 10^{-4}$
Cubic feet	Cords	<b>128</b>
Degree (angle)	Radians	57.2958
Ergs	Foot-pounds	$1.356 \times 10^7$
Feet	Miles	<b>5280</b>

To Obtain	Multiply	By <sup>a</sup>
Feet of water @ 4°C	Atmospheres	33.90
Foot-pounds	Horsepower-hours	$1.98 \times 10^6$
Foot-pounds	Kilowatt-hours	$2.655 \times 10^6$
Foot-pounds per minute	Horsepower	$3.3 \times 10^4$
Horsepower	Foot-pounds per second	$1.818 \times 10^{-3}$
Inches of mercury @ 0°C	Pounds per square inch	2.036
Joules	Btu	1054.8
Joules	Foot-pounds	1.355 82
Kilowatts	Btu per minute	$1.758 \times 10^{-2}$
Kilowatts	Foot-pounds per minute	$2.26 \times 10^{-5}$
Kilowatts	Horsepower	0.745 712
Knots	Miles per hour	0.868 976 24
Miles	Feet	$1.894 \times 10^{-4}$
Nautical miles	Miles	0.868 976 24
Radians	Degrees	$1.745 \times 10^{-2}$
Square feet	Acres	<b>43 560</b>
Watts	Btu per minute	17.5796

<sup>a</sup> Boldface numbers are exact; others are given to ten significant figures where so indicated by the multiplier factor.

# Temperature Factors

$$^{\circ}\text{F} = 9/5(^{\circ}\text{C}) + 32$$

$$\text{Fahrenheit temperature} = 1.8(\text{temperature in kelvins}) - 459.67$$

$$^{\circ}\text{C} = 5/9(^{\circ}\text{F} - 32)$$

$$\text{Celsius temperature} = \text{temperature in kelvins} - 273.15$$

$$\text{Fahrenheit temperature} = 1.8(\text{Celsius temperature}) + 32$$

## Conversion of Temperatures

From	To		From	To	
Fahrenheit	Celsius	$t_C = \frac{t_F - 32}{1.8}$	Celsius	Fahrenheit	$t_F = (t_C \times 1.8) + 32$
	Kelvin	$T_K = \frac{t_F - 32}{1.8} + 273.15$		Kelvin	$T_K = t_C + 273.15$
	Rankine	$T_R = t_F + 459.67$		Rankine	$T_R = (t_C + 273.15) \times 1.8$
			Kelvin	Celsius	$t_C = T_K - 273.15$
				Rankine	$T_R = T_K \times 1.8$
			Rankine	Fahrenheit	$t_F = T_R - 459.67$
				Kelvin	$T_K = \frac{T_R}{1.8}$

## A.4 Physical Constants

### General

Equatorial radius of the earth = 6378:388 km = 3963:34 miles (statute)

Polar radius of the earth = 6356:912 km = 3949:99 miles (statute)

1 degree of latitude at 40° = 69 miles

1 international nautical mile = 1:150 78 miles (statute) = 1852 m = 6076:115 ft

Mean density of the earth = 5:522 g=cm<sup>3</sup> = 344:7 lb=ft<sup>3</sup>

Constant of gravitation (6:673 ± 0:003) × 10<sup>-8</sup> g=cm<sup>3</sup>g<sup>-1</sup> s<sup>2</sup>

Acceleration due to gravity at sea level, latitude 45° = 980:6194 cm=s<sup>2</sup> = 32:1726 ft=s<sup>2</sup>

Length of seconds pendulum at sea level, latitude 45° = 99:3575 cm = 39:1171 in.

1 knot (international) = 101:269 ft=min = 1:6878 ft=s = 1:1508 miles (statute)=h

1 micron = 10<sup>-4</sup> cm

1 angstrom = 10<sup>-8</sup> cm

Mass of hydrogen atom = (1:673 39 ± 0:0031) × 10<sup>-24</sup> g

Density of mercury at 0°C = 13:5955 g=mL

Density of water at 3:98°C = 1:000 000 g=mL

Density, maximum, of water, at 3:98°C = 0:999 973 g=cm<sup>3</sup>

Density of dry air at 0°C, 760 mm = 1:2929 g=L

Velocity of sound in dry air at 0°C = 331:36 m=s ; 1087:1 f t=s  
 Velocity of light in v acuum = (2:997 925 S 0:000 002) £ 10<sup>10</sup> cm=s  
 Heat of fusion of water, 0°C = 79:71 cal=g  
 Heat of vaporization of water, 100°C = 539:55 cal=g  
 Electrochemical equivalent of silver 0:001 118 g=s international amp  
 Absolute wavelength of red cadmium light in air at 15°C, 760 mm pressure = 6438:4696  
 Wavelength of orange-red line of krypton 86 = 6057:802

## ¼ Constants

$$\begin{aligned}\frac{1}{4} &= 3:14159265368979323846264338327950288419716939937511 \\ 1=\frac{1}{4} &= 0:31830988618379067153776752674502872406891929148091 \\ \frac{1}{4}^2 &= 9:8690440108935861883449099987615113531369940724079 \\ \log_e \frac{1}{4} &= 1:14472988584940017414342735135305871164729481291531 \\ \log_{10} \frac{1}{4} &= 0:49714987269413385435126828829089887365167832438044 \\ \log_{10} \frac{1}{2\frac{1}{4}} &= 0:39908993417905752489250359150769595020993410292128\end{aligned}$$

## Constants Involving e

$$\begin{aligned}e &= 2:71828182845904523536028747135266249775724709369996 \\ 1=e &= 0:36787944117144232159552377016146086744581113103177 \\ e^2 &= 7:38905609893065022723042746057500781318031557055185 \\ M = \log_{10} e &= 0:43429448190325182765112891891660508229439700580367 \\ 1=M = \log_e 10 &= 2:30258509299404568401799145468436420760110148862877 \\ \log_{10} M &= 9:637784311300536789122967498645 ; 10\end{aligned}$$

## Numerical Constants

$$\begin{aligned}\frac{1}{2} &= 1:41421356237309504880168872420969807856967187537695 \\ {}^3\frac{1}{2} &= 1:25992104989487:16476721060727822835057025146470151 \\ \log_e 2 &= 0:69314718055994530941723212145817656807550013436026 \\ \log_{10} 2 &= 0:30102999566398119521373889472449302678818988146211 \\ \frac{1}{3} &= 1:73205080756887729352744634150587236694280525381039 \\ {}^3\frac{1}{3} &= 1:442244947030740838232163831078010958839186925349935 \\ \log_e 3 &= 1:09861228866810969139524523692252570464749055782275 \\ \log_{10} 3 &= 0:47712125471966243729502790325511530920012886419070\end{aligned}$$

## A.5 Symbols and Terminology for Physical and Chemical Quantities

Name	Symbol	Definition	SI Unit
Classical Mechanics			
Mass	$m$		kg
Reduced mass	$\mu$	$\mu = m_1 m_2 / (m_1 + m_2)$	kg
Density, mass density	$\rho$	$\rho = m/V$	$\text{kg} \cdot \text{m}^{-3}$
Relative density	$d$	$d = \rho/\rho^\theta$	1
Surface density	$\rho_A, \rho_S$	$\rho_a = m/A$	$\text{kg} \cdot \text{m}^{-2}$
Specific volume	$v$	$v = V/m = 1/\rho$	$\text{m}^3 \cdot \text{kg}^{-1}$
Momentum	$\mathbf{p}$	$\mathbf{p} = m\mathbf{v}$	$\text{kg} \cdot \text{m} \cdot \text{s}^{-1}$
Angular momentum, action	$\mathbf{L}$	$\mathbf{L} = \mathbf{r} \times \mathbf{p}$	$\text{J} \cdot \text{s}$
Moment of inertia	$I, J$	$I = \sum m_i r_i^2$	$\text{kg} \cdot \text{m}^2$
Force	$\mathbf{F}$	$\mathbf{F} = d\mathbf{p}/dt = m\mathbf{a}$	N
Torque, moment of a force	$\mathbf{T}, (\mathbf{M})$	$\mathbf{T} = \mathbf{r} \times \mathbf{F}$	$\text{N} \cdot \text{m}$
Energy	$E$		J
Potential energy	$E_p, V, \Phi$	$E_p = -\int \mathbf{F} \cdot d\mathbf{s}$	J
Kinetic energy	$E_k, T, K$	$E_k = (1/2)mv^2$	J
Work	$W, w$	$W = \int \mathbf{F} \cdot d\mathbf{s}$	J
Hamilton function	$H$	$H(q, p) = T(q, p) + V(q)$	J
Lagrange function	$L$	$L(q, \dot{q}) = T(q, \dot{q}) - V(q)$	J
Pressure	$p, P$	$p = F/A$	$\text{Pa}, \text{N} \cdot \text{m}^{-2}$
Surface tension	$\gamma, \sigma$	$\gamma = dW/dA$	$\text{N} \cdot \text{m}^{-1}, \text{J} \cdot \text{m}^{-2}$
Weight	$G (W, P)$	$G = mg$	N
Gravitational constant	$G$	$F = Gm_1 m_2 / r^2$	$\text{N} \cdot \text{m}^2 \cdot \text{kg}^{-2}$
Normal stress	$\sigma$	$\sigma = F/A$	Pa
Shear stress	$\tau$	$\tau = F/A$	Pa
Linear strain, relative elongation	$\varepsilon, e$	$\varepsilon = \Delta l/l$	1
Modulus of elasticity, Young's modulus	$E$	$E = \sigma/\varepsilon$	Pa
Shear strain	$\gamma$	$\gamma = \Delta x/d$	1
Shear modulus	$G$	$G = \tau/\gamma$	Pa
Volume strain, bulk strain	$\theta$	$\theta = \Delta V/V_0$	1
Bulk modulus, compression modulus	$K$	$K = -V_0(dp/dV)$	Pa
Viscosity, dynamic viscosity	$\eta, \mu$	$\tau_{x,z} = \eta(dv_x/dz)$	$\text{Pa} \cdot \text{s}$
Fluidity	$\phi$	$\phi = 1/\eta$	$\text{m}^2 \cdot \text{kg}^{-1} \cdot \text{s}$
Kinematic viscosity	$\nu$	$\nu = \eta/\rho$	$\text{m}^2 \cdot \text{s}^{-1}$
Friction coefficient	$\mu, (f)$	$F_{\text{frict}} = \mu F_{\text{norm}}$	1
Power	$P$	$P = dW/dt$	W
Sound energy flux	$P, P_a$	$P = dE/dt$	W
Acoustic factors			
Reflection factor	$\rho$	$\rho = P_r/P_0$	1
Acoustic absorption factor	$\alpha_a, (\alpha)$	$\alpha_a = 1 - \rho$	1
Transmission factor	$\tau$	$\tau = P_{tr}/P_0$	1
Dissipation factor	$\delta$	$\delta = \alpha_a - \tau$	1

Name	Symbol	Definition	SI Unit
Electricity and Magnetism			
Quantity of electricity, electric charge	$Q$		C
Charge density	$\rho$	$\rho = Q/V$	$\text{C}\cdot\text{m}^{-3}$
Surface charge density	$\sigma$	$\sigma = Q/A$	$\text{C}\cdot\text{m}^{-2}$
Electric potential	$V, \phi$	$V = dW/dQ$	$\text{V}, \text{J}\cdot\text{C}^{-1}$
Electric potential difference	$U, \Delta V, \Delta\phi$	$U = V_2 - V_1$	V
Electromotive force	$E$	$E = \int (F/Q) \cdot ds$	V
Electric field strength	$\mathbf{E}$	$\mathbf{E} = \mathbf{F}/Q = -\text{grad } V$	$\text{V}\cdot\text{m}^{-1}$
Electric flux	$\Psi$	$\Psi = \int \mathbf{D} \cdot d\mathbf{A}$	C
Electric displacement	$\mathbf{D}$	$\mathbf{D} = \epsilon \mathbf{E}$	$\text{C}\cdot\text{m}^{-2}$
Capacitance	$C$	$C = Q/U$	$\text{F}, \text{C}\cdot\text{V}^{-1}$
Permittivity	$\epsilon$	$\mathbf{D} = \epsilon \mathbf{E}$	$\text{F}\cdot\text{m}^{-1}$
Permittivity of vacuum	$\epsilon_0$	$\epsilon_0 = \mu_0^{-1} c_0^{-2}$	$\text{F}\cdot\text{m}^{-1}$
Relative permittivity	$\epsilon_r$	$\epsilon_r = \epsilon/\epsilon_0$	1
Dielectric polarization (dipole moment per volume)	$\mathbf{P}$	$\mathbf{P} = \mathbf{D} - \epsilon_0 \mathbf{E}$	$\text{C}\cdot\text{m}^{-2}$
Electric susceptibility	$\chi_e$	$\chi_e = \epsilon_r - 1$	1
Electric dipole moment	$p, \mu$	$\mathbf{p} = Q\mathbf{r}$	$\text{C}\cdot\text{m}$
Electric current	$I$	$I = dQ/dt$	A
Electric current density	$\mathbf{j}, \mathbf{J}$	$I = \int \mathbf{j} \cdot d\mathbf{A}$	$\text{A}\cdot\text{m}^{-2}$
Magnetic flux density, magnetic induction	$\mathbf{B}$	$\mathbf{F} = Q\mathbf{v} \times \mathbf{B}$	T
Magnetic flux	$\Phi$	$\Phi = \int \mathbf{B} \cdot d\mathbf{A}$	Wb
Magnetic field strength	$\mathbf{H}$	$\mathbf{B} = \mu \mathbf{H}$	$\text{A}\cdot\text{m}^{-1}$
Permeability	$\mu$	$\mathbf{B} = \mu \mathbf{H}$	$\text{N}\cdot\text{A}^{-2}, \text{H}\cdot\text{m}^{-1}$
Permeability of vacuum	$\mu_0$		$\text{H}\cdot\text{m}^{-1}$
Relative permeability	$\mu_r$	$\mu_r = \mu/\mu_0$	1
Magnetization (magnetic dipole moment per volume)	$\mathbf{M}$	$\mathbf{M} = \mathbf{B}/\mu_0 - \mathbf{H}$	$\text{A}\cdot\text{m}^{-1}$
Magnetic susceptibility	$\chi, \kappa, (\chi_m)$	$\chi = \mu_r - 1$	1
Molar magnetic susceptibility	$\chi_m$	$\chi_m = V_m \chi$	$\text{m}^3 \cdot \text{mol}^{-1}$
Magnetic dipole moment	$\mathbf{m}, \mu$	$E_p = -\mathbf{m} \cdot \mathbf{B}$	$\text{A}\cdot\text{m}^2, \text{J}\cdot\text{T}^{-1}$
Electrical resistance	$R$	$R = U/I$	$\Omega$
Conductance	$G$	$G = 1/R$	S
Loss angle	$\delta$	$\delta = (\pi/2) + \phi_I - \phi_U$	1, rad
Reactance	$X$	$X = (U/I) \sin \delta$	$\Omega$
Impedance (complex impedance)	$Z$	$Z = R + iX$	$\Omega$
Admittance (complex admittance)	$Y$	$Y = 1/Z$	S
Susceptance	$B$	$Y = G + iB$	S
Resistivity	$\rho$	$\rho = E/j$	$\Omega \cdot \text{m}$
Conductivity	$\kappa, \gamma, \sigma$	$\kappa = 1/\rho$	$\text{S}\cdot\text{m}^{-1}$
Self-inductance	$L$	$E = -L(dI/dt)$	H
Mutual inductance	$M, L_{12}$	$E_1 = L_{12}(dI_2/dt)$	H
Magnetic vector potential	$\mathbf{A}$	$\mathbf{B} = \nabla \times \mathbf{A}$	$\text{Wb}\cdot\text{m}^{-1}$
Poynting vector	$\mathbf{S}$	$\mathbf{S} = \mathbf{E} \times \mathbf{H}$	$\text{W}\cdot\text{m}^{-2}$



Name	Symbol	Definition	SI Unit
Electromagnetic Radiation			
Wavelength	$\lambda$		m
Speed of light			
In vacuum	$c_0$		$\text{m} \cdot \text{s}^{-1}$
In a medium	$c$	$c = c_0/n$	$\text{m} \cdot \text{s}^{-1}$
Wavenumber in vacuum	$\tilde{\nu}$	$\tilde{\nu} = \nu/c_0 = 1/n\lambda$	$\text{m}^{-1}$
Wavenumber (in a medium)	$\sigma$	$\sigma = 1/\lambda$	$\text{m}^{-1}$
Frequency	$\nu$	$\nu = c/\lambda$	Hz
Circular frequency, pulsance	$\omega$	$\omega = 2\pi\nu$	$\text{s}^{-1}, \text{rad} \cdot \text{s}^{-1}$
Refractive index	$n$	$n = c_0/c$	1
Planck constant	$h$		$\text{J} \cdot \text{s}$
Planck constant/ $2\pi$	$\hbar$	$\hbar = h/2\pi$	$\text{J} \cdot \text{s}$
Radiant energy	$Q, W$		J
Radiant energy density	$\rho, w$	$\rho = Q/V$	$\text{J} \cdot \text{m}^{-3}$
Spectral radiant energy density			
In terms of frequency	$\rho_\nu, w_\nu$	$\rho_\nu = d\rho/d\nu$	$\text{J} \cdot \text{m}^{-3} \cdot \text{Hz}^{-1}$
In terms of wavenumber	$\rho_{\tilde{\nu}}, w_{\tilde{\nu}}$	$\rho_{\tilde{\nu}} = d\rho/d\tilde{\nu}$	$\text{J} \cdot \text{m}^{-2}$
In terms of wavelength	$\rho_\lambda, w_\lambda$	$\rho_\lambda = d\rho/d\lambda$	$\text{J} \cdot \text{m}^{-4}$
Einstein transition probabilities			
Spontaneous emission	$A_{nm}$	$dN_n/dt = -A_{nm}N_n$	$\text{s}^{-1}$
Stimulated emission	$B_{nm}$	$dN_n/dt = -\rho_{\tilde{\nu}}(\tilde{\nu}_{nm}) \times B_{nm}N_n$	$\text{s} \cdot \text{kg}^{-1}$
Stimulated absorption	$B_{mn}$	$dN_n/dt = \rho_{\tilde{\nu}}(\tilde{\nu}_{nm})B_{mn}N_m$	$\text{s} \cdot \text{kg}^{-1}$
Radiant power, radiant energy per time	$\Phi, P$	$\Phi = dQ/dt$	W
Radiant intensity	$I$	$I = d\Phi/d\Omega$	$\text{W} \cdot \text{sr}^{-1}$
Radiant exitance (emitted radiant flux)	$M$	$M = d\Phi/dA_{\text{source}}$	$\text{W} \cdot \text{m}^{-2}$
Irradiance (radiant flux received)	$E, (I)$	$E = d\Phi/dA$	$\text{W} \cdot \text{m}^{-2}$
Emittance	$\varepsilon$	$\varepsilon = M/M_{\text{bb}}$	1
Stefan–Boltzmann constant	$\sigma$	$M_{\text{bb}} = \sigma T^4$	$\text{W} \cdot \text{m}^{-2} \cdot \text{K}^{-4}$
First radiation constant	$c_1$	$c_1 = 2\pi h c_0^2$	$\text{W} \cdot \text{m}^2$
Second radiation constant	$c_2$	$c_2 = hc_0/k$	$\text{K} \cdot \text{m}$
Transmittance, transmission factor	$\tau, T$	$\tau = \Phi_{\text{tr}}/\Phi_0$	1
Absorptance, absorption factor	$\alpha$	$\alpha = \Phi_{\text{abs}}/\Phi_0$	1
Reflectance, reflection factor	$\rho$	$\rho = \Phi_{\text{refl}}/\Phi_0$	1
(Decadic) absorbance	$A$	$A = \lg(1 - \alpha_i)$	1
Napierian absorbance	$B$	$B = \ln(1 - \alpha_i)$	1
Absorption coefficient			
(Linear) decadic	$a, K$	$a = A/l$	$\text{m}^{-1}$
(Linear) napierian	$\alpha$	$\alpha = B/l$	$\text{m}^{-1}$
Molar (decadic)	$\varepsilon$	$\varepsilon = a/c = A/cl$	$\text{m}^2 \cdot \text{mol}^{-1}$
Molar napierian	$\kappa$	$\kappa = \alpha/c = B/cl$	$\text{m}^2 \cdot \text{mol}^{-1}$
Absorption index	$k$	$k = \alpha/4\pi\tilde{\nu}$	1
Complex refractive index	$\hat{n}$	$\hat{n} = n + ik$	1
Molar refraction	$R, R_m$	$R = \frac{(n^2 - 1)}{(n^2 + 2)} V_m$	$\text{m}^3 \cdot \text{mol}^{-1}$
Angle of optical rotation	$\alpha$		1, rad

Name	Symbol	Definition	SI Unit
Solid State			
Lattice vector	$\mathbf{R}, \mathbf{R}_0$		m
Fundamental translation vectors for the crystal lattice	$\mathbf{a}_1; \mathbf{a}_2; \mathbf{a}_3,$ $\mathbf{a}; \mathbf{b}; \mathbf{c}$	$\mathbf{R} = n_1 \mathbf{a}_1 + n_2 \mathbf{a}_2 + n_3 \mathbf{a}_3$	m
(Circular) reciprocal lattice vector	$\mathbf{G}$	$\mathbf{G} \cdot \mathbf{R} = 2\pi m$	$\text{m}^{-1}$
(Circular) fundamental translation vectors for the reciprocal lattice	$\mathbf{b}_1; \mathbf{b}_2; \mathbf{b}_3,$ $\mathbf{a}^*; \mathbf{b}^*; \mathbf{c}^*$	$\mathbf{a}_i \cdot \mathbf{b}_k = 2\pi \delta_{ik}$	$\text{m}^{-1}$
Lattice plane spacing	$d$		m
Bragg angle	$\theta$	$n\lambda = 2d \sin \theta$	1, rad
Order of reflection	$n$		1
Order parameters			
Short range	$\sigma$		1
Long range	$s$		1
Burgers vector	$\mathbf{b}$		m
Particle position vector	$\mathbf{r}, \mathbf{R}_j$		m
Equilibrium position vector of an ion	$\mathbf{R}_0$		m
Displacement vector of an ion	$\mathbf{u}$	$\mathbf{u} = \mathbf{R} - \mathbf{R}_0$	m
Debye-Waller factor	$B, D$		1
Debye circular wavenumber	$q_D$		$\text{m}^{-1}$
Debye circular frequency	$\omega_D$		$\text{s}^{-1}$
Grüneisen parameter	$\gamma, \Gamma$	$\gamma = \alpha V / \kappa C_V$	1
Madelung constant	$\alpha, \mathcal{M}$	$E_{\text{coul}} = \frac{\alpha N_A z + z - e^2}{4\pi\epsilon_0 R_0}$	1
Density of states	$N_E$	$N_E = dN(E)/dE$	$\text{J}^{-1} \cdot \text{m}^{-3}$
(Spectral) density of vibrational modes	$N_\omega, g$	$N_\omega = dN(\omega)/d\omega$	$\text{s} \cdot \text{m}^{-3}$
Resistivity tensor	$\rho_{ik}$	$\mathbf{E} = \rho \cdot \mathbf{j}$	$\Omega \cdot \text{m}$
Conductivity tensor	$\sigma_{ik}$	$\sigma = \rho^{-1}$	$\text{S} \cdot \text{m}^{-1}$
Thermal conductivity tensor	$\lambda_{ik}$	$\mathbf{J}_q = -\lambda \cdot \text{grad } T$	$\text{W} \cdot \text{m}^{-1} \cdot \text{K}^{-1}$
Residual resistivity	$\rho_R$		$\Omega \cdot \text{m}$
Relaxation time	$\tau$	$\tau = l/v_F$	s
Lorenz coefficient	$L$	$L = \lambda/\sigma T$	$\text{V}^2 \cdot \text{K}^{-2}$
Hall coefficient	$A_H, R_H$	$\mathbf{E} = \rho \cdot \mathbf{j} + R_H (\mathbf{B} \times \mathbf{j})$	$\text{m}^3 \cdot \text{C}^{-1}$
Thermoelectric force	$E$		V
Peltier coefficient	$\Pi$		V
Thomson coefficient	$\mu, (\tau)$		$\text{V} \cdot \text{K}^{-1}$
Work function	$\Phi$	$\Phi = E_\infty - E_F$	J
Number density, number concentration	$n, (p)$		$\text{m}^{-3}$
Gap energy	$E_g$		J
Donor ionization energy	$E_d$		J
Acceptor ionization energy	$E_a$		J
Fermi energy	$E_F, \epsilon_F$		J
Circular wave vector, propagation vector	$\mathbf{k}, \mathbf{q}$	$\mathbf{k} = 2\pi/\lambda$	$\text{m}^{-1}$
Bloch function	$u_i(\mathbf{r})$	$\psi(\mathbf{r}) = u_i(\mathbf{r}) \exp(i\mathbf{k} \cdot \mathbf{r})$	$\text{m}^{-3/2}$
Charge density of electrons	$\rho$	$\rho(\mathbf{r}) = -e\psi^*(\mathbf{r})\psi(\mathbf{r})$	$\text{C} \cdot \text{m}^{-3}$
Effective mass	$m^*$		kg
Mobility	$\mu$	$\mu = v_{\text{drift}}/E$	$\text{m}^2 \cdot \text{V}^{-1} \cdot \text{s}^{-1}$
Mobility ratio	$b$	$b = \mu_n/\mu_p$	1
Diffusion coefficient	$D$	$dN/dt = -DA(dn/dx)$	$\text{m}^2 \cdot \text{s}^{-1}$
Diffusion length	$L$	$L = \sqrt{D\tau}$	m
Characteristic (Weiss) temperature	$\phi, \phi_w$		K
Curie temperature	$T_C$		K
Néel temperature	$T_N$		K

## A.6 Elementary Algebra and Geometry

---

### Fundamental Properties (Real Numbers)

$a + b = b + a$	Commutative law for addition
$(a + b) + c = a + (b + c)$	Associative law for addition
$a + 0 = 0 + a$	Identity law for addition
$a + (-a) = (-a) + a = 0$	Inverse law for addition
$a(bc) = (ab)c$	Associative law for multiplication
$a\left(\frac{1}{a}\right) = \left(\frac{1}{a}\right)a = 1, \quad a \neq 0$	Inverse law for multiplication
$(a)(1) = (1)(a) = a$	Identity law for multiplication
$ab = ba$	Commutative law for multiplication
$a(b + c) = ab + ac$	Distributive law

Division by zero is not defined.

### Exponents

For integers  $m$  and  $n$ ,

$$\begin{aligned}a^n a^m &= a^{n+m} \\a^n a^m &= a^{n+m} \\(a^n)^m &= a^{nm} \\(ab)^m &= a^m b^m \\(a=b)^m &= a^m = b^m\end{aligned}$$

### Fractional Exponents

$$a^{p/q} = (a^{1/q})^p$$

where  $a^{1/q}$  is the positive  $q$ th root of  $a$  if  $a > 0$  and the negative  $q$ th root of  $a$  if  $a$  is

negative and  $q$  is odd. Accordingly, the five rules of exponents given above (for integers) are also valid if  $m$  and  $n$  are fractions, provided  $a$  and  $b$  are positive.

## Irrational Exponents

If an exponent is irrational (e.g.,  $\sqrt{2}$ ), the quantity, such as  $a^{\sqrt{2}}$ , is the limit of the sequence  $a^{1.4}; a^{1.41}; a^{1.414}; \dots$

## Operations with Zero

$$0^m = 0 \quad a^0 = 1$$

## Logarithms

If  $x$ ,  $y$ , and  $b$  are positive and  $b \neq 1$ ,

$$\log_b(xy) = \log_b x + \log_b y$$

$$\log_b(x/y) = \log_b x - \log_b y$$

$$\log_b x^p = p \log_b x$$

$$\log_b(1/x) = -\log_b x$$

$$\log_b b = 1$$

$$\log_b 1 = 0 \quad \text{Note: } b^{\log_b x} = x$$

## Change of Base ( $a \neq 1$ )

$$\log_b x = \log_a x \log_b a$$

## Factorials

The factorial of a positive integer  $n$  is the product of all the positive integers less than or equal to the integer  $n$  and is denoted  $n!$ . Thus,

$$n! = 1 \cdot 2 \cdot 3 \cdot \dots \cdot n$$

Factorial 0 is defined:  $0! = 1$ .

## Stirling's Approximation

$$\lim_{n \rightarrow \infty} \frac{(n/e)^n \sqrt{2\pi n}}{n!} = 1$$

## Binomial Theorem

For positive integer  $n$

$$(x + y)^n = x^n + nx^{n-1}y + \frac{n(n-1)}{2!}x^{n-2}y^2 + \frac{n(n-1)(n-2)}{3!}x^{n-3}y^3 + \dots + nxy^{n-1} + y^n$$

## Factors and Expansion

$$(a + b)^2 = a^2 + 2ab + b^2$$

$$(a - b)^2 = a^2 - 2ab + b^2$$

$$(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$$

$$(a - b)^3 = a^3 - 3a^2b + 3ab^2 - b^3$$

$$(a^2 - b^2) = (a - b)(a + b)$$

$$(a^3 - b^3) = (a - b)(a^2 + ab + b^2)$$

$$(a^3 + b^3) = (a + b)(a^2 - ab + b^2)$$

## Progression

An *arithmetic progression* is a sequence in which the difference between any term and the preceding term is a constant ( $d$ ):

$$a; a + d; a + 2d; \dots; a + (n - 1)d$$

If the last term is denoted  $l [= a + (n - 1)d]$ , then the sum is

$$s = \frac{n}{2}(a + l)$$

A *geometric progression* is a sequence in which the ratio of any term to the preceding term is a constant  $r$ . Thus, for  $n$  terms,

$$a; ar; ar^2; \dots; ar^{n-1}$$

The sum is

$$S = \frac{a - ar^n}{1 - r}$$

## Complex Numbers

A complex number is an ordered pair of real numbers  $(a; b)$ .

*Equality:*  $(a; b) = (c; d)$  if and only if  $a = c$  and  $b = d$

*Addition:*  $(a; b) + (c; d) = (a + c; b + d)$

*Multiplication:*  $(a; b)(c; d) = (ac - bd; ad + bc)$

The first element  $(a; b)$  is called the *real* part, the second the *imaginary* part. An alternative notation for  $(a; b)$  is  $a + bi$ , where  $i^2 = (-1; 0)$ , and  $i = (0; 1)$  or  $0 + 1i$  is written for this complex number as a convenience. With this understanding,  $i$  behaves as a number, that is,  $(2 - 3i)(4 + i) = 8 - 12i + 2i - 3i^2 = 11 - 10i$ . The conjugate of  $a + bi$  is  $a - bi$ , and the product of a complex number and its conjugate is  $a^2 + b^2$ . Thus, *quotients* are computed by multiplying numerator and denominator by the conjugate of the denominator, as illustrated below:

$$\frac{2 + 3i}{4 + 2i} = \frac{(4 - 2i)(2 + 3i)}{(4 - 2i)(4 + 2i)} = \frac{14 + 8i}{20} = \frac{7 + 4i}{10}$$

## Polar Form

The complex number  $x + iy$  may be represented by a plane vector with components  $x$  and  $y$ :

$$x + iy = r(\cos \mu + i \sin \mu)$$

(See [Fig. A.1.](#)) Then, given two complex numbers  $z_1 = r_1(\cos \mu_1 + i \sin \mu_1)$  and  $z_2 = r_2(\cos \mu_2 + i \sin \mu_2)$ , the product and quotient are:

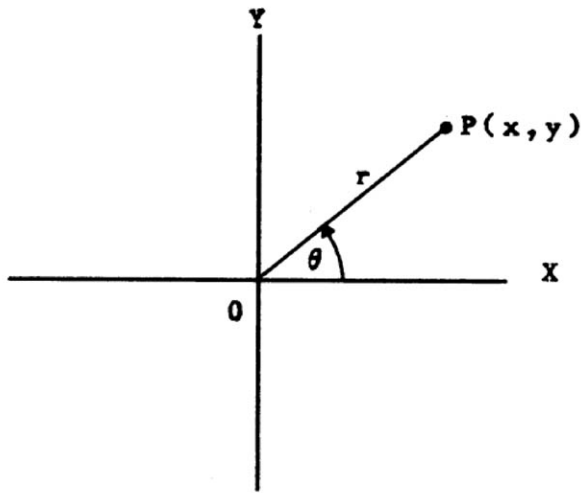
**Product:**  $z_1 z_2 = r_1 r_2 [\cos(\mu_1 + \mu_2) + i \sin(\mu_1 + \mu_2)]$

**Quotient:**  $z_1 / z_2 = (r_1 / r_2) [\cos(\mu_1 - \mu_2) + i \sin(\mu_1 - \mu_2)]$

**Powers:**  $z^n = [r(\cos \mu + i \sin \mu)]^n = r^n [\cos n\mu + i \sin n\mu]$

**Roots:**  $z^{1/n} = [r(\cos \mu + i \sin \mu)]^{1/n}$   
 $= r^{1/n} \left[ \cos \frac{\mu + k \cdot 360^\circ}{n} + i \sin \frac{\mu + k \cdot 360^\circ}{n} \right];$   
 $k = 0; 1; 2; \dots; n - 1$

**Figure A.1** Polar form of complex number.



## Permutations

A permutation is an ordered arrangement (sequence) of all or part of a set of objects. The number of permutations of  $n$  objects taken  $r$  at a time is

$$p(n; r) = n(n - 1)(n - 2) \cdots (n - r + 1)$$

$$= \frac{n!}{(n - r)!}$$

A permutation of positive integers is "even" or "odd" if the total number of inversions is an even integer or an odd integer, respectively. Inversions are counted relative to each integer  $j$  in the permutation by counting the number of integers that follow  $j$  and are less than  $j$ . These

are summed to give the total number of inversions. For example, the permutation 4132 has four inversions: three relative to 4 and one relative to 3. This permutation is therefore even.

## Combinations

A combination is a selection of one or more objects from among a set of objects regardless of order. The number of combinations of  $n$  different objects taken  $r$  at a time is

$$C(n; r) = \frac{P(n; r)}{r!} = \frac{n!}{r!(n-r)!}$$

## Algebraic Equations

### Quadratic

If  $ax^2 + bx + c = 0$ , and  $a \neq 0$ , then roots are

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

### Cubic

To solve  $x^3 + bx^2 + cx + d = 0$ , let  $x = y - \frac{b}{3}$ . Then the *reduced cubic* is obtained:

$$y^3 + py + q = 0$$

where  $p = c - \frac{1}{3}b^2$  and  $q = d - \frac{1}{3}bc + \frac{2}{27}b^3$ . Solutions of the original cubic are then in terms of the reduced cubic roots  $y_1, y_2, y_3$ :

$$x_1 = y_1 - \frac{1}{3}b \quad x_2 = y_2 - \frac{1}{3}b \quad x_3 = y_3 - \frac{1}{3}b$$

The three roots of the reduced cubic are

$$y_1 = (A)^{1/3} + (B)^{1/3}$$

$$y_2 = W(A)^{1/3} + W^2(B)^{1/3}$$

$$y_3 = W^2(A)^{1/3} + W(B)^{1/3}$$

where



$$A = \sqrt[3]{\frac{1}{2}q + \sqrt{\frac{1}{4}q^2 + (1=27)p^3}}$$

$$B = \sqrt[3]{\frac{1}{2}q - \sqrt{\frac{1}{4}q^2 + (1=27)p^3}}$$

$$W = \frac{1 + i\sqrt[3]{3}}{2}; \quad W^2 = \frac{1 - i\sqrt[3]{3}}{2}$$

When  $(1=27)p^3 + (1=4)q^2$  is negative,  $A$  is complex; in this case  $A$  should be expressed in trigonometric form:  $A = r(\cos \mu + i \sin \mu)$  where  $\mu$  is a first or second quadrant angle, as  $q$  is negative or positive. The three roots of the reduced cubic are

$$y_1 = 2(r)^{1=3} \cos(\mu=3)$$

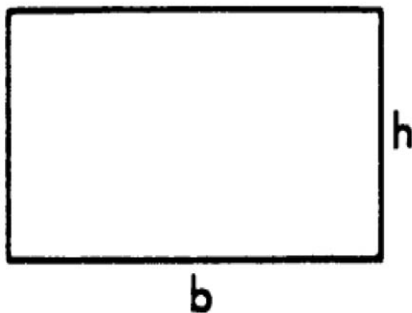
$$y_2 = 2(r)^{1=3} \cos \frac{\mu}{3} + 120^\circ$$

$$y_3 = 2(r)^{1=3} \cos \frac{\mu}{3} + 240^\circ$$

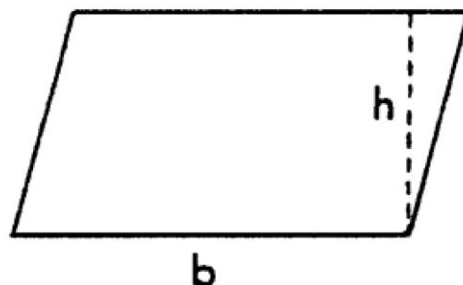
## Geometry

Figures A.2 to A.12 are a collection of common geometric figures. Area ( $A$ ), volume ( $V$ ), and other measurable features are indicated.

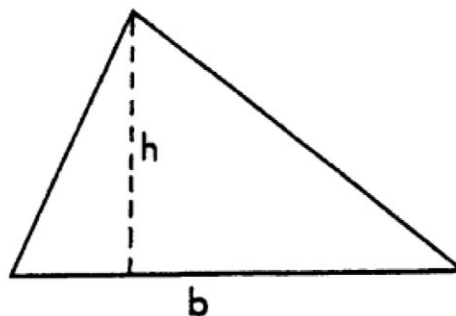
**Figure A.2** Rectangle.  $A = bh$ .



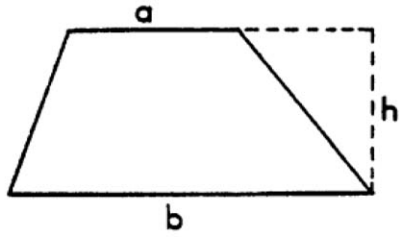
**Figure A.3** Parallelogram.  $A = bh$ .



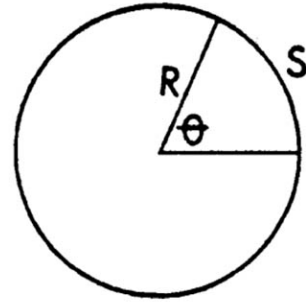
**Figure A.4** Triangle.  $A = \frac{1}{2}bh$ .



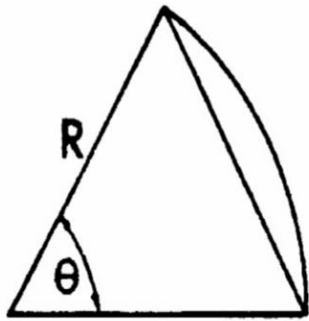
**Figure A.5** Trapezoid.  $A = \frac{1}{2}(a + b)h$ .



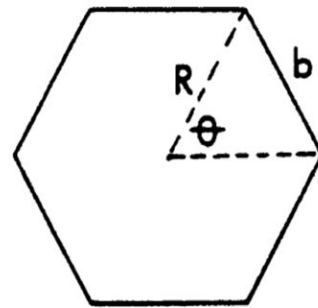
**Figure A.6** Circle.  $A = \frac{1}{4}R^2$ .



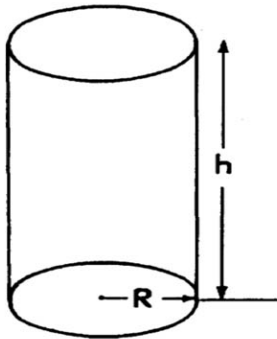
**Figure A.7** Sector of circle.  $A_{\text{sector}} = \frac{1}{2}R^2(\mu \text{ ; } \sin \mu)$ .



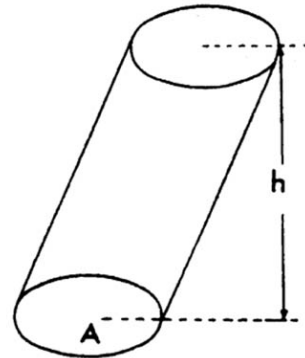
**Figure A.8** Regular polygon of  $n$  sides.  $A = (n-4)b^2 \cot(\frac{1}{2}\pi/n)$ ;  $R = (b/2) \csc(\frac{1}{2}\pi/n)$ .



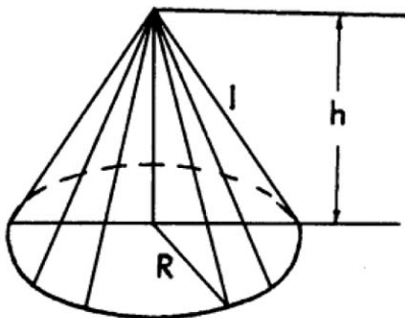
**Figure A.9** Right circular cylinder.  $V = \pi R^2 h$ ; lateral surface area  $= 2\pi R h$ .



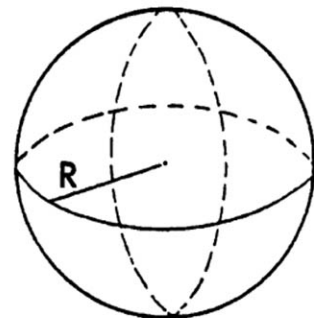
**Figure A.10** Cylinder (or prism) with parallel bases.  $V = Ah$ .



**Figure A.11** Right circular cone.  $V = \frac{1}{3}\pi R^2 h$ ; lateral surface area  $= \pi R l = \pi R \sqrt{R^2 + h^2}$ .



**Figure A.12** Sphere.  $V = \frac{4}{3}\pi R^3$ ; surface area  $= 4\pi R^2$ .



“Appendix: Mathematical Tables and Formulae”  
*The Engineering Handbook.*  
Ed. Richard C. Dorf  
Boca Raton: CRC Press LLC, 2000

## A.7 Table of Derivatives

In the following table,  $a$  and  $n$  are constants,  $e$  is the base of the natural logarithms, and  $u$  and  $v$  denote functions of  $x$ .

1. $\frac{d}{dx}(a) = 0$	20. $\frac{d}{dx} \cos^{-1} u = \frac{-1}{\sqrt{1-u^2}} \frac{du}{dx}, \quad (0 \leq \cos^{-1} u \leq \pi)$
2. $\frac{d}{dx}(x) = 1$	21. $\frac{d}{dx} \tan^{-1} u = \frac{1}{1+u^2} \frac{du}{dx}$
3. $\frac{d}{dx}(au) = a \frac{du}{dx}$	22. $\frac{d}{dx} \operatorname{ctn}^{-1} u = \frac{-1}{1+u^2} \frac{du}{dx}$
4. $\frac{d}{dx}(u+v) = \frac{du}{dx} + \frac{dv}{dx}$	23. $\frac{d}{dx} \sec^{-1} u = \frac{1}{u\sqrt{u^2-1}} \frac{du}{dx},$ $(-\pi \leq \sec^{-1} u < -\frac{1}{2}\pi; \quad 0 \leq \sec^{-1} u < \frac{1}{2}\pi)$
5. $\frac{d}{dx}(uv) = u \frac{dv}{dx} + v \frac{du}{dx}$	24. $\frac{d}{dx} \csc^{-1} u = \frac{-1}{u\sqrt{u^2-1}} \frac{du}{dx},$ $(-\pi < \csc^{-1} u \leq -\frac{1}{2}\pi; \quad 0 < \csc^{-1} u \leq \frac{1}{2}\pi)$
6. $\frac{d}{dx}(u/v) = \frac{v \frac{du}{dx} - u \frac{dv}{dx}}{v^2}$	25. $\frac{d}{dx} \sinh u = \cosh u \frac{du}{dx}$
7. $\frac{d}{dx}(u^n) = nu^{n-1} \frac{du}{dx}$	26. $\frac{d}{dx} \cosh u = \sinh u \frac{du}{dx}$
8. $\frac{d}{dx} e^u = e^u \frac{du}{dx}$	27. $\frac{d}{dx} \tanh u = \operatorname{sech}^2 u \frac{du}{dx}$
9. $\frac{d}{dx} a^u = (\log_e a) a^u \frac{du}{dx}$	28. $\frac{d}{dx} \operatorname{ctnh} u = -\operatorname{csch}^2 u \frac{du}{dx}$
10. $\frac{d}{dx} \log_e u = (1/u) \frac{du}{dx}$	29. $\frac{d}{dx} \operatorname{sech} u = -\operatorname{sech} u \tanh u \frac{du}{dx}$
11. $\frac{d}{dx} \log_a u = (\log_a e)(1/u) \frac{du}{dx}$	30. $\frac{d}{dx} \operatorname{csch} u = -\operatorname{csch} u \operatorname{ctnh} u \frac{du}{dx}$
12. $\frac{d}{dx} u^v = v u^{v-1} \frac{du}{dx} + u^v (\log_e u) \frac{dv}{dx}$	31. $\frac{d}{dx} \sinh^{-1} u = \frac{1}{\sqrt{u^2+1}} \frac{du}{dx}$
13. $\frac{d}{dx} \sin u = \cos u \frac{du}{dx}$	32. $\frac{d}{dx} \cosh^{-1} u = \frac{1}{\sqrt{u^2-1}} \frac{du}{dx}$
14. $\frac{d}{dx} \cos u = -\sin u \frac{du}{dx}$	33. $\frac{d}{dx} \tanh^{-1} u = \frac{1}{1-u^2} \frac{du}{dx}$
15. $\frac{d}{dx} \tan u = \sec^2 u \frac{du}{dx}$	34. $\frac{d}{dx} \operatorname{ctnh}^{-1} u = \frac{-1}{u^2-1} \frac{du}{dx}$
16. $\frac{d}{dx} \operatorname{ctn} u = -\operatorname{csc}^2 u \frac{du}{dx}$	35. $\frac{d}{dx} \operatorname{sech}^{-1} u = \frac{-1}{u\sqrt{1-u^2}} \frac{du}{dx}$
17. $\frac{d}{dx} \sec u = \sec u \tan u \frac{du}{dx}$	36. $\frac{d}{dx} \operatorname{csch}^{-1} u = \frac{-1}{u\sqrt{u^2+1}} \frac{du}{dx}$
18. $\frac{d}{dx} \csc u = -\csc u \operatorname{ctn} u \frac{du}{dx}$	
19. $\frac{d}{dx} \sin^{-1} u = \frac{1}{\sqrt{1-u^2}} \frac{du}{dx}, \quad (-\frac{1}{2}\pi \leq \sin^{-1} u \leq \frac{1}{2}\pi)$	

## Additional Relations with Derivatives

$$\frac{d}{dt} \int_a^t f(x) dx = f(t) \quad \frac{d}{dt} \int_t^a f(x) dx = -f(t)$$

$$\text{If } x = f(y), \text{ then } \frac{dy}{dx} = \frac{1}{\frac{dx}{dy}}$$

$$\text{If } y = f(u) \text{ and } u = g(x), \text{ then } \frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx} \text{ (chain rule)}$$

$$\text{If } x = f(t) \text{ and } y = g(t), \text{ then } \frac{dy}{dx} = \frac{g'(t)}{f'(t)}, \text{ and } \frac{d^2y}{dx^2} = \frac{f'(t)g''(t) - g'(t)f''(t)}{[f'(t)]^3}$$

(Note: Exponent in denominator is 3.)

## A.8 Integrals

---

### Elementary Forms

1.  $\int a dx = ax$
2.  $\int a f(x) dx = a \int f(x) dx$
3.  $\int A(y) dx = \int \frac{A(y)}{y'} dy$ ; where  $y' = \frac{dy}{dx}$
4.  $\int (u + v) dx = \int u dx + \int v dx$ , where  $u$  and  $v$  are any functions of  $x$
5.  $\int u dv = uv - \int v du$
6.  $\int u \frac{dv}{dx} dx = uv - \int v \frac{du}{dx} dx$
7.  $\int x^n dx = \frac{x^{n+1}}{n+1}$ , except  $n = -1$
8.  $\int \frac{f'(x) dx}{f(x)} = \log f(x)$ ; [ $df(x) = f'(x) dx$ ]
9.  $\int \frac{dx}{x} = \log x$
10.  $\int \frac{f'(x) dx}{f^2(x)} = -\frac{1}{f(x)}$ ; [ $df(x) = f'(x) dx$ ]
11.  $\int e^x dx = e^x$

$$\begin{aligned}
12. \quad \int e^{ax} dx &= e^{ax} = a \\
13. \quad \int b^{ax} dx &= \frac{b^{ax}}{a \log b}; \quad (b > 0) \\
14. \quad \int \log x dx &= x \log x - x \\
15. \quad \int a^x \log a dx &= a^x; \quad (a > 0) \\
16. \quad \int \frac{dx}{a^2 + x^2} &= \frac{1}{a} \tan^{-1} \frac{x}{a} \\
17. \quad \int \frac{dx}{a^2 - x^2} &= \frac{1}{2a} \log \frac{a+x}{a-x}; \quad (a^2 > x^2) \\
18. \quad \int \frac{dx}{x^2 - a^2} &= \frac{1}{2a} \log \frac{x-a}{x+a}; \quad (x^2 > a^2) \\
19. \quad \int \frac{dx}{a^2 + x^2} &= \frac{1}{a} \tan^{-1} \frac{x}{a} \\
20. \quad \int \frac{dx}{x^2 - a^2} &= \log \left( x + \sqrt{x^2 - a^2} \right) \\
21. \quad \int \frac{dx}{x \sqrt{x^2 + a^2}} &= \frac{1}{a} \sec^{-1} \frac{x}{a} \\
22. \quad \int \frac{dx}{x \sqrt{a^2 - x^2}} &= -\frac{1}{a} \log \frac{a + \sqrt{a^2 - x^2}}{x}
\end{aligned}$$

## Forms Containing (a + bx)

For forms containing a + bx, but not listed in the table, the substitution u = (a + bx) = x may prove helpful.

$$23. \quad \int (a + bx)^n dx = \frac{(a + bx)^{n+1}}{(n+1)b}; \quad (n \neq -1)$$

$$24. \int x(a + bx)^n dx = \frac{1}{b^2(n+2)} (a + bx)^{n+2} - \frac{a}{b^2(n+1)} (a + bx)^{n+1}; \quad (n \neq -1; -2)$$

$$25. \int x^2(a + bx)^n dx = \frac{1}{b^3} \frac{(a + bx)^{n+3}}{n+3} - 2a \frac{(a + bx)^{n+2}}{n+2} + a^2 \frac{(a + bx)^{n+1}}{n+1} \\ - \frac{x^{m+1}(a + bx)^n}{m+n+1} + \frac{an}{m+n+1} \int x^m(a + bx)^{n-1} dx$$

$$26. \int x^m(a + bx)^n dx = \frac{1}{a(n+1)} \int x^{m+1}(a + bx)^{n+1} dx + (m+n+2) \int x^m(a + bx)^{n+1} dx \\ \text{or} \\ \frac{1}{b(m+n+1)} \int x^m(a + bx)^{n+1} dx - \frac{a}{b} \int x^{m-1}(a + bx)^n dx$$

$$27. \int \frac{dx}{a + bx} = \frac{1}{b} \log(a + bx)$$

$$28. \int \frac{dx}{(a + bx)^2} = -\frac{1}{b(a + bx)}$$

$$29. \int \frac{dx}{(a + bx)^3} = -\frac{1}{2b(a + bx)^2}$$

$$30. \int \frac{x dx}{a + bx} = \frac{1}{b^2} [a + bx - a \log(a + bx)]$$

$$\text{or} \\ \frac{x}{b} - \frac{a}{b^2} \log(a + bx)$$

$$31. \int \frac{x dx}{(a + bx)^2} = \frac{1}{b^2} \log(a + bx) + \frac{a}{a + bx}$$

$$32. \int \frac{x dx}{(a + bx)^n} = \frac{1}{b^2} \frac{1}{(n-2)(a + bx)^{n-2}} + \frac{a}{(n-1)(a + bx)^{n-1}}; \quad n \neq 1; 2$$

$$33. \int \frac{x^2 dx}{a + bx} = \frac{1}{b^3} \left[ \frac{1}{2}(a + bx)^2 - 2a(a + bx) + a^2 \log(a + bx) \right]$$

$$34. \int \frac{x^2 dx}{(a + bx)^2} = \frac{1}{b^3} \left[ a + bx - 2a \log(a + bx) - \frac{a^2}{a + bx} \right]$$

$$35. \int \frac{x^2 dx}{(a + bx)^3} = \frac{1}{b^3} \left[ \log(a + bx) + \frac{2a}{a + bx} - \frac{a^2}{2(a + bx)^2} \right]$$

$$\begin{aligned}
36. \quad \int \frac{x^2 dx}{(a+bx)^n} &= \frac{1}{b^3} \frac{i-1}{(n-i-3)(a+bx)^{n-i-3}} \\
&\quad + \frac{2a}{(n-i-2)(a+bx)^{n-i-2}} + i \frac{a^2}{(n-i-1)(a+bx)^{n-i-1}}; \quad n \neq 1; 2; 3 \\
37. \quad \int \frac{dx}{x(a+bx)} &= i \frac{1}{a} \log \frac{a+bx}{x} \\
38. \quad \int \frac{dx}{x(a+bx)^2} &= \frac{1}{a(a+bx)} + i \frac{1}{a^2} \log \frac{a+bx}{x} \quad \# \\
39. \quad \int \frac{dx}{x(a+bx)^3} &= \frac{1}{a^3} + \frac{1}{2} \frac{2a+bx}{a+bx} + \log \frac{x}{a+bx} \\
40. \quad \int \frac{dx}{x^2(a+bx)} &= i \frac{1}{ax} + \frac{b}{a^2} \log \frac{a+bx}{x} \\
41. \quad \int \frac{dx}{x^3(a+bx)} &= \frac{2bx-i}{2a^2x^2} + \frac{b^2}{a^3} \log \frac{x}{a+bx} \\
42. \quad \int \frac{dx}{x^2(a+bx)^2} &= i \frac{a+2bx}{a^2x(a+bx)} + \frac{2b}{a^3} \log \frac{a+bx}{x}
\end{aligned}$$

## A.9 The Fourier Transforms

For a piecewise continuous function  $F(x)$  over a finite interval  $0 \leq x \leq \frac{1}{4}$ , the *finite Fourier cosine transform* of  $F(x)$  is

$$f_c(n) = \int_0^{\frac{1}{4}} f(x) \cos nx \, dx \quad (n = 0; 1; 2; \dots) \quad (\text{A:1})$$

If  $x$  ranges over the interval  $0 \leq x \leq L$ , the substitution  $x^0 = \frac{1}{4}x = L$  allows the use of this definition also. The inverse transform is written

$$\overline{F}(x) = \frac{1}{4} f_c(0) + \sum_{n=1}^{\infty} f_c(n) \cos nx \quad (0 < x < \frac{1}{4}) \quad (\text{A:2})$$

where  $\overline{F}(x) = [F(x+0) + F(x-0)]/2$ . We observe that  $\overline{F}(x) = F(x)$  at points of continuity. The formula

$$\begin{aligned}
f_c^{(2)}(n) &= \int_0^{\frac{1}{4}} F''(x) \cos nx \, dx \\
&= -n^2 f_c(n) + F'(0) + (-1)^n F'(\frac{1}{4})
\end{aligned} \quad (\text{A:3})$$

makes the finite Fourier cosine transform useful in certain boundary value problems.

Analogously, the *finite Fourier sine transform* of  $F(x)$  is

$$f_s(n) = \int_0^{\frac{1}{4}} F(x) \sin nx \, dx \quad (n = 1; 2; 3; \dots) \quad (\text{A:4})$$



and

$$\bar{F}(x) = \frac{2}{\pi} \sum_{n=1}^{\infty} f_s(n) \sin nx \quad (0 < x < \pi/4) \quad (\text{A:5})$$

Corresponding to Eq. (A.6), we have

$$\begin{aligned} f_s^{(2)}(n) &= \int_0^{\pi/4} F''(x) \sin nx \, dx \\ &= (-1)^n f_s(n) - n F'(0) - n(-1)^n F(\pi/4) \end{aligned} \quad (\text{A:6})$$

## Fourier Transforms

If  $F(x)$  is defined for  $x \geq 0$  and is piecewise continuous over any finite interval, and if

$$\int_0^{\infty} F(x) \, dx$$

is absolutely convergent, then

$$f_c(\xi) = \int_0^{\infty} F(x) \cos(\xi x) \, dx \quad (\text{A:7})$$

is the *Fourier cosine transform* of  $F(x)$ . Furthermore,

$$\bar{F}(x) = \int_0^{\infty} f_c(\xi) \cos(\xi x) \, d\xi \quad (\text{A:8})$$

If  $\lim_{x \rightarrow \infty} d^n F/dx^n = 0$ , an important property of the Fourier cosine transform,

$$\begin{aligned} f_c^{(2r)}(\xi) &= \int_0^{\infty} \frac{d^{2r} F}{dx^{2r}} \cos(\xi x) \, dx \\ &= (-1)^r \sum_{n=0}^{\infty} \frac{(-1)^n a_{2r-2n}}{(2n)!} \xi^{2n} + (-1)^r \xi^{2r} f_c(\xi) \end{aligned} \quad (\text{A:9})$$

where  $\lim_{x \rightarrow 0} d^r F/dx^r = a_r$ , makes it useful in the solution of many problems.

Under the same conditions,

$$f_s(\xi) = \int_0^{\infty} F(x) \sin(\xi x) \, dx \quad (\text{A:10})$$

defines the *Fourier sine transform* of  $F(x)$ , and

$$\bar{F}(x) = \int_0^1 f_s(\xi) \sin(\xi x) d\xi \quad (A:11)$$

Corresponding to Eq. (A.9) we have

$$\begin{aligned} f_s^{(2r)}(\xi) &= \int_0^1 \frac{d^{2r} F}{dx^{2r}} \sin(\xi x) dx \\ &= i \int_0^1 \sum_{n=1}^{\infty} (i\xi)^{2n-1} a_{2n} + (i\xi)^{2r-1} f_s(\xi) \end{aligned} \quad (A:12)$$

Similarly, if  $F(x)$  is defined for  $-1 < x < 1$ , and if  $\int_{-1}^1 F(x) dx$  is absolutely convergent, then

$$f(\xi) = \int_{-1}^1 F(x) e^{i\xi x} dx \quad (A:13)$$

is the *Fourier transform* of  $F(x)$ , and

$$\bar{F}(x) = \int_{-1}^1 f(\xi) e^{i\xi x} d\xi \quad (A:14)$$

Also, if

$$\lim_{|x| \rightarrow 1} \left| \frac{d^n F}{dx^n} \right| = 0 \quad (n = 1, 2, \dots, r-1)$$

then

$$f^{(r)}(\xi) = \int_{-1}^1 F^{(r)}(x) e^{i\xi x} dx = (i\xi)^r f(\xi) \quad (A:15)$$

## Finite Sine Transforms

$f_s(n)$	$F(x)$
1. $f_s(n) = \int_0^\pi F(x) \sin nx dx \quad (n = 1, 2, \dots)$	$F(x)$
2. $(-1)^{n+1} f_s(n)$	$F(\pi - x)$
3. $\frac{1}{n}$	$\frac{\pi - x}{\pi}$
4. $\frac{(-1)^{n+1}}{n}$	$\frac{x}{\pi}$
5. $\frac{1 - (-1)^n}{n}$	1

$f_s(n)$	$F(x)$
6. $\frac{2}{n^2} \sin \frac{n\pi}{2}$	$\begin{cases} x & \text{when } 0 < x < \pi/2 \\ \pi - x & \text{when } \pi/2 < x < \pi \end{cases}$
7. $\frac{(-1)^{n+1}}{n^3}$	$\frac{x(\pi^2 - x^2)}{6\pi}$
8. $\frac{1 - (-1)^n}{n^3}$	$\frac{x(\pi - x)}{2}$
9. $\frac{\pi^2(-1)^{n-1}}{n} - \frac{2[1 - (-1)^n]}{n^3}$	$x^2$
10. $\pi(-1)^n \left( \frac{6}{n^3} - \frac{\pi^2}{n} \right)$	$x^3$
11. $\frac{n}{n^2 + c^2} [1 - (-1)^n e^{c\pi}]$	$e^{cx}$
12. $\frac{n}{n^2 + c^2}$	$\frac{\sinh c(\pi - x)}{\sinh c\pi}$
13. $\frac{n}{n^2 - k^2} \quad (k \neq 0, 1, 2, \dots)$	$\frac{\sinh k(\pi - x)}{\sin k\pi}$
14. $\begin{cases} \frac{\pi}{2} & \text{when } n = m \\ 0 & \text{when } n \neq m \end{cases} \quad (m = 1, 2, \dots)$	$\sin mx$
15. $\frac{n}{n^2 - k^2} [1 - (-1)^n \cos k\pi] \quad (k \neq 1, 2, \dots)$	$\cos kx$
16. $\begin{cases} \frac{n}{n^2 - m^2} [1 - (-1)^{n+m}] & \text{when } n \neq m = 1, 2, \dots \\ 0 & \text{when } n = m \end{cases}$	$\cos mx$
17. $\frac{n}{(n^2 - k^2)^2} \quad (k \neq 0, 1, 2, \dots)$	$\frac{\pi \sin kx}{2k \sin^2 k\pi} - \frac{x \cos k(\pi - x)}{2k \sin k\pi}$
18. $\frac{b^n}{n} \quad ( b  \leq 1)$	$\frac{2}{\pi} \arctan \frac{b \sin x}{1 - b \cos x}$
19. $\frac{1 - (-1)^n}{n} b^n \quad ( b  \leq 1)$	$\frac{2}{\pi} \arctan \frac{2b \sin x}{1 - b^2}$

## Finite Cosine Transforms

$f_c(n)$	$F(x)$
1. $f_c(n) = \int_0^\pi F(x) \cos nx \, dx \quad (n = 0, 1, 2, \dots)$	$F(x)$
2. $(-1)^n f_c(n)$	$F(\pi - x)$
3. 0 when $n = 1, 2, \dots$ ; $f_c(0) = \pi$	1
4. $\frac{2}{n} \sin \frac{n\pi}{2}$ ; $f_c(0) = 0$	$\begin{cases} 1 & \text{when } 0 < x < \pi/2 \\ -1 & \text{when } \pi/2 < x < \pi \end{cases}$
5. $-\frac{1 - (-1)^n}{n^2}$ ; $f_c(0) = \frac{\pi^2}{2}$	$x$
6. $\frac{(-1)^n}{n^2}$ ; $f_c(0) = \frac{\pi^2}{6}$	$\frac{x^2}{2\pi}$
7. $\frac{1}{n^2}$ ; $f_c(0) = 0$	$\frac{(\pi - x)^2}{2\pi} - \frac{\pi}{6}$
8. $3\pi^2 \frac{(-1)^n}{n^2} - 6 \frac{1 - (-1)^n}{n^4}$ ; $f_c(0) = \frac{\pi^4}{4}$	$x^3$
9. $\frac{(-1)^n e^c \pi - 1}{n^2 + c^2}$	$\frac{1}{c} e^{cx}$

$f_c(n)$	$F(x)$
10. $\frac{1}{n^2 + c^2}$	$\frac{\cosh c(\pi - x)}{c \sinh c\pi}$
11. $\frac{k}{n^2 - k^2} [(-1)^n \cos \pi k - 1] \quad (k \neq 0, 1, 2, \dots)$	$\sin kx$
12. $\frac{(-1)^{n+m} - 1}{n^2 - m^2}$ ; $f_c(m) = 0 \quad (m = 1, 2, \dots)$	$\frac{1}{m} \sin mx$
13. $\frac{1}{n^2 - k^2} \quad (k \neq 0, 1, 2, \dots)$	$-\frac{\cos k(\pi - x)}{k \sin k\pi}$
14. 0 when $n = 1, 2, \dots$ ; $f_c(m) = \frac{\pi}{2} \quad (m = 1, 2, \dots)$	$\cos mx$

## Fourier Sine Transforms

$F(x)$	$f_s(\alpha)$
1. $\begin{cases} 1 & (0 < x < a) \\ 0 & (x > a) \end{cases}$	$\sqrt{\frac{2}{\pi}} \left[ \frac{1 - \cos \alpha}{\alpha} \right]$
2. $x^{p-1} \quad (0 < p < 1)$	$\sqrt{\frac{2}{\pi}} \frac{\Gamma(p)}{\alpha^p} \sin \frac{p\pi}{2}$
3. $\begin{cases} \sin x & (0 < x < a) \\ 0 & (x > a) \end{cases}$	$\frac{1}{\sqrt{2\pi}} \left[ \frac{\sin[a(1-\alpha)]}{1-\alpha} - \frac{\sin[a(1+\alpha)]}{1+\alpha} \right]$
4. $e^{-x}$	$\sqrt{\frac{2}{\pi}} \left[ \frac{\alpha}{1+\alpha^2} \right]$
5. $xe^{-x^2/2}$	$\alpha e^{-\alpha^2/2}$
6. $\cos \frac{x^2}{2}$	$\sqrt{2} \left[ \sin \frac{\alpha^2}{2} C\left(\frac{\alpha^2}{2}\right) - \cos \frac{\alpha^2}{2} S\left(\frac{\alpha^2}{2}\right) \right]^*$
7. $\sin \frac{x^2}{2}$	$\sqrt{2} \left[ \cos \frac{\alpha^2}{2} C\left(\frac{\alpha^2}{2}\right) + \sin \frac{\alpha^2}{2} S\left(\frac{\alpha^2}{2}\right) \right]^*$

\*  $C(y)$  and  $S(y)$  are the Fresnel integrals

$$C(y) = \frac{1}{\sqrt{2\pi}} \int_0^y \frac{1}{\sqrt{t}} \cos t \, dt$$

$$S(y) = \frac{1}{\sqrt{2\pi}} \int_0^y \frac{1}{\sqrt{t}} \sin t \, dt$$

## Fourier Cosine Transforms

$F(x)$	$f_c(\alpha)$
1. $\begin{cases} 1 & (0 < x < a) \\ 0 & (x > a) \end{cases}$	$\sqrt{\frac{2}{\pi}} \frac{\sin a \alpha}{\alpha}$
2. $x^{p-1} \quad (0 < p < 1)$	$\sqrt{\frac{2}{\pi}} \frac{\Gamma(p)}{\alpha^p} \cos \frac{p\pi}{2}$
3. $\begin{cases} \cos x & (0 < x < a) \\ 0 & (x > a) \end{cases}$	$\frac{1}{\sqrt{2\pi}} \left[ \frac{\sin[a(1-\alpha)]}{1-\alpha} + \frac{\sin[a(1+\alpha)]}{1+\alpha} \right]$

$F(x)$	$f_c(\alpha)$
4. $e^{-x}$	$\sqrt{\frac{2}{\pi}} \left( \frac{1}{1+\alpha^2} \right)$
5. $e^{-x^2/2}$	$e^{-\alpha^2/2}$
6. $\cos \frac{x^2}{2}$	$\cos \left( \frac{\alpha^2}{2} - \frac{\pi}{4} \right)$
7. $\sin \frac{x^2}{2}$	$\cos \left( \frac{\alpha^2}{2} - \frac{\pi}{4} \right)$

# Fourier Transforms

$F(x)$	$f(\alpha)$
1. $\frac{\sin ax}{x}$	$\begin{cases} \sqrt{\frac{\pi}{2}} &  \alpha  < a \\ 0 &  \alpha  > a \end{cases}$
2. $\begin{cases} e^{iwx} & (p < x < q) \\ 0 & (x < p, x > q) \end{cases}$	$\frac{i}{\sqrt{2\pi}} \frac{e^{ip(w+\alpha)} - e^{iq(w+\alpha)}}{(w+\alpha)}$
3. $\begin{cases} e^{-cx+iwx} & (x > 0) \\ 0 & (x < 0) \end{cases} \quad (c > 0)$	$\frac{i}{\sqrt{2\pi}(w+\alpha+ic)}$
4. $e^{-px^2} \quad R(p) > 0$	$\frac{1}{\sqrt{2p}} e^{-\alpha^2/4p}$
5. $\cos px^2$	$\frac{1}{\sqrt{2p}} \cos \left[ \frac{\alpha^2}{4p} - \frac{\pi}{4} \right]$
6. $\sin px^2$	$\frac{1}{\sqrt{2p}} \cos \left[ \frac{\alpha^2}{4p} + \frac{\pi}{4} \right]$
7. $ x ^{-p} \quad (0 < p < 1)$	$\sqrt{\frac{2}{\pi}} \frac{\Gamma(1-p) \sin \frac{p\pi}{2}}{ \alpha ^{(1-p)}}$
8. $\frac{e^{-a x }}{\sqrt{ x }}$	$\frac{\sqrt{\sqrt{(a^2 + \alpha^2)} + a}}{\sqrt{a^2 + \alpha^2}}$
9. $\frac{\cosh ax}{\cosh \pi x} \quad (-\pi < a < \pi)$	$\sqrt{\frac{2}{\pi}} \frac{\cos \frac{a}{2} \cosh \frac{\alpha}{2}}{\cosh \alpha + \cos a}$
10. $\frac{\sinh ax}{\sinh \pi x} \quad (-\pi < a < \pi)$	$\frac{1}{\sqrt{2\pi}} \frac{\sin a}{\cosh \alpha + \cos a}$
11. $\begin{cases} \frac{1}{\sqrt{a^2 - x^2}} & ( x  < a) \\ 0 & ( x  > a) \end{cases}$	$\sqrt{\frac{\pi}{2}} J_0(a\alpha)$
12. $\frac{\sin[b\sqrt{a^2 + x^2}]}{\sqrt{a^2 + x^2}}$	$\begin{cases} 0 & ( \alpha  > b) \\ \sqrt{\frac{\pi}{2}} J_0(a\sqrt{b^2 - \alpha^2}) & ( \alpha  < b) \end{cases}$

$F(x)$	$f(\alpha)$
13. $\begin{cases} P_n(x) & ( x  < 1) \\ 0 & ( x  > 1) \end{cases}$	$\frac{i^n}{\sqrt{\alpha}} j_{n+1/2}(\alpha)$
14. $\begin{cases} \frac{\cos[b \sqrt{a^2 - x^2}]}{\sqrt{a^2 - x^2}} & ( x  < a) \\ 0 & ( x  > a) \end{cases}$	$\sqrt{\frac{\pi}{2}} J_0(a \sqrt{a^2 + b^2})$
15. $\begin{cases} \frac{\cosh[b \sqrt{a^2 - x^2}]}{\sqrt{a^2 - x^2}} & ( x  < a) \\ 0 & ( x  > a) \end{cases}$	$\sqrt{\frac{\pi}{2}} J_0(a \sqrt{a^2 - b^2})$

The following functions appear among the entries of the tables on transforms.

Function	Definition	Name
$\text{Ei}(x)$	$\int_{-\infty}^x \frac{e^v}{v} dv$ ; or sometimes defined as $-\text{Ei}(-x) = \int_x^{\infty} \frac{e^{-v}}{v} dv$	Exponential integral function
$\text{Si}(x)$	$\int_0^x \frac{\sin v}{v} dv$	Sine integral function
$\text{Ci}(x)$	$\int_{\infty}^x \frac{\cos v}{v} dv$ ; or sometimes defined as negative of this integral	Cosine integral function
$\text{erf}(x)$	$\frac{2}{\sqrt{\pi}} \int_0^x e^{-v^2} dv$	Error function
$\text{erfc}(x)$	$1 - \text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-v^2} dv$	Complementary function to error function
$L_n(x)$	$\frac{e^x}{n!} \frac{d^n}{dx^n} (x^n e^{-x})$ , $n = 0, 1, \dots$	Laguerre polynomial of degree $n$



## A.10 Bessel Functions

---

### Bessel Functions of the First Kind, $J_n(x)$ (Also Called Simply *Bessel Functions*) (Fig. A.13)

Domain:  $[x > 0]$

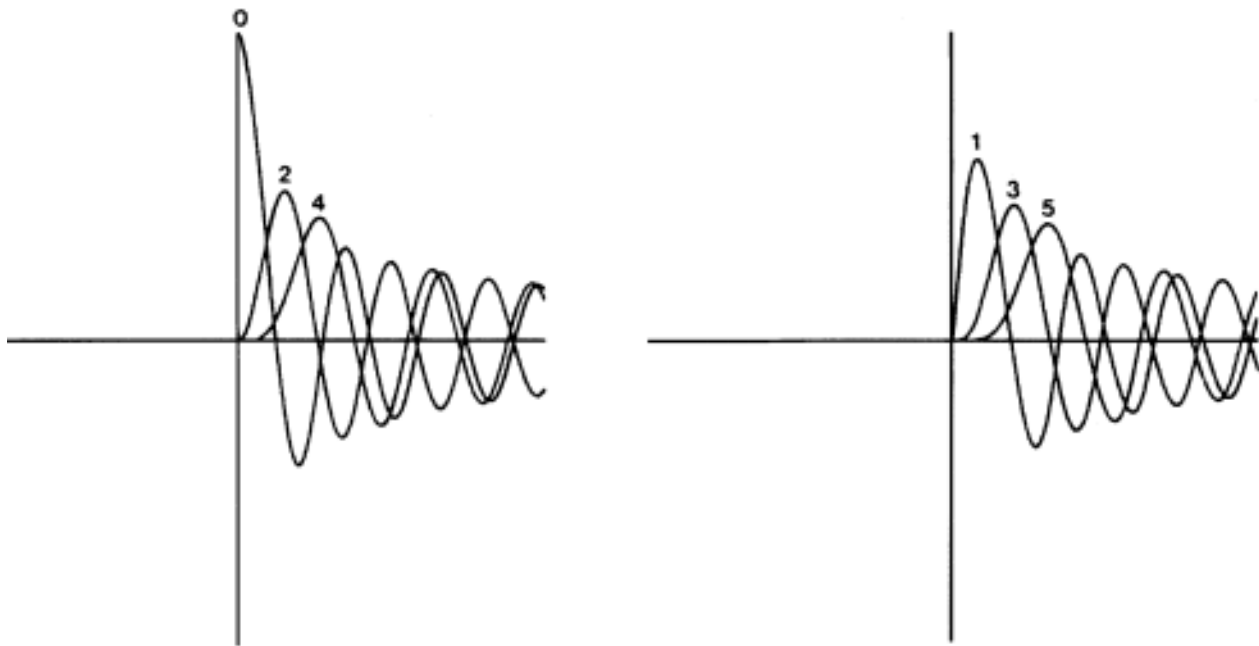
Recurrence relation:

$$J_{n+1}(x) = \frac{2n}{x} J_n(x) - J_{n-1}(x); \quad n = 0; 1; 2; \dots$$

Symmetry:  $J_{-n}(x) = (-1)^n J_n(x)$

0.  $J_0(20x)$
1.  $J_1(20x)$
2.  $J_2(20x)$
3.  $J_3(20x)$
4.  $J_4(20x)$
5.  $J_5(20x)$

**Figure A.13** Bessel functions of the first kind.



### Bessel Functions of the Second Kind, $Y_n(x)$ (Also Called *Neumann Functions* or *Weber Functions*) (Fig. A.14)

Domain:  $[x > 0]$

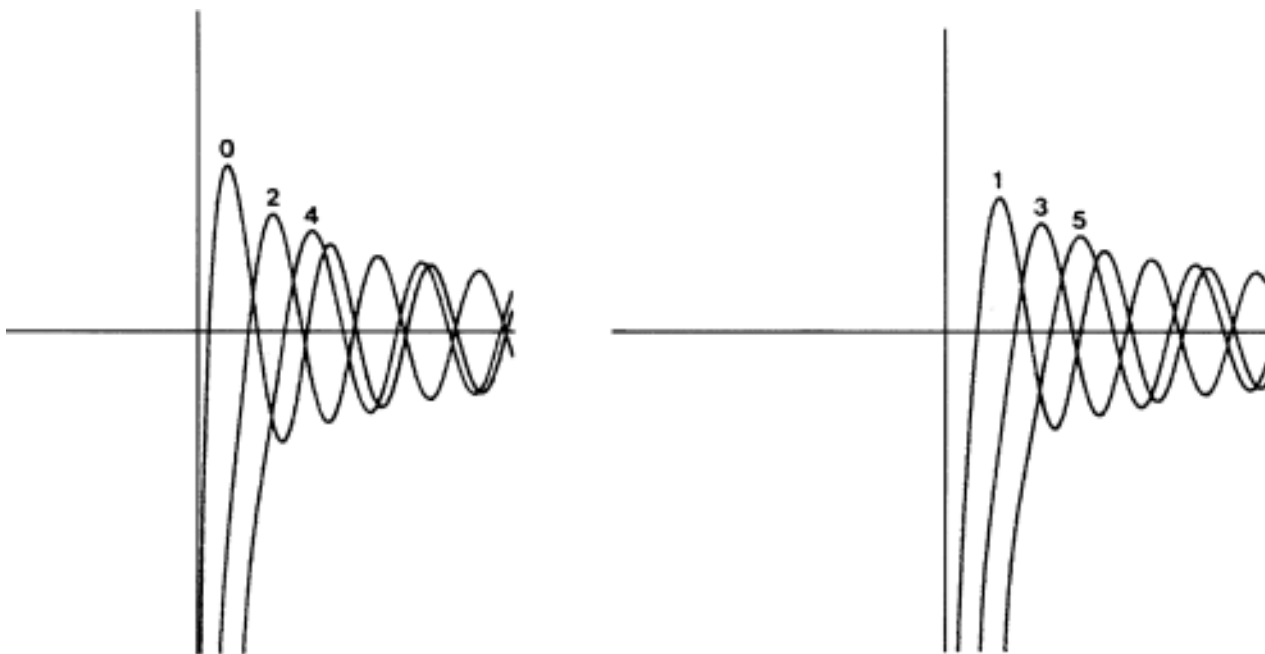
Recurrence relation:

$$Y_{n+1}(x) = \frac{2n}{x} Y_n(x) - Y_{n-1}(x); \quad n = 0; 1; 2; \dots$$

$$\text{Symmetry: } Y_{i-n}(x) = (-1)^n Y_n(x)$$

0.  $Y_0(20x)$
1.  $Y_1(20x)$
2.  $Y_2(20x)$
3.  $Y_3(20x)$
4.  $Y_4(20x)$
5.  $Y_5(20x)$

**Figure A.14** Bessel functions of the second kind.



## A.11 Legendre Functions

### Associated Legendre Functions of the First Kind, $P_n^m(x)$ (Fig. A.15)

Domain:  $[-1 < x < 1]$

Recurrence relations:

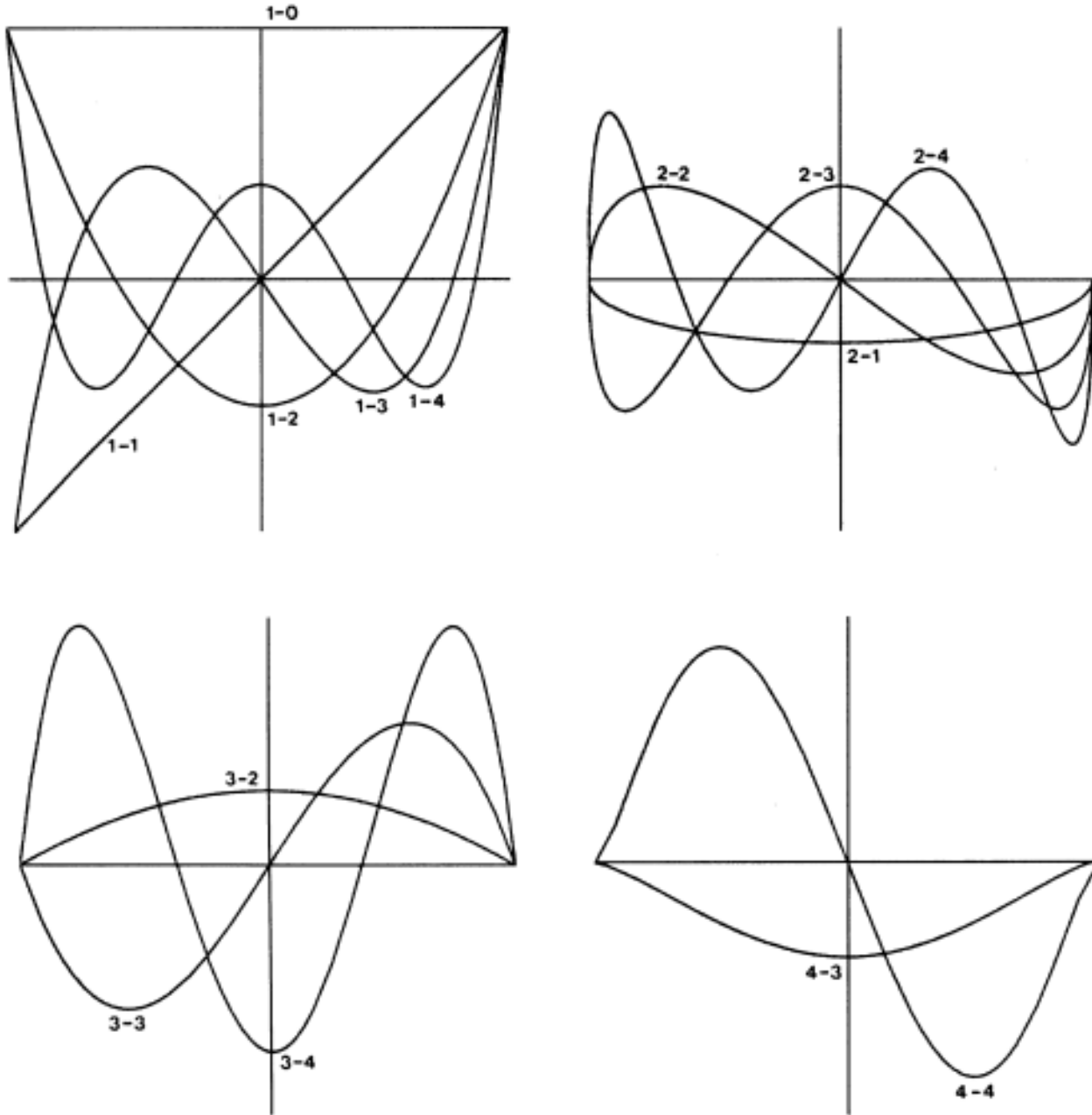
$$P_{n+1}^m(x) = \frac{(2n+1)xP_n^m - (n+m)P_{n-1}^m(x)}{n-m+1}; \quad n = 1; 2; 3; \dots$$

$$P_n^{m+1}(x) = (x^2 - 1)^{1/2} [(n-m)P_n^m(x) - (n+m)P_{n-1}^m(x)]; \quad m = 0; 1; 2; \dots$$

with

$$P_0^0 = 1 \quad P_1^0 = x$$

**Figure A.15** Bessel functions of the second kind.



Special case:  $P_n^0 =$  Legendre polynomials

1-0.  $P_0^0(x)$

1-1.  $P_1^0(x)$

1-2.  $P_2^0(x)$

1-3.  $P_3^0(x)$

1-4.  $P_4^0(x)$

2-1.  $0.25 P_1^1(x)$

2-2.  $0.25 P_2^1(x)$

2-3.  $0.25 P_3^1(x)$

2-4.  $0.25 P_4^1(x)$

3-2.  $0.10 P_2^2(x)$

3-3.  $0.10 P_3^2(x)$

3-4.  $0.10 P_4^2(x)$

4-3.  $0.025 P_3^3(x)$

4-4.  $0.025 P_4^3(x)$

## A.12 Table of Differential Equations

	Equation	Solution
1.	$y' = \frac{dy}{dx} = f(x)$	$y = \int f(x) dx + c$
2.	$y' + p(x)y = q(x)$	$y = \exp[-\int p(x) dx] \{c + \int \exp[\int p(x) dx] q(x) dx\}$
3.	$y' + p(x)y = q(x)y^\alpha$ $\alpha \neq 0, \alpha \neq 1$	Set $z = y^{1-\alpha} \rightarrow z' + (1-\alpha)p(x)z = (1-\alpha)q(x)$ and use 2
4.	$y' = f(x)g(y)$	Integrate $\frac{dy}{g(y)} = f(x) dx$ (separable)
5.	$\frac{dy}{dx} = f(x/y)$	Set $y = xu \rightarrow u + x \frac{du}{dx} = f(u)$ $\int \frac{1}{f(u) - u} du = \ln x  + c$
6.	$y' = f\left(\frac{a_1x + b_1y + c_1}{a_2x + b_2y + c_2}\right)$	Set $x = X + \alpha, y = Y + \beta$ Choose $\begin{cases} a_1\alpha + b_1\beta = -c_1 \\ a_2\alpha + b_2\beta = -c_2 \end{cases} \rightarrow Y' = f\left(\frac{a_1X + b_1Y}{a_2X + b_2Y}\right)$ If $a_1b_2 - a_2b_1 \neq 0$ , set $Y = Xu \rightarrow$ separable form $u + Xu' = f\left(\frac{a_1 + b_1u}{a_2 + b_2u}\right)$ If $a_1b_2 - a_2b_1 = 0$ , set $u = a_1x + b_1y \rightarrow$ $\frac{du}{dx} = a_1 + b_1f\left(\frac{u + c_1}{ku + c_2}\right)$ since $a_2x + b_2y = k(a_1x + b_1y)$
7.	$y'' + a^2y = 0$	$y = c_1 \cos ax + c_2 \sin ax$
8.	$y'' - a^2y = 0$	$y = c_1 e^{ax} + c_2 e^{-ax}$
9.	$y'' + ay' + by = 0$	Set $y = e^{-(a/2)x} u \rightarrow u'' + \left(b - \frac{a^2}{4}\right)u = 0$
10.	$y'' + a(x)y' + b(x)y = 0$	Set $y = e^{-(1/2)\int a(x) dx} \rightarrow u'' + \left[b(x) - \frac{a^2}{4} - \frac{a'}{2}\right]u = 0$
11.	$x^2y'' + xy' + (x^2 - a^2)y = 0$ $a \geq 0$ (Bessel)	i. If $a$ is not an integer $y = c_1 J_a(x) + c_2 J_{-a}(x)$ (Bessel functions of first kind) ii. If $a$ is an integer (say, $n$ ) $y = c_1 J_n(x) + c_2 Y_n(x)$ ( $Y_n$ is Bessel function of second kind)
12.	$(1 - x^2)y'' - 2xy' + a(a + 1)y = 0$ , $a$ is real (Legendre)	$y(x) = c_1 p_a(x) + c_2 q_a(x)$ (Legendre functions)
13.	$y' + ay^2 = bx^n$ (integrable Riccati) $a, b, n$ real	Set $u' = ayu \rightarrow u'' - abx^n u = 0$ and use 14

